Project Number	IST-2006-033789
Project Title	Planets
Title of Deliverable	*The concept of significant properties*. (Part one of a three-part final report from the Digital Object Properties Working Group)
Deliverable Number	D23A
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

Contributors

Person	Role	Partner	Contribution
Lynne Montague		TNA	Author
Eleonora Nicchiarelli		ONB	Author
Henk Matthezing		KBNL	Contributor
Robert Kummer		UZK	Contributor
Johanna Puhl		UZK	Contributor
Bill Roberts		NANETH	Contributor
DOPWG		Various	Contributor

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 seeks to investigate the classification of digital object properties for digital preservation, both in the context of the PLANETS project and in that of the wider digital preservation community. The report aims to address the diverse approaches employed when looking at classifying digital object properties and will, consolidate and summarise the different approaches and give recommendations for future work.

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 (this report));
- *Planets components for the extraction and evaluation of digital object properties* (PLANETS deliverable PC3 – D23B); 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

The aim of this document is to investigate the classification of digital object properties for digital preservation, both in the context of the PLANETS project and in that of the wider digital preservation community. The report aims to address the diverse approaches employed when looking at classifying digital object properties and will consolidate and summarise the different approaches through a state-of–the-art review and give recommendations for future work.

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) (this report);
- *Planets components for the extraction and evaluation of digital object properties* (PLANETS deliverable PC3 – D23B); 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[1]. 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.

[1] See associated Planets report, *Planets components for the extraction and evaluation of significant properties* (deliverable no. PC3-D23B), for more detail

2. The role of properties in digital preservation

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. Chen explains that there are failings in our information infrastructure and a lack of proven methods to ensure that digital information will continue to exist, that we will be able to access it using the technology tools available, or that the information that is available will be authentic and reliable[a].

Wilson observes that a successful preservation strategy must weigh the need to preserve the fixity/integrity of the digital object against the inevitable changes to the technical environment within which it resides[b]. This therefore involves defining the degree to which a digital object may be altered by any preservation actions, whilst still remaining authentic and accessible, and necessitates the question of what should be preserved.

It is not possible, within the scope of this document, to fully discuss the concept of authenticity and its importance in relation to digital properties, as it is an area which continues to be surrounded by much discussion and many definitions. However, the essence of the concept can be seen in the JISC definition which states that:

'an authentic digital resource is one that is what it purports to be, is free from corruption, and is intact in all essential respects.'[c]

Wilson further asserts that authenticity requires:

•Integrity/accuracy - there should have been no unauthorised changes;
•Reliability - the object is what it says it is; and
•Usability - it should be able to be retrieved and rendered[d].

According to guidance published by The National Archives of the UK (TNA), *'a record is considered to be essentially complete and uncorrupted if the message that it is meant to communicate in order to achieve its purpose is unaltered.'*[e]

This concept of authenticity is at the core of assessing whether a preservation action has been successful i.e. whether the migrated digital object retained its authenticity?

2.1 Properties and their relative significance

In the digital preservation field, a property has been defined by Dappert as:

'An abstract attribute, trait or peculiarity suitable for describing preservation objects, actions or environments' [f].

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. As Wilson states:

'*Unless such properties can be defined in a rigorous and measurable manner, cultural memory institutions have no objective framework for identifying, implementing, and validating appropriate preservation strategies, nor for asserting the continued authenticity of their digital collections*.'[b]

Accepting the premise that it is not possible, in most circumstances, to preserve all the elements of a digital object, it becomes a matter not of identifying all the properties of a particular object, but rather of identifying the most important ones, in order to ensure that they are maintained throughout the preservation process and that the authenticity of the digital record is retained. These properties have become known as Significant Properties. Hedstrom and Lee point out that:

'A formal expression of significant properties of complex digital objects has many general and practical applications. Such a model can be applied to appraisal and selection of digital materials, to assessing the risk of information loss associated with various preservation strategies, to the development of preservation metadata, to documenting the basis for preservation decisions, and to the automated management of complex digital objects'[g]*.*

However, this is a far from simple process. As Hedstrom and Lee go on to point out, making decisions about the significant properties of digital objects is extremely complex, due to the different levels of abstraction at which properties can be expressed, the range of available options for creating migrated versions, and the range of behaviours and features that a digital object can exhibit. Further, as Knight and Pennock say, the term *significant* is a relative one and therefore, in order to assess the significance of particular properties, there needs to be criteria against which to measure them[h]. In the InSPECT project, Knight states that the overall purpose of the digital object itself should be assessed in looking at the significance of individual properties and that this would include consideration of the needs of the designated community,[2] compliance with business and legal standards, preservation tool availability, the importance of the property to the digital object as a whole and the capabilities of the institution preserving the object[i].

As Dappert and Farquhar assert, whilst there is a perception that significant properties relate only to intellectual content, it is not always possible to decide out of context whether some properties relate to the intellectual content or are merely circumstantial[j]. They give the example of a number that has been formatted to be red and point out that this may have been done merely to be aesthetically pleasing and would, in this case, be deemed as circumstantial. However it may be red in order to indicate that it is a negative number and therefore it is semantically significant. Its significance, and the significance of any contextual constraints, would need to be explicitly determined by the stakeholder.

Further detail of these concepts will be included within chapter 3 and in companion report, *Planets components for the extraction and evaluation of significant properties,* when discussing the state of the art and the work done within the PLANETS project respectively.

2.2 Terminology

Many terms and definitions have been used for describing the concept of significant properties. For example, significant characteristics, significant properties, essential characteristics, essential properties, essential attributes, aspects and essence have all been used to name the same broad concept in the relevant literature. As regards the definition of these terms, new definitions are frequently put forward and, as Giaretta highlights, these definitions are sometimes inconsistent or unclear, thus creating confusion[k].

The following two chapters set out some of the major work done in the field, and when discussing each project, the terminology used by the individual project has been used. However, in terms of the DOPWG, a shared vocabulary and model for digital object properties was agreed during a face-to-face meeting in Den Hague in October 2009[l]. Firstly, five concepts were formulated:

1. A Record (within archives) or Publication (within libraries) consists of Metadata and a (digital) Preservation Object;
2. A (digital) Preservation Object (the thing we want to preserve) is represented in a Bytestream plus Metadata[3];
3. A Performance consists of a Preservation Object and its Environment i.e. a rendering;

[2] Within an Open Archival Information System (OAIS) a designated community is defined as 'An identified group of potential Consumers who should be able to understand a particular set of information. The designated community may be composed of multiple user communities.' The term 'stakeholder' is often used instead of designated community. Both terms can, but don't necessarily, mean the user of a preservation object.

[3] The metadata mentioned in points one and two is different. For example, the metadata in point one may include administrative metadata such as what kind of document it is or how many pages it has plus descriptive metadata about the intellectual content of the object e.g. the author, title, keywords etc. The metadata in point two may also contain descriptive metadata, such as how the object was produced and on what machinery, plus technical metadata about the structure and architecture of the object.

4. There is a conceptual distinction between (technically) Extractable Properties and Observational Properties;
5. A Preservation Object has Characteristics which consist of Properties with a Value

These were then backed up by four further statements where the idea of a Creator, whose digital creation we are preserving, and of a User, who is the audience, were introduced:

6. We perform (Performance) a Preservation Object to meet the requirements for how the Stakeholder[4] wishes the Record or Publication to be experienced;
7. The Creator and his/her Context specifies all the Characteristics of a Preservation Object but not all of their subsequent significance;
8. The Creator and his/her Context specifies some Characteristics as significant in the Context of asserting Authenticity and Intent;
9. The User and his/her Context specifies some Characteristics as significant in interpreting the Performance.

The interpretation set out at point 5 (and elaborated in section 3.1.11 below) means that it is the Characteristic, i.e. the property/value pairing, that is preserved and which can therefore have significance assigned to it, rather than the abstract property on its own. Technically this, therefore, makes the term significant properties wrong and is the reason that the word ‘characteristics’ was used instead of ‘properties’ when formulating the nine statements above. However in reflection, it has been observed during the course of researching this paper, that the term ‘significant properties’ appears to be the more widely used and understood term within the domain and for this reason it will be used in this report. It should be understood to relate to the significance of both the properties and their associated values.

2.3 Significant Properties and Representation Information

Another area of confusion in the field is between the concepts of significant properties and representation information. Put simply, as set out by Brown[m] [n], they are two distinct concepts; with significant properties being attributes of the OAIS defined Information Object, and representation information being relevant to the Data Object[5]. Knight and Pennock assert that there has been inconsistency in the way they are interpreted[h], and this is supported by Dappert and Farquhar. However, as the latter point out, the difference between the two becomes clear when it is realised that representation information is needed to make sense of a particular data object for a specific designated community at a specific time. Unlike significant properties, it doesn’t indicate the constraints that should be placed on the object during transformations over time, nor specify what the characteristics of the resultant transformed object should be[j]. Both concepts are classed as digital object properties.

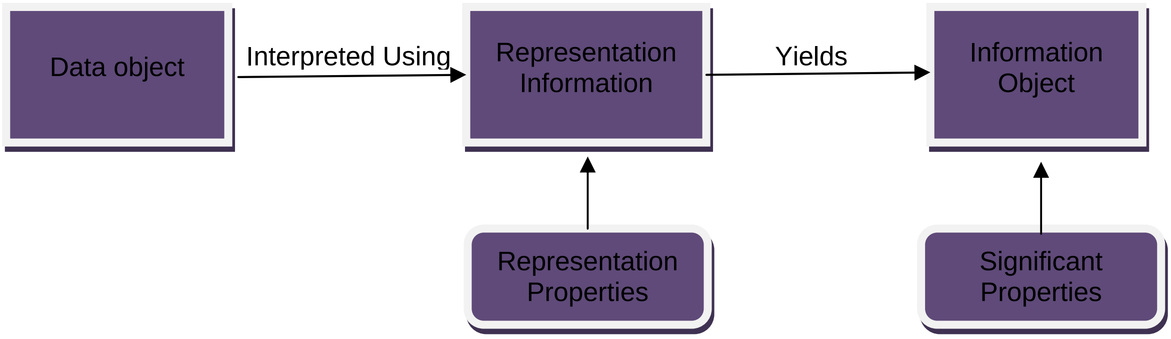

Figure 1: The positioning of Representation and Significant Properties within the OAIS model[6].

[4] Whilst the Stakeholder could be the user, this is not necessarily so.

[5] See section 3.1.6 below for more detail on OAIS terms.

[6] Adapted from Brown, A. (2008)[n].

3. The State of the Art

3.1 Previous work

The concept of significant properties has been the focus of a considerable amount of work over the last decade. This section does not intend to provide a comprehensive review of all of this work. Rather it provides an overview of some of the major milestones in thinking and summarises the current state of the art in the varying approaches to significant properties.

3.1.1 Rothenberg and Bikson - Carrying Authentic, Understandable and Usable Digital Records Through Time (1999)°

In 1999 Rothenberg and Bikson undertook a study to consider the technical issues surrounding the long-term digital preservation of Dutch government records. They produced one of the first reports in which significant properties (although not called this in the report) were identified as an important consideration for a digital preservation strategy. In doing this they developed three key outputs; a digital preservation strategy, a framework within which preservation could be undertaken and a testbed design to support related experimentation.

The strategy uses both top-down and bottom-up approaches. Using a top-down approach, the strategy identifies archiving requirements by analysing relevant organisational processes and functions in order to produce a set of essential characteristics that need to be preserved. From this, the criteria needed to ensure authenticity in a preserved digital record can be specified. Their paper asserts that although these authenticity criteria may vary with different types of record, the intent behind them is to ensure that the original content, context, appearance, structure and behaviour of the record is preserved as necessary.

Through analysis of the authenticity criteria, the attributes of a digital record that require preservation can be identified and, using a bottom-up approach, these attributes are mapped to the technical properties of preservation methods. In this way technical processes capable of preserving the record attributes are identified and a suitable preservation approach selected.

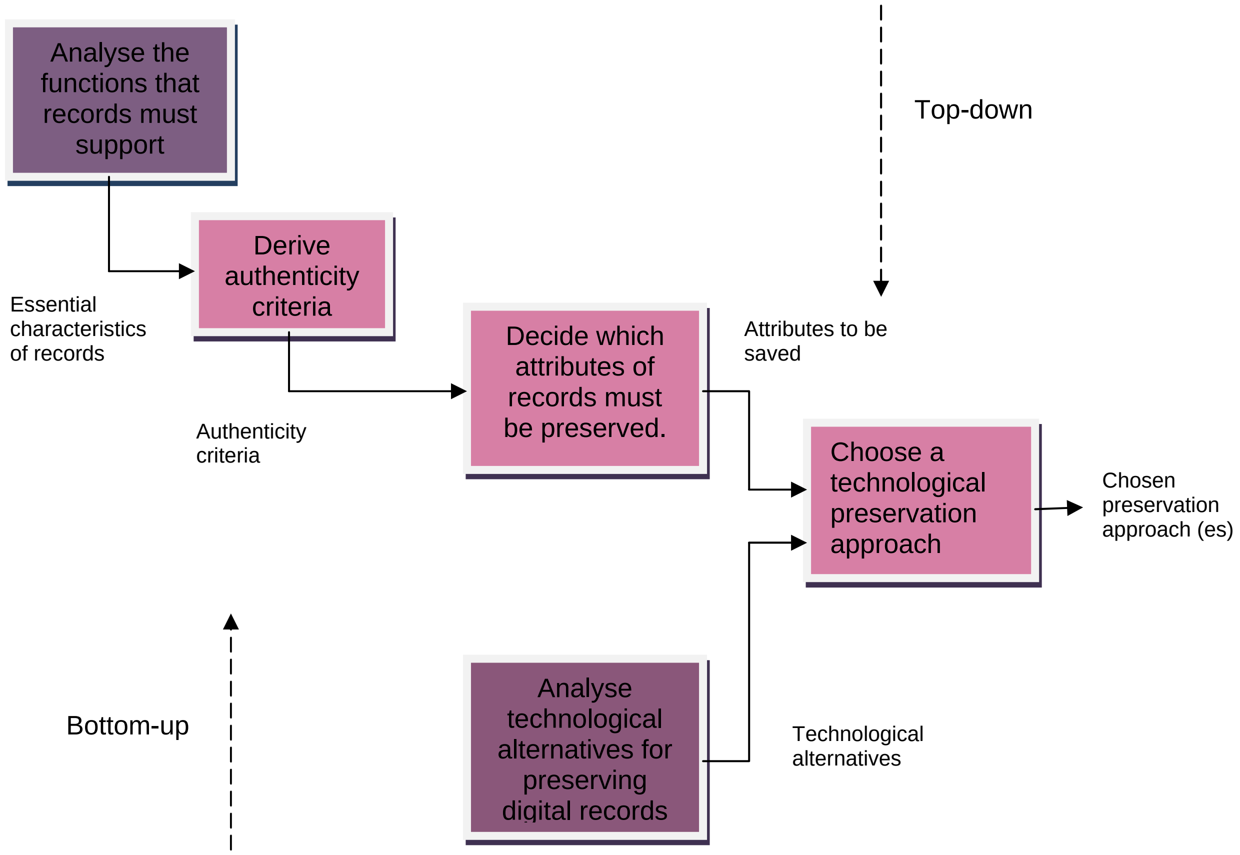
Analyse the functions that records must support
Top-down
Derive authenticity criteria
Essential characteristics of records
Decide which attributes of records must be preserved.
Attributes to be saved
Authenticity criteria
Choose a technological preservation approach
Chosen preservation approach (es)
Bottom-up
Analyse technological alternatives for preserving digital records
Technological alternatives

Figure 2: The Rothenberg and Bikson preservation strategy

3.1.2 InterPARES 1 (1999-2001)[p]

The International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 1 project looked at preserving the authenticity of documents throughout the preservation process, from a diplomatics perspective[7]. A Template for Analysis was produced, in order to test the hypothesis that traditional archival diplomatic principles could be applied to digital records. The template was used to categorise the properties of a digital record into four broad categories (with sub-categories), in order to assess which properties are important for maintaining authenticity. The four categories were: Documentary Form, Annotations, Context and Medium.

The categorisation of properties in this way was a useful early example of thinking about properties and their significance for a specific purpose, i.e. ensuring authenticity. However, the general diplomatics–specific approach was seen to have limitations as an analytical tool. It was also recognised that it was difficult to apply the traditional ideas of 'what a record is' to the complex and dynamic electronic records and systems that we commonly use now.

3.1.3 Cedars (1998-2002)[q]

The CURL Exemplars in Digital Archives (Cedars) project was set up to investigate a variety of digital preservation issues, with particular focus on university libraries. In considering suitable technical approaches, the project talked about the need to preserve 'all the significant properties of the original' when migrating a digital object for preservation. They asserted that by identifying these significant properties, key preservation format decisions could then be evaluated against the need to preserve all of these properties. However, whilst the project gave some examples of what may be considered significant, it did not formalise a definition of what it perceived significant properties to be.

3.1.4 CAMiLEON (1999 – 2002)[r]

The Creative Archiving at Michigan and Leeds: Emulating the Old on the New (CAMiLEON) project was tasked with developing technical strategies for the long-term preservation of digital objects. As with the overlapping Cedars project above, the importance of significant properties was highlighted, and the Cedars concept was built upon. In a related paper in 2002, significant properties were defined as:

'those properties of digital objects that affect their quality, usability, rendering, and behaviour'

The paper goes on to point out that by identifying the properties that affect the 'look-and-feel' of the digital object, and those that are regarded as important by the relevant designated communities, the most appropriate preservation methods can be chosen to preserve the properties.

As a way of formally expressing significant properties, the project defined a conceptual model for complex digital objects and their components, and then mapped significant properties to common digital object component types. It was deemed that such a model would have numerous applications in appraising and selecting digital materials, assessing preservation strategies and the associated risk of loss of information, developing metadata for preservation, documenting preservation decisions and helping with the automated management of digital objects.[g]

3.1.5 Digital Preservation Testbed (Testbed Digitale Bewaring) (2000-2003)[s]

This Dutch government project was set up to develop the testbed proposed by Rothenberg and Bikson above, and in doing so, further investigated the categorisation of properties relevant to authenticity. The assumption was made that different types of digital record have different preservation and authenticity requirements. The project tested three different preservation strategies; migration, emulation and XML, on four distinct record types; text, email, databases and spreadsheets. Records ingested into the testbed were analysed according to the categories of

[7] Archival Diplomatics were first used in the 17th century as a way of analysing and determining the authenticity of a document. Originally used specifically to determine legal authenticity, the principles were later applied to ensure historical legitimacy.

content, context, appearance, structure and behaviour, as specified by Rothenberg and Bikson, and authenticity requirements were developed from these[t]. It was noted that the primary influence on the importance of the categories was business process, but that for each type of record the importance of an individual category might vary. For example, while appearance may not be deemed important for an email, it may be vital for a text document.[u]

3.1.6 OAIS (2002)[v]

This is an International Organization for Standardization specification for an Open Archival Information System (OAIS) Reference Model.

'An OAIS is an archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community'.

The reference model provides a guidance framework of tools, terminology, concepts and processes for the preservation of, predominantly digital, information. It is particularly pertinent to organisations with a long-term responsibility for keeping information accessible.

The OAIS model for obtaining information from data consists of a Data Object, which in a digital context will be a sequence of bits; Representation Information which accompanies the Data Object and interprets and gives meaning to the bits in the Data Object; and a resultant Information Object which is the rendered, user-readable recreation of the Data Object.

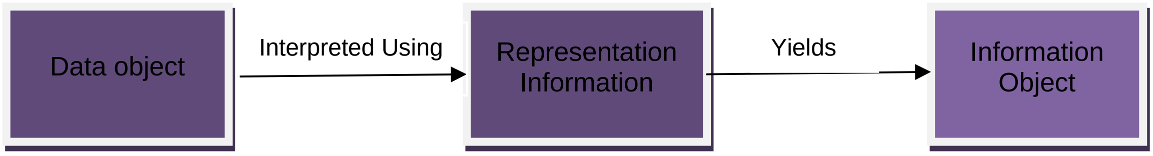

Figure 3: OAIS Model – Obtaining information from data

Whilst the model does not specifically talk about significant properties, the InSPECT project took the view that they are the characteristics of an Information Object that must be preserved in order to maintain its meaning and accessibility through its recreation or the transformation of the associated data object[w]. This viewpoint is supported by Hedstrom and Lee[g].

3.1.7 The National Archives of Australia's Performance Model (2002)[x]

In 2002 the National Archives of Australia (NAA) developed The Performance Model for digital records. As Brown points out, this model closely corresponds to the OAIS model outlined above.[m] The idea behind the Performance Model is that each experience of a digital record e.g. a viewing of a document, is a performance of technology and data interacting, and becomes a new 'original copy'. In this way, several viewers of a record experience equivalent performances of the original, but not the original itself.

The source of the record is the fixed data, which provides it with meaning, but only when combined with the technology to render it on an output device such as a screen. Under the model, this technology is known as the process. It is this interaction between source and process that creates the performance.

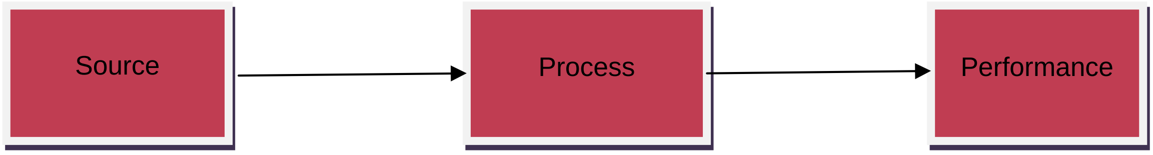

Figure 4: Adapted from the NAA Performance model

Under this model, it is not necessary for the process to stay the same (and in fact due to rapid technological obsolescence and the deterioration of storage media it is not reasonable to expect that the process will stay the same), as long as the 'essential' parts of the performance can be recreated. To this end the NAA articulated the concept of the 'essence' of a record as being the essential characteristics that need to be preserved in order to maintain the meaning of a performance. By determining the essential characteristics at the outset, resources are not expended on maintaining characteristics that do not affect a record's meaning in archival terms.

3.1.8 DELOS (2005)[y]

In 2005, the preservation cluster of the Delos Association for Digital Libraries developed an eight-step process, Utility Analysis, to test and evaluate digital preservation strategies. The Utility Analysis tool, which is traditionally used for infrastructure projects, was adapted for the digital preservation field as part of the project.

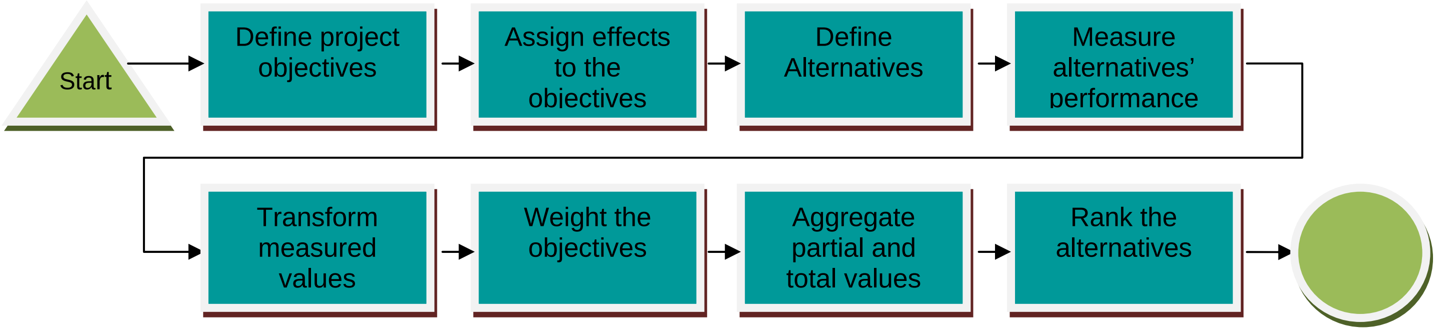
Start
Define project objectives
Assign effects to the objectives
Define Alternatives
Measure alternatives' performance
Transform measured values
Weight the objectives
Aggregate partial and total values
Rank the alternatives

Figure 5: The Delos Utility Analysis workflow

The first step of the Utility Analysis is to define the objectives of the project in question, and this is done by building an 'objective tree', i.e. a hierarchical tree-based structure within which both the high- and low-level goals and characteristics of the project are organised. Information within the objective tree is categorized into two major groups, Digital Object Characteristics, which are deemed by the project to include files, software packages and operating systems; and Process Characteristics, which include any other characteristics that are not directly part of the digital object itself[8]. It can be seen that in considering both groups of characteristics, the significance of a given characteristic is key to whether and where it should be included in the hierarchy. In considering file characteristics, the project suggests using three of the Rothenberg and Bikson categories of Structure, Appearance and Behaviour. For the Process Characteristics, authenticity, stability, scalability and usability are suggested as the four sub-goals within which they should be considered.

Top level	**Level 2**	**Level 3**	**Selected level 4 criteria**
File Characteristics	Appearance	Page	borders, numbering, . . .
		Paragraph	formation, . . .
		Character	font style, colour, . .
		Sound	bit rate, . . .
		Video	frame rate, . . .
	Structure	Caption, tag description, . . .	

[8] Although strictly a Process characteristic, the project regarded cost as important enough to be considered as a third high level category.

	Behaviour	Reaction on user inputs, search, links, . . .	
Process Characteristics	Authenticity	Traceability of changes, . . .	
	Stability	Supplier independency, . . .	
	Scalability	Data increase, format range increase, . . .	
	Usability	Process complexity, functionality, . . .	

Figure 6: Objective tree: Hierarchical order of goals[9]

By completing the remaining steps in the workflow, alternative preservation strategies can be measured, weighted and evaluated in a numerical form. The team undertook two case studies, one on the preservation of a journal and one on audio files, to illustrate how this approach can work in practice.

3.1.9 InSPECT (2007-2009)[z]

InSPECT (Investigating the Significant Properties of Electronic Content over Time) was a two-year, JISC funded project which aimed to expand the concept of what it called Significant Properties; assessing what makes a property significant, developing a methodology to identify significant properties for particular digital object types, and testing these properties through file-format migrations using example format types.

The term ‘significant properties’ was initially defined for the project by Wilson as:

‘The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record.’

However this definition was later adapted, and the term Information Object introduced, to ally it with the OAIS Reference Model.

‘The characteristics of an Information Object that must be maintained over time to ensure its continued access, use, and meaning, and its capacity to be accepted as evidence of what it purports to record’[aa]

In developing a practical methodology for evaluating the significance of properties, the project acknowledged that it is not possible to provide a single catch-all interpretation of what is significant. Rather, as pointed out by the OAIS Reference Model, the process of assigning significance is a subjective and changeable one, dependent on the number and type of stakeholders within the designated community and their needs[i].

The methodology and subsequent assessment framework, which was developed in two iterations, aim to provide a process through which an evaluator can assess the stakeholders, analyse the properties of a digital object, and make appropriate decisions about which digital object properties are significant in any given case. This takes account of the level of loss that is acceptable during the preservation process, i.e. whether to take a risk-averse or a risk-tolerant position to loss. The concept of performance (as defined by the NAA above), is key to the work done within the project.

The first iteration of the assessment framework provided a structured process and template. This could be used by an evaluator of a digital object, firstly to evaluate the digital object as a whole, and then to break it down into its sub-components, and finally its properties. In this way, when considering the set objectives of stakeholders from the relevant Designated Community, the importance of individual properties and acceptable levels of tolerance to loss can be established.

[9] Adapted from *A framework for documenting the behaviour and functionality of digital objects and preservation strategies.* [y]

This was the process followed by the project team in the evaluation of the significant properties of four digital object types; email, audio, raster images and structured text[10]. Once sets of properties for each object type had been defined, the team carried out format migrations on representation test files, and compared the characterised source and migration manifestations to see how successfully the significant properties were migrated. Four test reports setting out the process were produced[bb]

The second iteration of the assessment framework adopted a revised version of the Function-Behaviour-Structure framework developed by John Gero in 1990. Originally developed to help engineers and designers create and reengineer systems, it was revised within the InSPECT project to enable the analysis of requirements, i.e. the analysis of the object's current functionality and the stakeholder's desired functionality in future manifestations. In this way, the digital object can be reformulated as appropriate.

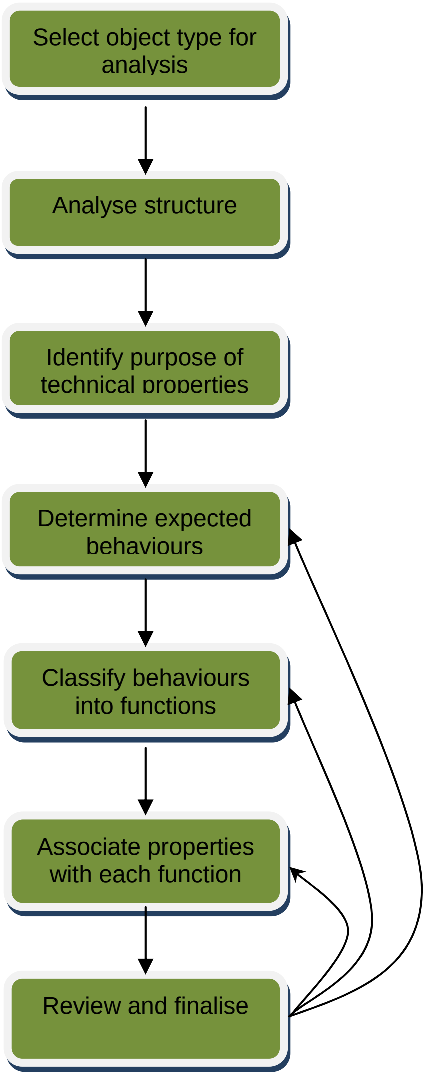
Select object type for analysis
Analyse structure
Identify purpose of technical properties
Determine expected behaviours
Classify behaviours into functions
Associate properties with each function
Review and finalise

Figure 7: The object analysis stage of the InSPECT framework for determining significance.[aa]

In order to support the management of digital objects, the project also developed a Significant Properties Data Dictionary, to provide a structured way of recording, evaluating and assigning value to digital objects.[cc]

[10] See also the work of four other JISC-funded projects, completed before the end of the Inspect project, which looked at significant properties for the specific digital object classes of vector images, moving images, software and e-learning objects, using different methodologies for each (http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/2008sigprops.aspx), plus the paper written by Gareth Knight and Maureen Pennock which compares the significant properties identified by each of these 4 projects[h].

One of the main findings of the project is that consideration of significant properties needs to be routinely included in digital curation workflows. In order for this to be possible, further work needs to take place to ensure that there are appropriate characterisation and metadata extraction tools available to adequately measure significant properties.

3.1.10 Significant Properties In the Laboratory (SPIL) (2009)[dd]

This is a JISC-funded follow-on project from InSPECT that will look at applying the InSPECT methodology to specific scientific data, in order to show how significant properties can be used practically in the curation process of a specific use case. The project aims to produce two main deliverables: firstly a set of software components and services to extract significant properties, validate significant properties, and carry out format conversions, alongside a demonstrator that uses these services; and secondly a final report, incorporating a case study. At the time of writing, information about the status of these deliverables is not available.

3.1.11 IIPC - Long-term Preservation of Web Archives – Experimenting with Emulation and Migration Methodologies (2009)[ee]

Although looking predominantly at Web Archives, this 2009 International Internet Preservation Consortium (IIPC) paper also mentions the conceptual issues involved with the interpretation of digital preservation policy in general (using the policy of the National Library of Australia (NLA) as an example). It also states that further qualification of terms is needed to implement preservation policy practically.

Accepting that it is not possible to preserve a digital object over the long-term without some change in the object, author Stawowczyk Long points out that an institution must clearly understand which aspects of any digital object it intends to preserve. In looking at this, the Digital Preservation department of the NLA has introduced the high- level concept of ‘preservation intent’, of which significant properties will be a part. Stawowczyk Long suggests that rather than going through the lengthy and complex process of defining all the significant properties of a digital object, it would be more practical for a collecting institution to look at them ‘through a prism’ of preservation intent.

He goes on to point out that, in addition to the institution’s preservation intent, understanding the creator’s intent, contextual information and technical information can help to determine the important aspects to preserve.

3.1.12 Extractors

In addition to the work highlighted above, much work has been undertaken to develop file property extractors. Please see associated Planets report, *Planets components for the extraction and evaluation of significant properties* (deliverable no. PC3-D23B) for a summary of the major work done in this area.

3.2 Sources of properties

Digital object properties are derived from a wide variety of different sources, thus adding an extra layer of complexity when considering their preservation. This section describes the different sources of properties, both conceptually and logically, and explains what is included within each type.

3.2.1 Conceptual Digital Object Property Types

During the course of the work done by the DOPWG, the concepts of Extractable, Observational and Performance properties have been discussed at length. This section attempts to summarise these concepts and illustrate the issues that have arisen in attempting to provide authoritative definitions.

The simple definitions below are given in order to provide a general idea of what the concepts involve. They are not regarded as definitive and should be read in conjunction with the discussion of issues that follows.

Extractable Properties: These can be defined as properties that can be extracted by software from the file itself.

Observational Properties: These can be defined as properties that can be determined by human observation.

Performance Properties: These can be defined as properties that relate to a Performance, i.e. the rendered combination of the Preservation Object and its environment

Institutional/Process/Policy Properties[11]: These can be defined as non-technical properties related to organisational processes or policy. They do not fall into either the extractable or observational categories[12].

Issues and observations highlighted during DOPWG discussions:

- Using these definitions, the observational and extractable categories may overlap, i.e. it may be possible to also observe certain extractable properties once they are rendered. Therefore, it may be deemed necessary to refine the definition of observable properties to properties which can be determined by human observation but which are not extractable by software, in order to clarify which category a property falls in.

- Even refining the definition of observational properties does not solve the issue in all circumstances. It was noted during the work on the PLANETS ontology that there are cases where a property can be described differently depending on the format. For example, humans can observe the property, 'page number', but because it can also be extracted (e.g. for specific formats within the Microsoft Office family), it would be regarded as an extractable property under the above definition. However, it is not extractable from PDF files, and therefore would be observational for this format.

- When considering extractable properties, it needs to be decided whether a property is deemed extractable if it can be extracted in principle (even if there is no software available to do this) or only if there is a working extractor available.

  If using the former option, the question arises as to how to determine what is or isn't extractable in principle, a difficult question without a large degree of computing knowledge. It also raises the question of how useful it is to know something is extractable in principle. The latter option makes it easier to definitively state whether a property is classed as extractable, but it has the disadvantage of having to reclassify properties when a new extractor becomes available. It also means that there is a class of properties that are neither extractable nor observational.

- During the DOPWG discussions it was stated that observational properties are always performance properties because humans will only be able to observe properties that have been performed.

- Whilst the DOPWG has typically talked about extractable properties as being those which are extracted from the file alone, it was envisaged that a new class or sub-class of 'extractable performance' properties would also be needed for properties that can only be determined from the combination of file and rendering software but which can be measured by software. Whilst these may be quite rare this category would be applicable where the rendering software carries out a processing algorithm. For example the location of page breaks in a Word file is not contained in the file but is calculated by the application when the file is viewed. This can alter according to the printer that Word is currently 'planning' to print to. Another example would be an HTML page with embedded javascript which runs on loading the page to e.g. adjust the size

[11] A definitive name for this type of property has yet to be decided

[12] See section 2.4.1 of associated Planets report, *Planets components for the extraction and evaluation of significant properties* (deliverable no. PC3-D23B), for more detail about these types of property.

of an element on the page according to the browser window size, or to insert the current date and time.

3.2.2 Properties within the PLANETS-wide ontology[13]

Due to the conceptual issues surrounding the definition of properties, as illustrated above, it was necessary for the team undertaking the development of the PLANETS-wide ontology to take a view of how they were going to categorise and define properties, in order to progress the work of mapping them within the ontology. They defined the two categories of observational and extractable properties identified above, as follows:

Extractable Properties: These are defined as properties which are found within the file itself and which can be extracted by a common piece of software e.g. Image resolution.

Observational Properties: These are defined as properties that, whilst related to the file itself, are not found within it and cannot be extracted by commonly used software, but which are observable by humans e.g. Image quality. It was observed that Performance Properties are not contained within the file itself and are classed as observational properties.

This categorisation of properties within the ontology is designed to indicate to the user whether properties should be compared using software or human observation. Due to resource limits within the project, a decision was made to create a third option for the ontology called, Uncategorised Properties, in order to take into account both the Institutional/Process/Policy properties (mentioned above), and any properties which have yet to be assessed.

3.2.3 Logical digital object property types

Notwithstanding the conceptual issues involved in defining the meaning of certain terms, it can be seen that the properties relevant to the preservation of a digital object can be found at a variety of different levels, both intrinsic and extrinsic to the object itself.

In their paper, *Significance is in the Eye of the Stakeholder*,[j] [14] Dappert and Farquhar propose a conceptual model which asserts that properties and characteristics are present at the class levels of preservation object, environment and preservation action. The preservation object is deemed to have three subclasses; physical objects, representation objects and logical objects, and properties will be found at every level.

[13] See associated Planets report, *Specification of a Planets-wide Ontology of properties for digital preservation needs* (deliverable no. PC3-D23C), for the full ontology report

[14] As discussed in associated Planets report, *Planets components for the extraction and evaluation of significant properties* (deliverable no. PC3-D23B).

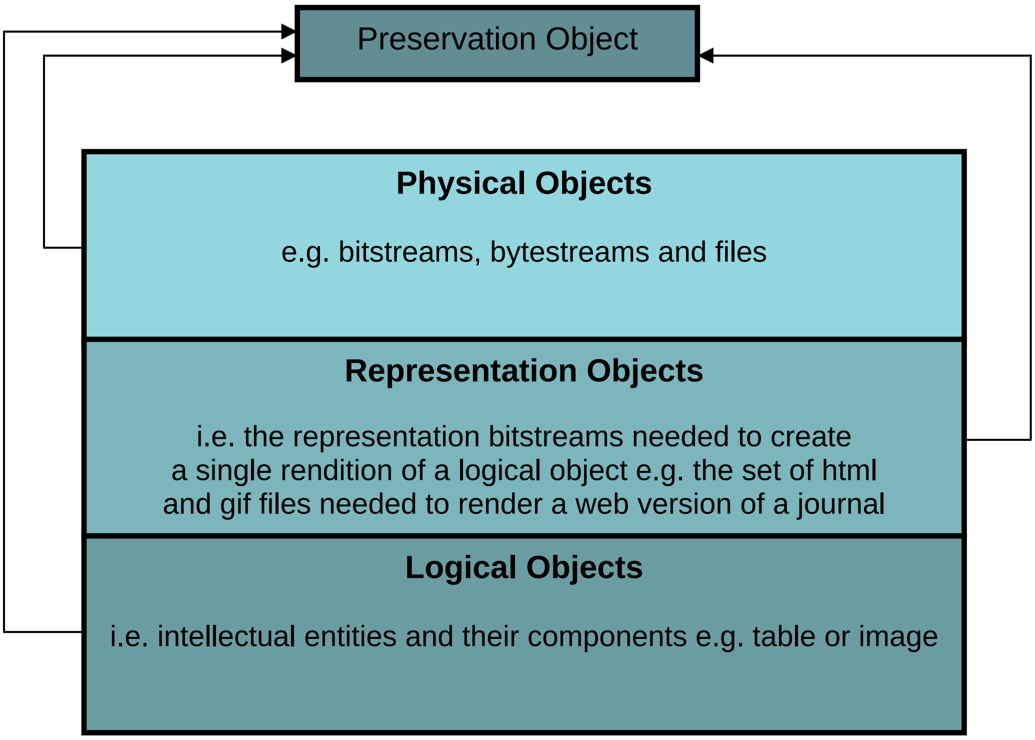
Preservation Object
Physical Objects
e.g. bitstreams, bytestreams and files
Representation Objects
i.e. the representation bitstreams needed to create a single rendition of a logical object e.g. the set of html and gif files needed to render a web version of a journal
Logical Objects
i.e. intellectual entities and their components e.g. table or image

Figure 8: Adapted from Dappert and Farquhar, Preservation Object Subclasses[j]

The preservation object will have an associated environment or environments within which it operates, and these will have sub-environments such as software, hardware, middleware, designated community, budgetary and legal. All of these sub-environments will have properties relevant to the interpretation and performance of the object. Finally, the preservation actions undertaken by an institution in order to protect their collection may also have properties that need to be complied with e.g. those associated with copyright restrictions.

Put simply:

Objects have properties: e.g. file size, page count, pixel depth….

Environments have properties: e.g. a TIFF viewer, a JP2K viewer, processor type, memory….

Actions have properties: e.g. cost, speed, memory consumption, single item or batch, metadata retention….[ff]

In a separate report, *Obtaining and Relating Digital Object Properties in Digital Preservation*, which was undertaken as part of the work of the DOPWG[gg], Dappert points out that alignment problems can arise when different preservation services express properties at different levels. This can be exacerbated when different techniques are used for ascertaining a value for the same property, or when file formats use fundamentally different paradigms.

Through analysis of preservation plans and services, this paper sets out the eight different types of property expression that were found, and assesses the ease with which they can be derived. These properties are classified, according to the method by which their values are obtained, as follows:

1. Extractable, File-Based Properties
2. Extractable, Complex Properties
3. Non-Extractable, Complex Properties
4. Implicit Semantics Properties
5. Inferable Properties
6. Non-Deterministic Properties
7. Random Properties

8. Indeterminable Properties

In addition, two further sets of properties were identified as being outside the scope of the paper; Representation Independent Properties and User Experience Properties. The paper goes on to illustrate the difficulties in aligning some of these categories within a property ontology, which needs to be able to describe properties semantically in order that they can be compared or derived. For some categories of property, however, this is either impossible (category 8) or possible only with difficulty (categories 6 and 7).

In conclusion the report identifies several areas that need to be considered in undertaking future work, including:

- The need to consider '*incomplete, approximate and heuristic* values' in the assessment of characteristics;
- The need to define an expression language in order to define derived properties;
- The need for robust aggregate comparisons of digital object property values; and
- The need to capture the semantics of similar properties.

4. Recommendations for future development

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

4.1 Process-related vs. object-related properties

One of the main changes in thinking concerning digital object properties is the shift from thinking purely in terms of significant properties towards thinking in terms of observational vs. extractable properties. Part of the reason for this shift is the problem of inherent subjectivity, and the need to make the properties concept objective. Although the newly adapted observational vs. extractable dichotomy does not solve the subjectivity problem completely, these new concepts hint at a more fundamental issue.

While the observational (by humans) property concept relates to a digital object as part of a (rendering) process, the extractable property concept relates to the more static state of the digital object. An analogy can be found in the wave-particle duality concept as it is used in quantum mechanics to describe both the wave-like and particle-like properties of all matter, while at the same time addressing the inadequacy of classical concepts of 'wave' and 'particle'.[15] Within the digital preservation domain it may be interesting to further investigate the nature of this (computer) process - (digital) object duality.

[15] See wikipedia for an explanation: (http://en.wikipedia.org/wiki/Wave%E2%80%93particle_duality)

5. References

[a] Chen, S. (2001). The paradox of digital preservation. Retrieved 8th March 2010, from http://www.fpdigital.com/Resource/Files/ParadoxOfDigitalPreservation.pdf

[b] Wilson, A. (2007). Significant properties report, InSPECT Work Package 2.2, Draft/Version 2. Retrieved 11th January 2010 from http://www.significantproperties.org.uk/documents/wp22_significant_properties.pdf

[c] Pennock, M. (2006). Digital Preservation - Continued access to authentic digital assets. Retrieved on 11th January 2010 from http://www.jisc.ac.uk/media/documents/publications/digitalpreservationbp.pdf

[d] Wilson, A. (2008). Significant Properties of Digital Objects. Retrieved on 8th March 2010 from http://www.dpconline.org/events/significant-properties.html

[e] The National Archives (UK) (2006). Generic requirements for sustaining electronic information over time: Defining the characteristics for authentic records. Retrieved on 22nd February 2010 from http://www.nationalarchives.gov.uk/documents/generic_reqs1.pdf

[f] 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

[g] Hedstrom, M. and Lee, C.A. (2002). Significant properties of digital objects: definitions, applications, implications, Proceedings of the DLM-Forum 2002. Retrieved on 19th January 2010 from http://www.ils.unc.edu/callee/sigprops_dlm2002.pdf.

[h] Knight, G., Pennock M. (2009). Data Without Meaning: Establishing the Significant Properties of Digital Research. The International Journal of Digital Curation, Issue 1, Volume 4, 2009. Retrieved 11th February 2010 from http://www.ijdc.net/index.php/ijdc/article/viewFile/110/87

[i] Knight, G. (2009). Framework for the definition of significant properties. Inspect Framework Report, first version. Retrieved on 2nd March 2010 from http://www.significantproperties.org.uk/wp33-propertiesreport-v1.pdf

[j] 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

[k] Giaretta, D., Matthews, B., Bicarregui, J., Lambert, S, Guercio, M., Michetti, G and Sawyer, D. (2009). Significant Properties, Authenticity, Provenance, Representation Information And OAIS. Retrieved on 2nd March 2010 from http://www.cdlib.org/services/uc3/iPres/presentations/GiarettaSigProps.pdf

[l] DOPWG Shared Model (2009). Retrieved on 26th February 2010 from http://planets-project.eu/private/pages/wiki/index.php?title=DOPWG_Shared_Model&action=history

[m] 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

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

[o] Rothenberg, J., and Bikson, T. (1999). Carrying authentic, understandable and usable digital records through time: Report to the Dutch National Archives and Ministry of the Interior. Retrieved on 22nd February 2010 from http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf

[p] MacNeil, H., Wei, C., Duranti, L., et al. Authenticity task force report. Retrieved on 18th December 2009, from http://www.interpares.org/book/interpares_book_d_part1.pdf

[q] Retrieved on 18th January 2010 from http://www.webarchive.org.uk/wayback/archive/20050410120000/http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html.

[r] Retrieved on 18th January 2010 from http://www.si.umich.edu/CAMILEON/about/aboutcam.html.

[s] Retrieved on 18th January 2010 from http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=185&lang=en.

[t] Potter, M. (2002). Researching long term digital preservation approaches in the digital preservation testbed. RLG DigiNews, Vol.6(3). Retrieved on 23rd February 2010, from http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070511/viewer/file1521.html#feature2

[u] Digital Preservation Testbed. (2003). From digital volatility to digital permanence: Preserving text documents. Retrieved on 22nd February 2010, from http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=185

[v] Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System (OAIS) (2002). Retrieved on 29/1/2010 from http://public.ccsds.org/publications/archive/650x0b1.pdf

[w] Knight, G. (2009). Inspect Framework Report, final version. Retrieved on 2nd March 2010 from http://www.significantproperties.org.uk/inspect-framework.html

[x] Heslop, H., Davis, S. and Wilson, A. (2002) An Approach to the Preservation of Digital Records http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm2-888.pdf

[y] Rauch, C., Strodl, S., and Rauber, A. (2005). A framework for documenting the behaviour and functionality of digital objects and preservation strategies. Retrieved on 16th February 2010, from http://www.dpc.delos.info/private/output/DELOS_WP6_d641_final__vienna.pdf

[z] http://www.significantproperties.org.uk/index.html. Retrieved on 18/1/2010.

[aa] Grace, S., Knight, G., and Montague, L. (2009). Inspect Final Report. Retrieved on 2nd March 2010 from http://www.significantproperties.org.uk/inspect-finalreport.pdf

[bb] http://www.significantproperties.org.uk/testingreports.html. Retrieved on 18/1/2010.

[cc] Knight, G. (2010) Significant Properties Data Dictionary. Retrieved on 2nd March 2010 from http://www.significantproperties.org.uk/sigprop-dictionary.pdf

[dd] http://www.kcl.ac.uk/iss/cerch/projects/portfolio/spil.html Retrieved on 18/1/2010.

[ee] Stawowczyk Long, A. (2009). Long-term Preservation of Web Archives – Experimenting with Emulation and Migration Methodologies. Retrieved from on 29th January 2010 from http://www.netpreserve.org/publications/NLA_2009_IIPC_Report.pdf.

[ff] Dappert, A. and Farquhar, A. (2009). Implementing Metadata that Guides Digital Preservation Services. Retrieved on 2nd March 2010 from http://www.cdlib.org/services/uc3/iPres/presentations/Farquhar.pdf

[gg] Dappert, A. (2010) Obtaining and Relating Digital Object Properties in Digital Preservation. PLANETS deliverable PC2-D17.