Project Information			
Project Acronym	InSPECT		
Project Title	Investigating the Significant Properties of Electronic Content over Time		
Start Date	October 2006	**End Date**	December 2009
Lead Institution	King's College London		
Project Director	Stephen Grace		
Project Manager & contact details	Stephen Grace Centre for e-Research, King's College London stephen.grace@kcl.ac.uk		
Partner Institutions	The National Archives		
Project Web URL	www.significantproperties.org.uk		
Programme Name (and number)	Repositories and Preservation Programme 2007-2008		
Programme Manager	Neil Grindley		

Document Name			
Document Title	Final Report		
Reporting Period			
Author(s) & project role	Stephen Grace, Project Manager Gareth Knight, Technical Officer Lynne Montague, Project Officer		
Date	1 December 2009	**Filename**	
URL			
Access	☑ Project and JISC internal	☐ General dissemination	

Document History		
Version	**Date**	**Comments**
1.0	21 Dec 09	Draft final report

Table of Contents

Table of Contents 2
Acknowledgements 2
Executive Summary 3
Background 4
Aims and Objectives 4
Methodology 5
Implementation 14
    Audio 15
    Emails 16
    Raster images 18
    Structured Text 19
Outputs and Results 21
Outcomes 22
Conclusions 22
Implications 23
Recommendations 23
References 24
  Appendix A: Workflow for determining Significant Properties 27

Acknowledgements

This project was funded by JISC under its Repositories and Preservation programme. Neil Grindley the programme manager gave support and encouragement throughout the project's lifetime.

InSPECT was led by the Arts and Humanities Data Service (AHDS) Executive until its demise in March 2008, and then by the new Centre for e-Research (CeRch) at King's College London. CeRch was assisted by project partners at The National Archives (TNA), which undertook two of the four format case studies; our thanks go to Lynne Montague the TNA project officer, and Adrian Brown, Ian Hodges and Tim Gollins her successive managers during the project duration.

We also thank Andrew Wilson (formerly of AHDS, now of National Archives of Australia) for writing the Significant Properties Report, and for helpful feedback and insights received from Susan Thomas, Patricia Caplan, Angela Dappert and the PREMIS Editorial Board. CeRch project staff delivered workshops at Archiving 2008 (Bern, June 2008) and DCC 101 (Edinburgh, November 2008) events, and thanks go to Neil Grindley and Grant Young for collaborating on the Bern workshop.

Executive Summary

Significant Properties (SPs) have been in the digital preservation literature for a decade or more. There has been widespread agreement on the important role they play, but no consensus on exactly **what** is significant.

Investigating the Significant Properties of Electronic Content over Time (InSPECT) aimed to expand and articulate the concept of 'significant properties', determine sets of significant properties for a specified group of digital object types, evaluate methods for measuring these properties for a sample of relevant representation formats, investigate and test the mapping and comparison of these properties between different representation formats, and identify any issues requiring further research. Significant Properties are defined by InSPECT 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***

The definition was tested by identifying Significant Properties (SPs) in particular digital object types. The project chose audio, email, raster image and structured text objects as distinct classes of digital objects, and typically three formats for each type. After developing a framework for considering SPs, they were determined for each object type using a consistent methodology and extracted from sample files. Transformations (in this case through file format migration) were undertaken and SPs were extracted from the resultant objects. Comparisons were then made of the source and target objects to see how well the target retained the significant properties of the source.

The framework combined teleological and epistemological approaches in its analysis methodology, requiring the evaluator to determine the context of an object's creation (the purpose it was created for, how it was created, and so on) and the information necessary to communicate the intrinsic knowledge to a designated community. The framework then used a modified version of the Function-Behaviour-Structure (FBS) framework commonly used in engineering complex structures to analyse digital object types. Here, Function relates to the design intention or purpose that is performed; Behaviour covers the epistemological outcome derived from the function and structure that is obtained by a stakeholder; and Structure to the structural elements of the Object that enables the stakeholder to achieve the stated behaviour. In the first phase (Object Analysis), the evaluator analysed a representative sample of an object type, identified a set of functions and behaviours that may be achieved by the object, together with the properties that were necessary for their performance. Next in Stakeholder Analysis, the evaluator identified one or more stakeholders that have some relationship with the Information Object and analysed the functions that they wish to perform. Finally in the Reformulation phase, the evaluator performs preservation actions on files (here, by migrating file format) and compares the result with the original objects. The framework is available as a separate document on the Significant Properties website (www.significantproperties.org.uk).

Four testing reports detail the investigations into determining and identifying SPs in audio, email, raster images and structured texts. In each case, technical documentation was inspected to create a list of properties; each was then analysed as to its function and determined as significant (or not) if it afforded the accessibility, usability or meaning of the object. Sample files had their SPs extracted – typically using characterisation tools like JHOVE – and recorded, using a data dictionary created by project staff. Migration to different formats was performed and the SPs extracted from the new transformed versions, to compare with the originals. Each testing report gives a list of the properties deemed as significant for a given object type.

InSPECT defined Significant Properties in OAIS terms, and delineated them from Representation Information. A data dictionary was complied which could be used in PREMIS, or work with curation tools such as JHOVE and UDFR (as the successor to PRONOM). Finally, the report makes recommendations to continue the testing and evaluation work of InSPECT, while talking to related parties to maximise the uptake of SPs work.

Background

> ***"Preserving digital objects requires an understanding of the significant properties that need to survive in order to keep the object usable, meaningful, authentic and reliable"***
> Digital Preservation Europe, 2007, p27

Significant Properties (SPs) as a concept has been a focus of analysis and reference by several projects during the last decade, and it was identified as one of the ten challenges for future work by Digital Preservation Europe in its Research Roadmap (Digital Preservation Europe, 2007). The concept first emerged through work performed by the CEDARS project in the late 1990s to describe the elements of a digital object that should be maintained through preservation action (Cedars Project, 2002). It is built upon the underlying belief that it may be impractical, due to technical issues, cost or other factors to reproduce all elements of an object over time. For example, a new format may not support all features of the original or an emulator may introduce anomalies into the recreation. Therefore, selection criteria should be developed that enables a curator to determine the elements of an object that must be maintained and distinguish them from those that may be abandoned. Since its development, the concept has been described using several different synonyms (essence, essential characteristics, core features, properties of conceptual object and others) and been subject to different, although semi-compatible interpretations. In an early work package for the InSPECT project, Wilson (2007) reviewed related work and proposed a revised definition. Significant properties, Wilson states, refer to:

> ***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***

In later stages of the project, the InSPECT project team attempted to ally the interpretation of significant properties more closely with the OAIS Reference Model and define its relationship with Representation Information. This has resulted in the incorporation of the following reference to the OAIS Information Object in the definition.

> ***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.***

The concept of significant properties using either definition is thus associated with the notions of authenticity (that a record is what it purports to be) and integrity (that there has not been corruption at the bit-level or deliberate alteration at the semantic level that has caused the original meaning to be lost) (Bearman & Trant, 1998; Digital Preservation Testbed, 2003; Wilson, 2007). The requirements to maintain the authenticity and integrity of an object are useful in determining the properties that may be considered significant. However, as evidenced in the definition of significance and associated terms, the interpretations of authenticity and integrity may vary in different contexts.

Aims and Objectives

The InSPECT project had five objectives:

- expand and articulate the concept of 'significant properties'
- determine sets of significant properties for a specified group of digital object types
- evaluate methods for measuring these properties for a sample of relevant representation formats
- investigate and test the mapping and comparison of these properties between different representation formats
- identify issues which will require further research

Methodology

Literature review – significant properties

An early task of the project was to "explore the current versions, views, and visions" of significant properties by collecting earlier discussions of the issues, analysing them and "developing a clear, consistent articulation" of the concept (Wilson, 2007). Wilson noted that while benign neglect of physical objects might not precipitate their loss, in the digital domain this might result in the loss and/or destruction of the data. Preservation of digital objects needs to be undertaken at an early stage, and continuously thereafter, often by means of active interventions. Migration-based interventions (called 'transformations' in the OAIS Reference Model) result in the altering and possibly loss of data, so the question remains, What to preserve?

Wilson summarises the treatment of significant properties (SPs) in major projects such as CEDARS and CAMiLEON, and in digital preservation programmes at Cornell University, at national archives in Australia, Britain and America, and in the ongoing PLANETS project at the European level. Finally he articulates the concept of SPs in the following definition:

> ***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***

Properties are considered to exist in one of five categories: **content, context, appearance, structure** and **behaviour**. As the project developed, and sought to ally its work with the information model of OAIS, the definition was slightly altered (to say 'an Information Object' rather than 'digital objects') and the appearance category was amended to rendering, to take account of non-visual and non-static forms of digital object such as sound files.

Literature review – assessment frameworks

In the first year of the project, the project team reviewed a broad range of literature written on the topic to identify if any had developed frameworks that may be used for identification and evaluation (Wilson, 2007; Knight, 2008a). Notable work that was examined includes that written by Rothenberg & Bikson (1999), the CEDARS Project, the CAMiLEON project, the National Archives of Australia, RLG, Digital Preservation Testbed, DELOS, as well as more recent developments by the CASPAR, PLANETS and four JISC-funded Significant Properties projects. Many of these studies describe the process by which they identified the significant properties of various object types, describing formalised and semi-formal methods.

The more formal frameworks outline a set of activities that an evaluator should perform to obtain a list of properties that are necessary for preservation. In an early study of the topic, Rothenberg & Bikson developed a Needs Analysis model and identify four keys stages that should be followed to determine the elements of an Object that must and can be maintained. This was followed by the InterPARES1 project, which applied the principles of archival diplomatics to digital records as a method of determining its authenticity. The methodology is built upon the premise that many of the authenticity requirements of a record can only be determined by considering its intended purpose in an organizational setting and, as a result, cannot be easily understood by examining the record in isolation. However, the InterPARES Authenticity Task Force recognized that archival diplomatics as a methodology is tailored to requirements that may only be identified in a known organizational environment and, as a result, it is difficult to apply them to digital records that do not contain textual information, are dynamically produced, or published in a system where the context cannot be known (MacNeil et al, 2000). Finally, the conceptual Utility Analysis and Objective Tree (Rauch, Strodl & Rauber, 2005) was applied in the DELOS and PLANETS projects as a metric to test and evaluate digital preservation strategies. A key development is the specification of four main groups of characteristics: object, record, process and costs as a basis for evaluating different preservation options.

The frameworks developed by each project were useful in informing the development of the project, but were considered to be insufficient in isolation. In developing its methodology, the InSPECT project drew upon elements expressed in several projects. The needs analysis approach in the RAND Europe study (Rothenburg & Bikson, 1999) was recognised as being particularly useful for determining the requirements of stakeholders within specific environments. The expression of specific requirements as an Objective Tree by the PLANETS project was also seen as an interesting

approach, but at the time it was evaluated, it was recognised that a different approach may be necessary that placed a greater emphasis upon the role that properties performed for the recreation of the Information Object.

Analysis methodology

A methodology was developed on an ongoing basis during the lifetime of the project to determine the properties that had to be maintained in subsequent manifestations of a digital object. Experience gained by the InSPECT project team and the four related JISC-funded studies when examining the significant properties of different object types were noted and incorporated into subsequent versions of the analysis methodology. The project also drew upon feedback provided by email and through various workshops.

It was recognised early in the project that it is impractical to present a single, definitive interpretation of significance. Instead, it should adopt a methodology that enables the evaluator to identify the stakeholders that have some investment in the Information Object and define the subjective decision-making process that contributes to their evaluation of significance. Key to this approach was the recognition of three factors:

- Many stakeholders may be associated with an Object, e.g. creator, researcher
- The type of stakeholder associated with the Object may vary and change at different stages of its lifecycle. For example, a creator may use an Information Object in the initial stages of its lifecycle and subsequently make it available for use by other researchers.
- Each stakeholder may possess a distinct knowledge base and have specific needs for the task they wish to perform

By adopting a relativistic approach, an evaluator operating in a curatorial institution can determine the properties that they consider to be essential based upon their interpretation of acceptable loss. They may accept that some loss of functionality is necessary if it is to simplify the preservation process or, alternatively may adopt a risk adverse approach and adopt a preservation strategy that enables them to maintain all properties of the Information Object.

The methodology was underpinned by a joint teleological and epistemological approach. Teleology is the philosophical study of the design and purpose of an object. This conceptualises an author as a designer that creates an object as the result of an intellectual process to fulfil specific objectives or to address a problem. Epistemology is the branch of philosophy concerned with the meaning of knowledge and the process by which knowledge is acquired. In combination, the two philosophical branches require the evaluator to determine the context of the object's creation (the purpose it was created for, how it was created, and so on) and the information necessary to communicate the intrinsic knowledge to a designated community.

Assessment Framework

A formal framework was required to guide the process of identifying, analysing and recording the elements of an Information Object that are essential or beneficial to maintain over subsequent manifestations of a digital object. The framework sought to be rational in order to support the decision-making process and consistent in its application, while offering sufficient flexibility to meet the needs of the evaluator. The framework developed by the InSPECT project was used to determine sets of requirements for raster images, audio recordings, structured text and e-mail. By applying the framework, InSPECT evaluators were able to make an informed choice based upon consideration of associated factors, rather than haphazard decisions that cannot be supported at a later date.

The first version of the assessment framework (Knight, 2008b) outlined a set of activities that an evaluator could follow and offered a template to record values. To follow the instructions, an evaluator would start by examining the Information Object in its entirety (e.g. an email, raster image, audio recording) and progress through the sub-components until they can identify the technical properties that are necessary to recreate it. Once the evaluator has reached the property-level, they would work with one or more stakeholders in the Designated Community to analyse the acceptable boundaries necessary to achieve their stated objectives. This framework formed the basis of the four test activities, which sought to identify and measure SPs both before and after file-format migrations.

A revised version of the assessment framework (Knight, 2009c) utilises design methods to identify and evaluate the functions performed by an Object in its current manifestation and re-develop it to meet the needs of other stakeholders, such as a curator. To structure the analysis process, the project modified the Function-Behaviour-Structure (FBS) framework. The framework was initially developed by John Gero in 1990 to assist engineers and designers with the process of creating and re-engineering systems and has been revised and refined on several occasions since. Gero interprets the design process as an intentional, intellectual process by which a designer takes a set of designated functions and transforms them into a design description for an artefact structure that can fulfil these requirements. In the FBS model, the behaviour that is exhibited by an artefact (e.g. the operation of a motor engine) is a product of the functional purpose established by the designer and the physical structure which compose the artefact. The design method may also be used to reverse-engineer and re-design an existing product to perform one or more new functions (Takeda et al, n.d). For example, the structure of a bicycle may be re-designed to enable it to fit into a small space for transport when not in use.

Although the role of an engineer that is responsible for re-designing an artefact may initially appear to conflict with the curatorial duty to maintain the Information Object, they are not so dissimilar. As Rusbridge (2006) notes, it may not be necessary to be faithful to the original object in all respects. Many stakeholders may be willing to accept an Object that omits specific content or functionality. By considering the purpose of an object in conjunction with the stakeholder that uses it, a curator may identify the functions required by the creator in the early stages of the object lifecycle and evaluate if they continue to be necessary when used by a different community of users. This may result in the curator recognizing that some functionality is not required and adopting a preservation strategy that is faster, simpler to perform and less costly than alternative strategies that maintain all elements of the object.

In applying the FBS model to the analysis of digital objects, the InSPECT project has reinterpreted the base terminology used and the set of activities necessary to perform the requirements gathering process. The following definitions are used for the titular components of the FBS framework:

- *Function:* The design intention or purpose that is performed.
- *Behaviour:* The epistemological outcome derived from the function and structure that is obtained by the stakeholder. E.g. an interpretation of the meaning contained in the Content Information.
- *Structure:* The structural elements of the Object that enables the stakeholder to achieve the stated behaviour.

The interpretation of Behaviour differs from the definition initially provided by Gero, which examines the behaviour that an artefact exhibit. Instead, an interpretation has been made similar to that provided by Stalker (2002) when examining the lifecycle of an artefact.

Applying the FBS design method to the curation lifecycle

The design method serves as a useful metaphor for understanding the decisions made and activities performed through the lifecycle of a digital object. These similarities are illustrated in figure 1, which maps the FBS model work performed by Gero (1990) and Stalker (2002) onto the Digital Curation Centre (DCC) curation lifecycle model.

The initial FBS steps of Formulations and Synthesis may be mapped to the DCC Conceptualise stage, in which the creator conceives and plans the creation of data. During the creation process they may analyse and evaluate a prototype and revise it accordingly to fit their needs. At some stage when the creator is satisfied with the prototype it is ‘constructed’ into a final version of the digital object. The creation of a design document, as expressed in the FBS model may not be present with the creation process of many types of object. However, by re-interpreting the activity as the documentation of the final iteration of the object, design documentation may be used to refer to the creation of Representation Information.

A curatorial institution may also make decisions at a later data that involve some type of reformulation of the object. For example, a digital repository may perform corrective actions at Ingest to resolve issues found in the received object, e.g. modify the encoding structure to conform to a specification or standard) (Modification/repair). In Preservation Action/Transform a new encoding structure may be chosen to represent the set of behaviours of the original (Reformulation 1). Subsequently, it may be necessary to revise existing behaviours or define new ones that the preservation manifestation of the

Information Object should demonstrate (Reformulation 2). Finally, it may be recognized over time that the Designated Community use the object differently than that envisaged and intervention is necessary to offer new functionality or to re-design the structure (Reformulation 3). These various tasks are shown mapped to the stages of the DCC lifecycle in figure 1:

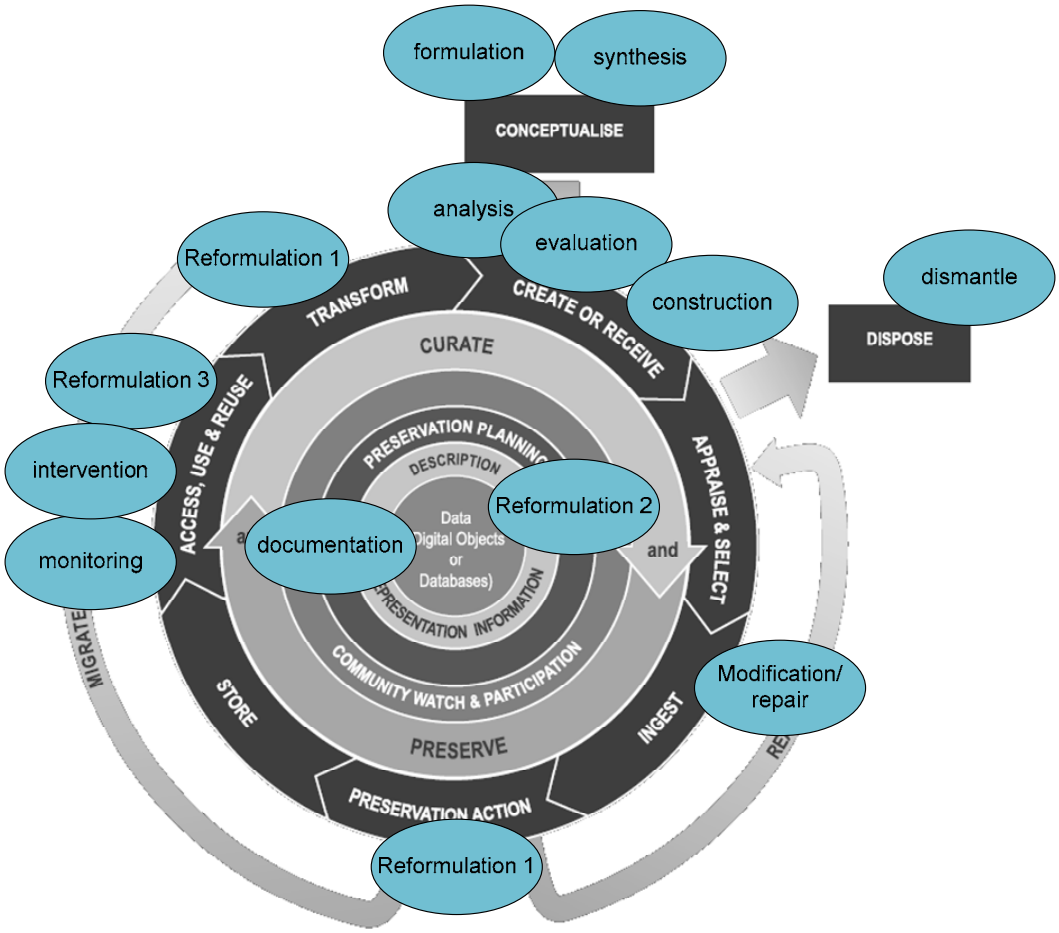


Figure 1: FBS design steps mapped onto the DCC curation lifecycle model (© DCC, adapted)

Requirements Analysis

The revised assessment framework developed by the InSPECT project utilised the FBS design method to identify the functions that have been defined by the creator of a digital object and evaluate if it is necessary to recreate them in subsequent manifestations of the Information Object. The workflow was composed of three sets of activities (each composed of several sub-tasks): Object Analysis, Stakeholder Analysis (both part of Requirements Analysis) and Reformulation. In Requirements Analysis the evaluator is required to gather information on the existing functionality that is provided by a digital object and understand the tasks that a stakeholder wishes to perform. The activities outlined in this stage may be classified in the 'Define Requirements' stage of the PLANETS preservation planning workflow (Rauch, Strodl & Rauber, 2005). The information gathered during the stage may be subsequently used to determine the properties of the Data Object that are significant for the recreation of an Information Object (e.g. as part of a DIP), that is accessed and used by stakeholders in a particular environment.

Requirements Analysis is composed of two streams of activity that each possesses a sequential set of sub-tasks to be performed:

- Object Analysis: The evaluator analyses a representative sample of an object type, identifies a set of functions and behaviours that may be achieved, and the properties that are necessary for their performance.
- Stakeholder Analysis: The evaluator identifies one or more stakeholders that have some relationship with the Information Object and analyse the functions that they wish to perform.

The two streams of activities may be performed in parallel or at different time periods. The latter is recommended, to enable the evaluator to gain a greater understanding of the functionality that is provided by the object type, which may be used as a basis for understanding the functions that may practically be provided to a stakeholder.

Object analysis

In the Object Analysis stage the evaluator selects an Object type for examination and develops their understanding of its technical composition and the purpose for which it may be used. The Object analysis workflow is composed of seven sub-tasks that may be performed sequentially (figure 2).

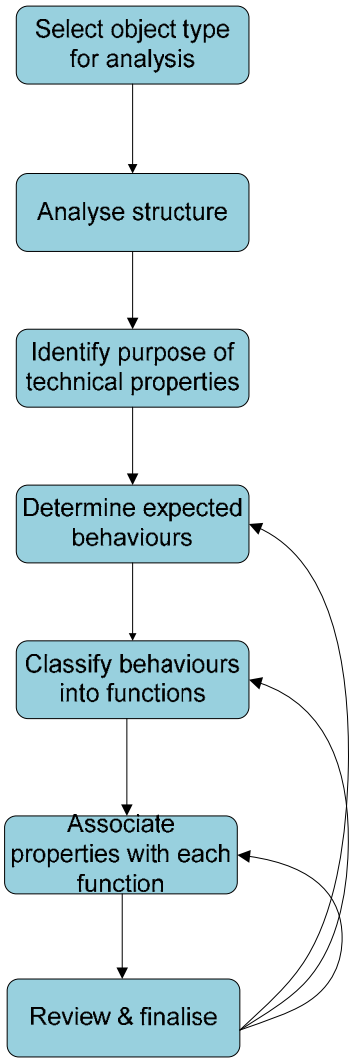


Figure 2: Workflow for the object analysis stage

Requirements

To perform this stage the evaluator must posses the following:

- A representative sample of objects for analysis
- Technical specifications or standards that describe the composition of the object
- Characterisation tools for analysis of the objects

1. Select object type for analysis

The first step is to select the object type to be analysed. The evaluator may choose to select a high-level object type (raster images, audio recordings, web pages, e-mail) or a sub-type that contains specific characteristics.

2. Analyse structure
Second, the evaluator should analyse the object and obtain a complete list of technical properties. The objective of the task is to develop an understanding of the type of technical properties and value types that are contained within the object type. Each property will be analysed in further detail in the following step. The task may be performed using several methods. A characterisation tool may be used to analyse and extract information on the technical composition of the object for storage as Representation Information. The evaluator may review technical specifications or standards associated with the object type and identify the technical information that is used to construct the Data Object.

3. Identify purpose of technical properties
Third, the evaluator should determine the purpose of each technical property that composes the object type/sub-type. The purpose of the activity is to determine the role that the property performs within the Data Object. If the technical property contributes to the recreation of the Information Object, it is considered useful to record the property value, for later evaluation after preservation action has been performed. When analysing technical properties that may be associated with the Information Object of raster images, audio recordings, presentational mark-up and e-mail, the InSPECT project used the following categories:

- *Content:* Information content within the Information Object. For example, text, still and moving images, audio, and other intellectual productions. Examples: duration, character count.
- *Context*: Any information that describes the environment in which the Content was created or that affect its intended meaning. Examples: Creator name, date of creation.
- Rendering: Any information that contributes to the re-creation of the performance. For example, font type, colour and size, bit depth.
- *Structure:* Information that describes the extrinsic or intrinsic relationship between two or more types of content, as required to reconstruct the performance. E.g. e-mail attachments.
- *Behaviour:* Properties that indicate the method in which content interacts with other stimuli. For example, hyperlinks.

The five terms may be used as high-level categories to distinguish properties of the Information Object from those of the Data Object. Each term may be further decomposed into sub-elements. E.g. Context: provenance, context: descriptive

4. Determine expected behaviours
Fourth, the evaluator should consider the different types of activities that a user – any type of user – may wish to perform. The list of activities should be recorded as a set of expected behaviours. At this stage of analysis, the evaluator should consider all uses of the Object type rather than those limited to a particular stakeholder. To produce a list of expected behaviours, the evaluator may draw upon their own experiences, the list of property descriptions performed in the previous step, formal standards and specifications, or other information sources. It may also be beneficial to consider the purpose for which the Information Object was utilized in its original creation environment.

5. Classify behaviours into functions
Fifth, the evaluator should classify the set of behaviours identified in the previous stage into a set of functions. The functions may be used as a basis for tailoring future manifestations of the Information Object to the needs of the stakeholder. In performing the activity, the evaluator may recognize that two or more behaviours may be associated with a single function. Alternatively, they may recognize that other behaviours emerge that should be recorded. For example, the recreation of the visual appearance of a message body may result in the recipient understanding contextual information that is implicit in the visual layout. InSPECT staff used content, context, rendering, structure and behaviour as the top-level functions.

6. Associate structure with each behaviour
The purpose of the sixth step is to link the technical properties that establish the structure of the Data Object with the set of expected behaviours. By performing the task, the evaluator may identify and list the subset of technical properties found within the Data Object that contribute to the recreation of the Information Object. The subset of Information Object properties may subsequently be measured and validated when performing format normalization, format migration, or other types of preservation action.

7. Review and finalise
Finally, the evaluator should review the information gathered in the previous steps and consider if any revisions should be made. Pertinent questions to be asked at this stage include:

- Are there any other behaviours that may be exhibited?
- Can any of the Functions identified be de-composed into two or more Functions that are more accurate?
- Are there any other properties that should be associated with a Function?

Stakeholder requirements analysis

The objective of the stakeholder requirements analysis is to identify the stakeholder categories that may have some relationship with the object type/sub-type and determine the set of functions that they require when using it. The set of functions associated with the stakeholder may be subsequently cross-matched with the object type functions and a list of significant properties developed for each context. When performing the analysis of raster images, structured text, digital audio recordings and e-mail, the InSPECT project team examined the requirements of a curatorial institution. However, the stakeholder requirements analysis may be performed on other stakeholders, such as a creator, researcher, as required by the evaluator.

To perform the analysis the evaluator must posses the following:

- A clear understanding of the relationship between the stakeholder that is the target of analysis and the object type (e.g. researcher, creator, curator)
- One or more people that have been identified as representatives of the stakeholder category

The workflow for the stakeholder requirements analysis is composed of seven steps. Although they are presented in a sequential order, the evaluator may choose to return to earlier steps at any time to revise it (figure 3).

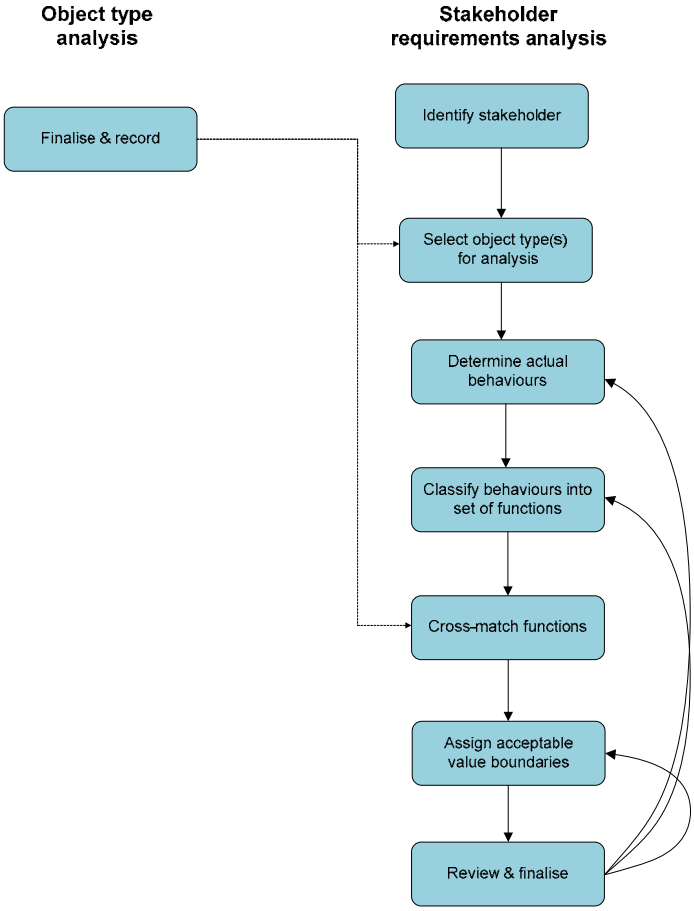


Figure 3: Workflow for the stakeholder requirements analysis stage

1. Identify stakeholders
The first step is to determine the stakeholders that will be the target of analysis and obtain their co-operation. A digital object may be associated with several types of stakeholder throughout its lifecycle,

each of which will have different aims and objectives. To identify potential stakeholders for analysis, the evaluator may wish to consult policies, procedures, or legal documents that establish the community that they are intended to serve. Several methods may be used to obtain details of the actions that a stakeholder will perform when using the object type. Examples may include the use of questionnaires, unstructured/semi-structured/structured interviews and/or observational study. The assessment of a large number of stakeholders may be time-consuming to perform. It may therefore be useful to establish boundaries upon the community that will be examined. The InSPECT project limited the stakeholder analysis to specific curatorial institutions that wish to maintain the authenticity and integrity of a digital object.

2. Select object type for analysis
This step is concerned with the selection of an appropriate object type that is used by a stakeholder. As noted in step 1 of the Object type analysis, an object may be classified into a single high-level type (e.g. raster images, audio recordings, web pages, e-mail) or a sub-type that contains specific characteristics. When interviewing the stakeholder, the evaluator may choose to examine the functions required of a specific high-level object type (e.g. raster images, email) or several different sub-types (raster images for scientific use).

3. Determine actual behaviours
The objective of the third step is to determine the activities that a specific category of stakeholder will likely perform when using the object. In the FBS model, the actions that occur in a real-world environment are referred to as ‘actual behaviours’ and are distinct from the ‘expected behaviours’ that were defined in step 4 of the object analysis. The actual behaviours may represent a subset of the expected behaviours that were identified (e.g. some users will need to view the Information Object, but may not wish to manipulate or edit the content), or may include new behaviours that were not previously recognized. To determine the behaviours exhibited by a stakeholder, the evaluator may wish to adopt an epistemological approach by considering the knowledge base that the stakeholder will draw upon and the method in which the content will be interpreted. The set of previously defined expected behaviours may be used to guide a semi-structured interview. Alternatively, they may ask the stakeholder to demonstrate how they would use the Information Object.

4. Classify behaviours into set of functions
The purpose of the fourth step is to classify the set of behaviours identified in the previous stage into a set of functions that subsequent manifestations of the Information Object should perform. A function refers to a specific design intention or purpose to be performed. In performing the activity, the evaluator may recognize that two or more behaviours should be associated with a single function. The evaluator should use the functional classification performed in step 5 of the object analysis as a classification guide.

5. Cross-match functions
The objective of the fifth step is to develop a list of the technical properties that are significant in performing the functions required by a stakeholder. To achieve the task, the set of functions identified for the stakeholder in step 4 should be cross-matched with the set of object type functions developed previously. The set of properties that are developed will enable the stakeholder to perform basic functions associated with the reproduction of the information content (e.g. view a still image, listen to an audio recording) and may include additional functions, as influenced by the type of activities that they wish to perform (e.g. verify authenticity).

6. Assign acceptable value boundaries
The objective of the sixth step is to determine the value boundaries for properties that are acceptable to the stakeholder. The acceptable value boundaries may be used to assess the success of preservation action when creating an Information Object for use by a particular stakeholder. In some circumstances it may be impossible or impractical to transfer all properties when re-formulating an object. However, it may be questioned if the degree of accuracy is expected or required. Although the Information Object is not reproduced exactly, it may be sufficiently accurate to perform the functions required by a stakeholder. It may be feasible to assign quality thresholds for a minimum and maximum value for properties that are beneficial but not essential to the understanding of the Information Object’s meaning or that allow some value variation without having a noticeable impact.

Four boundary constraints are anticipated:

- *Equality:* the property stored in the Record must be equal to one or more values stored in the metadata.
- *Minimum:* if a numeric measurement is used, minimum indicates the lowest numeric value that is allowed. The minimum and maximum measurement types must be used in combination.
- *Maximum:* if a numeric measurement is used, maximum indicates the highest number value that is allowed. For example, the highest sampling rate of an audio recording.
- *Range:* the value is one of several that are recorded.

An evaluator may take one of several approaches to obtaining information to populate the acceptable value boundary fields. They may observe the method in which the stakeholder uses the information and determine the quality level required to achieve a specific function. Alternatively, if the stakeholder is technically inclined, the evaluator may choose to explain the purpose of each property and ask them to consider the acceptable variation to perform the identified functions.

7. Review and finalise

Finally, the evaluator should review the information gathered in the previous steps and consider if any revisions should be made. Pertinent questions to be raised at this stage may include:

- Are there other behaviours that may be exhibited?
- Would it be more appropriate to de-compose one behaviour into two sub-behaviours to provide a more accurate description of the activity?
- Are there other functions that may be identified?

Once the evaluator is satisfied that they have completed the task, they may record the stakeholder functional requirements.

Reformulation

Reformulation in a design context refers to a process of re-developing an artefact to perform a revised set of functions or enable different behaviours. A digital object may be reformulated at several stages in a curation lifecycle. In the FBS-based data model developed for the project, functions required in manifestations of the OAIS AIP and DIP are associated with one or more stakeholders. By adopting this approach, a list of technical properties may be developed that are significant in the context that the Information Object is used. For example, the functions established by Curatorial institution A may specify that properties associated with the header and message text are significant, while the functions specified for Curatorial institution B may take a more risk-adverse approach by specifying that all properties associated with the header, text and visual appearance of the message are significant.

To illustrate the re-formulation process, figure 4 indicates a simple workflow in which an Information Object created for use by Stakeholder A is re-formulated into an Information Object for use by stakeholder B. To perform the process, it is necessary to identify the set of high-level functions required by the stakeholder (as identified in the Stakeholder requirements analysis) and cross-match it with the Object properties necessary to perform each function. The re-formulation specification may be subsequently used to evaluate the success of the conversion.

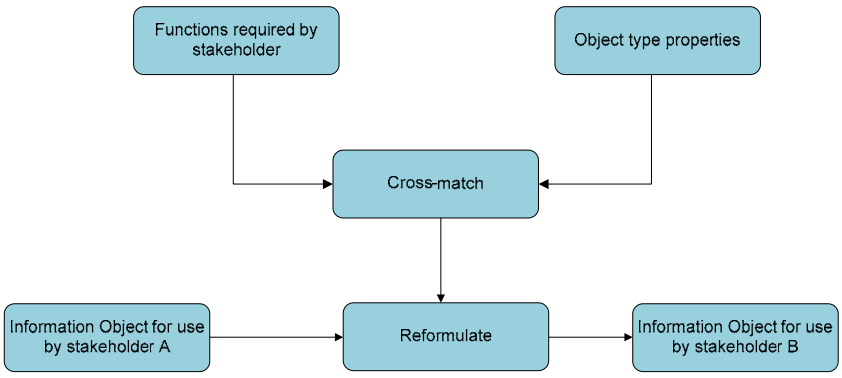


Figure 4: A simple re-formulation workflow

The workflow begins when a user requests an object conforming to a specific object type/sub-type (e.g. an email). If the user can be classified as one of the previously defined stakeholders (e.g. researcher, curator, tutor) a set of functions associated with the stakeholder may be obtained from an appropriate information source. If the user does not conform to an existing stakeholder classification or wishes to perform functions that cross-over two different stakeholders (e.g. genealogy) they may be given the option to select a set of behaviours associated with the object type, which is used as a basis for defining a new set of functions.

In the second stage of the workflow, the requested object is analysed to ascertain that it can perform the set of functions requested by the user. One approach may involve the comparison of a set of properties associated with selected functions to metadata generated by a characterisation tool (e.g. JHOVE). Finally, the re-formulated object is validated that it contains the requisite properties to perform each function. An invalid object may be deleted, at which point the workflow would return to an earlier stage. A valid object may be subsequently made available to the user for use. An indicative workflow showing all the stages of the assessment framework is at Appendix 2.

Significant Properties Data Dictionary

The InSPECT project also developed a data dictionary for recording the properties of an Information Object that are significant to a stakeholder. It was recognized that many institutions possess procedural lists and guidelines that indicate the properties associated with specific types of digital object. However, they are often stored in an unstructured form within an electronic document. The SPs data dictionary supports the management of digital objects by enabling institutions to:

- Record the components and properties of a digital object
- Evaluate the subjective value of each component and property that represent the information content.
- Assign quantitative and qualitative quality thresholds for the recreation of information content
- Evaluate the recreation of the Information Object by comparing properties stored in a source and destination Data Object.
- Obtain information regarding the ability to maintain each property when converting to a different encoding format, by querying a third-party service.

The use of a data dictionary in its various implementations (e.g. XML, database) will enable an institution to store object type information and stakeholder requirements to be stored in the same environment as the digital object objects themselves.

The data model developed for use in the Significant Properties Data Dictionary adopts the Information Object approach specified in OAIS and draws upon work undertaken by the PREMIS Working Group and National Archives Seamless Flow programme among others, to define a set of sub-units. In its final version developed for the InSPECT project, the data model defines four entities: an Information Object that represents a compound of many types of information (e.g. text, images, sounds, etc.) consisting of intellectual or technical components; Component that refers to a subset of the Information Object with which multiple properties are associated (e.g. a shape within a vector diagram; and Properties which represent the technical or semantic characteristics required to recreate the Object in part or in full. Each property will possess one or more values. Finally, an Agent must be associated with each entity to identify the stakeholder associated with the definition, classification and/or evaluate of a property within an Information Object. Future work on the Data Dictionary could re-develop the data model to better represent the FBS-derived methodology. Such work would likely draw upon related work being performed in the PLANETS project (Dappert & Farquhar, 2008: University at Cologne, 2008), and explore its reconciliation with the data model used in PREMIS.

Implementation

InSPECT applied the framework and methodology to four object types

- Audio
- Emails
- Raster images
- Structured text

Each of the four object types were analysed for their significant properties using the assessment template found at Appendix A and the first version of the assessment framework (Knight, 2008a). They were then characterised to record the properties found in sample files. Format migrations were

carried out and characterisation undertaken on the new versions to compare properties with those of the source files.

Audio

As a digital asset, an audio file may be considered a compound object that is able to encapsulate two ore more distinct types of information. The type of information and the method in which it is structured will often vary between audio recordings stored in different Representation Formats. Encoding algorithms used for the storage of audio data have different forms, which differ in the method that they store information. Some (like MIDI) are instruction sets to indicate the musical notes to be reproduced, others use musical notation, but the analysis of Representation Formats was limited to those encoded as continuous waveforms.

Several specifications were considered in compiling the set of significant properties, including the draft AES-X098B specification, Harvard University Library's DRS administrative metadata for digital audio schema, PBCore and the Library of Congress AudioMD schema, as well as the preservation guidance provided by the Indiana University Digital Library Sound Directions project, Council on Library and Information Resources & Library of Congress, Arts and Humanities Data Service and CDP Digital Audio Working Group.

The significant properties of audio text that need to be maintained are:

1. **Duration**
2. **Bit depth**
3. **Sample rate**
4. **Number of channels**
5. **Sound field**
6. **Sound map location for each channel**

If the audio recording contains BEXT-formatted metadata the following information should also be retained:

7. **Description**
8. **Originator**
9. **OriginatorReference**
10. **OriginationDate**
11. **OriginationTime**
12. **Coding History**
13. **Quality Report**
14. **Cue Sheet**

Representation formats
MPEG-1 Audio Layer 3 (MP3), Microsoft Waveform (.wav) and Broadcast Wave (BWF) were selected as audio representation formats for testing. The project team were unable to locate software that provided a detailed analysis for all of the analysed formats, so several tools were chosen for different aspects of the analysis. To assist with the analysis, the investigator merged the output of each tool into a single text record for analysis.

Experiment
The majority of experiments consisted of four distinct stages. The initial characterisation stage examined the source object and extracts appropriate representation information, to provide a baseline against which later characterisation activities were compared. Each source object was converted into several different file formats, including FLAC, OGG Vorbis, MP3, AIFF and MS Wave. The second characterisation stage examined the converted objects and extracts appropriate representation information. Finally, the results of the format conversion were evaluated through a combination of automated and manual comparison. A comparison was made between Representation Information extracted from the source and converted objects and an auditory assessment was made of each recording to identify any noticeable differences.

- Convert MS Wave to other formats using FFMPEG
- Convert MP3 to other formats using FFMPEG
- Convert BWF and MS Wave to other formats using SoX
- Convert MP3 to other formats using SoX
- Extract BEXT metadata from BWF using JHOVE

For the fifth experiment we sought to discover if JHOVE was able to extract all descriptive metadata embedded in the extension chunk of a Broacast Wave object. To establish a baseline for the

experiment, we selected four Broadcast Wave files and examined their content using three different tools. The analysis identified eight elements within the 'bext' chunk that provided provenance information about the recording, indicating its purpose and origin. The conversion of embedded metadata from Broadcast Wave to XML was a two-stage process of text extraction and formatting using JHOVE. Each of the BWF files were selected in turn and the results were saved in the JHOVE XML format. An analysis of each of the XML files indicated a common set of metadata elements had been extracted from the BWF files and stored in a consistent metadata scheme. JHOVE utilised the same element names for the majority of elements with the exception of 'Sample Count' which had been relabelled timeReference.

Conclusion
There were surprisingly few variations in the audio objects when the original and converted audio files were compared. In all of the experiments, the sample rate and number of channels remained the same. There was some minor differences in the duration reported for converted files which differed from that measured or obtained from the source file. However, an examination of the same audio file using different software appeared to confirm that the variation was caused by different handling of milliseconds by each tool.

The primary difference between original and converted objects was caused by the encoding algorithm and the capabilities of the container format to store different types of metadata. When converting from a lossless to lossy format or visa versa, it was possible to measure duration, sample rate and no. of channels. However, the variable bit depth of lossy encoding made it difficult to identify if quality loss had occurred. The analysis was limited by the availability of tools able to analyse each object at the required level of granularity and difference in handling the information.

Emails

An email is a compound object that may contain a diverse set of structured and semi-structured information. The specification for email messages is defined in several documents, collectively called the Multipurpose Internet Mail Extensions (MIME) that indicate a base set of information and an structure for its organization. An email is likely to consist of two mandatory Components, the Header and Body, with an optional third Component of Attachment(s). The visual recreation of an email message between two software applications differs considerably, as Figure 5 shows. This raises the question of what is the essence of the Performance that must be retained. What are the requirements of the email object to understand the information contained within or, to reverse the question, what can be removed before it becomes non-intelligible?

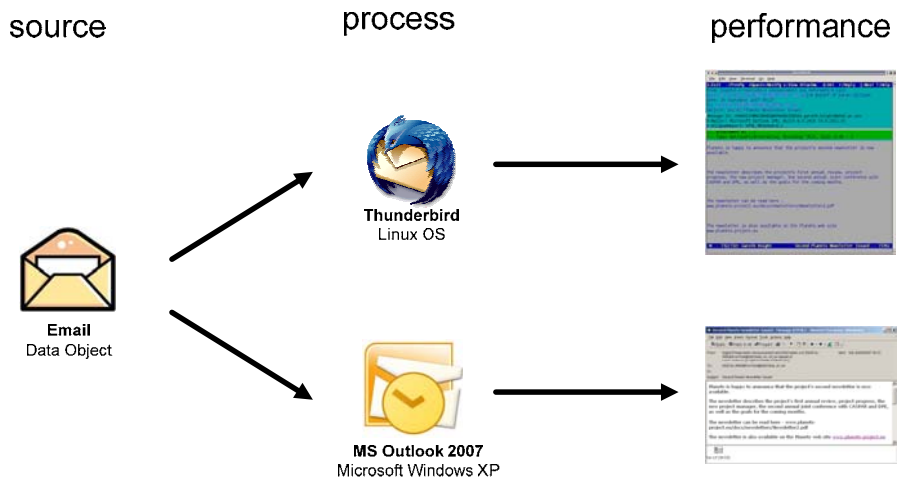


Figure 5 Application of the Performance Model to emails

Testing requirements that must be maintained
The authenticity requirements of an email share many similarities with other types of object, such as unstructured and structured text. However, the function that an email performs in a digital environment requires the consideration of additional requirements for authenticity. Each email may possess

complex inter-relationships with other emails (multiple emails that compose a thread) and intra-relationships between sub-Components (e.g. the relationship between an email and attached objects).

To demonstrate that the authenticity and integrity of an email object has been maintained, the project team recognised the need to determine two types of information, information created by an author that is intended for communication to a designated community and information that establishes the provenance of the email, indicating its purpose and the processes through which it was created and transmitted. The specification of a designated community – one or more intended recipients – is a characteristic that is unique within the object types being examined within the InSPECT project. Raster images, sound recordings and presentation markup typically identify the Creator and Publisher, but do not indicate the intended audience. However, emails share similarity with other purpose-specific object types, such as Learning Objects that are created for a specific audience.

The message body contains information that has been inserted by a Creator – a person or software – that is intended to be communicated to one or more recipients. The designation of the significant properties of the message body is potentially ambiguous, which can be argued from two perspectives. A minimal position argues that the significant properties of the message body should be limited to the visible characters of the message and, to a lesser extent the hidden formatting of the text, e.g. line breaks. Additional mark-up embedded within the message text, such as font size and colour are unnecessary and may be removed. The opposing, 'maximum' position is that each and every one of the mark-up elements contained within a message body is significant and should be maintained for accuracy. The mark-up of text may contain intrinsic information for communication. For example, an email may contain the statement, *"Please comment on the text highlighted in red in the following paragraphs*". Although the message text remains understandable, the statement indicates that text colour is used to communicate information which has been lost. InSPECT took both approaches into account in identifying significant properties.

A core set of significant properties is considered necessary to establish the authenticity and integrity of the email message:

1. **Local-part**
2. **Domain-part**
3. **Relationship**
4. **Subject**
5. **Trace-field**
6. **Message body with no mark-up**
7. **Attachments**

If message threads are considered important to preserve (rather than the individual messages) the **Message-ID** and **References** properties should also be considered significant.

Representation formats
Four representation formats were considered in the testing report: Microsoft Outlook Message (.msg), Microsoft Outlook Personal Folder (.pst); mbox and Maildir.

Experiment
The experiments consisted of four distinct stages. The first characterisation stage examined the source object and extracts appropriate representation information. The number of emails contained in the source object was counted through a combination of manual and automated verification. This provided a baseline against which later characterisation activities were compared. The source objects were then converted into several different file formats, including RDF-based XML, mbox and plain text. A second characterisation stage examined the converted objects and extracted appropriate representation information. The result of the format conversion was evaluated through a combination of automated and manual comparison. A comparison was made between information extracted from the source and converted object and a visual assessment was made of a sample of emails to identify noticeable differences.

- Convert Outlook PST to XML-based RDF using Aperture
- Convert Outlook PST to mbox using ReadPST
- Convert Outlook PST to text using Outport

Conclusions
As an object type that consists predominantly of semi-structured text, a presumption might be made that the audit and conversion of email objects would be a simple process. Email formats are in essence text-based formats that could encapsulate any type of information. However, the experiments outlined in this report demonstrate that email management continues to cause problems. Due to the constraints imposed by existing software tools, the ability to convert between formats and assess the success of format conversion is limited.

The characterisation stage is necessary to audit that emails have been exported correctly, in terms of the number of emails and the significant properties contained by each. Although email formats are based upon published standards, the encoding formats in which emails are stored on client machines are difficult to audit. Many of the software tools examined required an active installation of Outlook to perform an audit or required some form of initial conversion before they could analyse the data. This made it difficult to audit the source objects using automated tools, requiring manual checks to be performed. This is likely to be impractical for an institution responsible for managing a large number of emails.

The conversion tools tested by the project were extensive in the information that they could export and diverse in the formats that they supported. However, no single application provided extensive functionality. The tools performed format conversion based upon a standard setting, rather than taking into consideration the specific requirements (e.g. mark-up) of each email object. They were also unable to record the relationship between the message body and attachments, for instance.

Raster images

The ANSI/NISO Z39.87 data dictionary was used as the basis for the project team's analysis of the significant properties of raster images. For the purposes of the InSPECT project raster images were considered at the highest level i.e. properties which apply to all raster images rather than format- or technology-specific properties or properties that apply to specific sub-classes of raster images. Notably, the project team ruled out of scope colour space since colours can be rendered accurately by different colour spaces. Similarly colour profiles (which map colour values of devices such as cameras and monitors to device-independent colour spaces) as a subordinate Z39.87 element were considered out of scope. Both these elements, however, are critically important as Representation Information, since they help to reproduce the original environment of the digital image.

In summary, the significant properties of raster images that need to be maintained, within the scope and definition of the InSPECT project are:

1. **Image Width**
2. **Image Height**
3. **X Sampling Frequency**
4. **Y Sampling Frequency**
5. **Bits per sample**
6. **Samples per pixel**
7. **Extra samples**

Representation Formats
For this project, TIFF (Version 6.0), JPEG (Version 1.02) and GIF (Version 89a) were the formats chosen for testing as these were all supported by the JHOVE tool chosen to perform file characterisation. They are also all widely used in different settings as raster image representation formats.

Experiment
Three experiments each followed a three-stage process. Initial characterisation determined a set of properties pre-defined by the relevant JHOVE module, and gave a value for each of these properties where present. The property values obtained from this characterisation served as the basis for comparison with our images once they were migrated in the next stage of the experiments. Secondly each of the test objects was migrated twice, from its original format to each of the other test formats. Thirdly, once each image was migrated, each of the two new format versions was characterised using JHOVE and the output used as the basis for comparison with the original image to see how well properties were retained through migration.

- Convert GIF 89a to JPEG 1.02 and TIFF 6.0 using Photoshop
- Convert JPEG 1.02 to GIF 89a and TIFF 6.0 using Photoshop

- Convert TIFF 6.0 to GIF 89a and JPEG 1.02 using Photoshop

Visual assessment of converted images
Once the automated parts of the process were carried out, a visual assessment of the images was performed. Photoshop was used to open each image so that the evaluator could visually compare them. A visual inspection of the image files in Photoshop resulted in a few obvious differences being noted in the images.

Conclusions
There were surprisingly few variations in the characterised migrated images when compared with the original files. In all cases, the Image Width and Height were maintained through migrations. This, or rather the ratio of Width to Height, can be seen as one of the most significant aspects of maintaining the integrity of an image.

The main differences are the constraints of the source and target formats, their ability to handle metadata or the software used the conduct the experiments. For example, colour space is not a property designated as significant by this project but the colour space metadata produced by the experiments was included in the experiment results for interest. Whilst, in theory it doesn't matter what the colour space of a migrated image is, as long as the colours are represented correctly from the original image, the colour space will dictate the number of bits per sample i.e. the number of colours that are possible, and this is a significant property.

Nearly all migrations resulted in a change in colour space. However, it is difficult to tell whether the visual colour changes noted were down to the change in colour space, or the fact that the number of bits used to represent colour had also changed. It is also difficult to assess this effect this other than by viewing the different versions of an image and making a visual judgement. As stated above, in terms of visual assessments of the converted images, with the aim of pinpointing changes in the images due to changes in the significant properties, a photographic expert is really needed.

Due to the constraints of the representation formats, particularly GIF, in handling metadata, as mentioned above, the only metadata consistently recorded across all three formats were image width, image length, bits per sample and colour space (which is not deemed as significant for our purposes). Therefore the ability to measure and assess the success of a file migration in an automated way between these three formats was limited. However, it is suggested that the significant properties of image width, image length, and bits per sample which were recorded, are the most vital aspects in maintaining an authentically rendered image.

Structured Text

The key characteristic that distinguishes structured and unstructured text is the presence of markup that provides additional information about the interpretation of text. The central premise of the Performance model is the distinction between the raw, uninterpreted data, defined as the Source, and the interpretation of the data as a Performance. Although this is a useful metaphor, its application for structured text documents will vary, as distinguished by the content type and the rendering method. During the analysis it was recognized that, when applied to certain types of structured text (e.g. XML documents that do not possess associated instructions on the preferred method of recreation), the Performance Model metaphor is unhelpful unless a distinction between the Source and Performance can be made. Many types of structured text may be 'performed' using several different methods. To illustrate, an XML-encoded text may be presented to the user as an RSS feed, processed and converted to an audio stream, or represented in several XHTML-compliant web pages that contain different types of information.

Assessment of the significant properties of structured text was based primarily on the latest W3C HTML 4.01 specification. Many elements were considered significant "in certain circumstances". Body colour text (Text=[colour]) illustrates the reasoning behind this decision: it is an attribute that specifies the foreground colour for text on the page. Web Accessibility Initiative guidelines deprecate the use of colour alone to convey information for accessibility reasons so it should not be considered significant. Some authors, however, use colour artistically and others choose to use it to convey semantic meaning, e.g. using red to indicate a negative number. These elements have not been included in the following list, but it may need to be augmented by some of them in specific circumstances.

In summary, the significant properties of structured text that need to be maintained are:

1. **Title**
2. **Creator**
3. **Date**
4. **Keywords**
5. **Rights**
6. **Div**
7. **Span**
8. **Language**
9. **Paragraph**
10. **Line break**
11. **Headings**
12. **Emphasis**
13. **Bold**
14. **Italics**
15. **Underline**
16. **Strong emphasis**
17. **Strikethrough**
18. **Horizontal Rule**
19. **Inserted text**
20. **Deleted text**
21. **Samp**
22. **Cite**
23. **Defined Terms (DFN)**
24. **Code**
25. **Abbreviation**
26. **Acronym**
27. **Quotations**
28. **Subscript / Superscript**
29. **Address**
30. **Button**
31. **List Elements**
32. **Table Elements**
33. **Image**
34. **Link**
35. **Applet**
36. **Frame**
37. **Frameset**

Representation formats
HTML 3.2, HTML 4.1 and XHTML 1.0 were selected as the formats for testing as these were all supported by the JHOVE tool which was chosen to do the file characterisation.

Experiment
Three experiments each followed a three-stage process similar to that used for raster images. Initial characterisation determined a set of properties from JHOVE, and gave a value for each of these properties where present. The property values obtained from this characterisation served as the basis for comparison with web pages once they were migrated in the next stage of the experiments. Secondly each of the test objects was migrated from its original format to three or four other test formats, as explained below. Thirdly, once each object was migrated, the new format versions were characterised using JHOVE for comparison with the source version, to see how well properties were retained through migration.

- Convert HTML 3.2 to HTML 4.1 and XHTML 1.0 using Dreamweaver
- Convert HTML 4.01 to HTML 3.22 and XHTML using Dreamweaver
- Convert XHTML to HTML 4.01 strict and transitional using Dreamweaver

Conclusions
Only 10 of the 37 suggested significant properties are potentially recorded by JHOVE (that is, Primary Language, Other Languages, Metatags, Frames, Images, Citations, Defined Terms, Abbreviations, Entities, Unicode Entity Blocks). As the major characterisation tool available, it would be helpful for more of these to be identifiable, with values, within JHOVE. Also, four of the significant properties that could be identified by JHOVE were not represented in the test samples assembled. However this could be a limitation of JHOVE in that they were possibly just not identified.

Of the 37 significant properties specified by the project team date, creator, rights and keywords were included in the Metatags section of JHOVE where relevant. In addition, the languages, frames, images, citations, defined terms (DFN element) and abbreviations metadata were also regarded as significant if recorded. However, information under abbreviations, citations, defined terms or frames was not recorded for any of the sample test images. This highlights two drawbacks with the experiments carried out. In all cases, where significant properties information was recorded, it remained the same across migrations. In all cases, the size of the file changes between migrations: file size is not a significant property but it can be seen that even where this and the format changes, it does not mean that significant property information will not be migrated correctly.

Outputs and Results

The chief outputs of the project are

1. a report on the concept of Significant Properties
2. a methodology for the project
3. a framework for assessing SPs
4. a data dictionary
5. the four testing reports outlined in the methodology section above

In addition, project staff gave presentations, workshops and papers at various events, and had fruitful discussions with others as the project progressed which helped shape the work and relate it to other current activities.

The Significant Properties Report (Wilson, 2007) offered a comprehensive overview of discussion to date in the literature. It offers an explicit definition of significant properties to guide further discussion of the subject:

> "**the characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects."**

Having a clear definition, which clarifies the hitherto overlapping talk of properties, characteristics and essence, will help future discussions of Significant Properties.

Following feedback from an expert workshop in April 2008, the notion of scoring each property was abandoned. This technique appeared to offer a scientific measure which was, in fact, subjective and unsustainable across a large quantity of elements. Weighted scoring is employed successfully in the PLATO preservation planning tool, but it did not offer any value to InSPECT's testing activity.

Testing reports for each of the four object types followed the methodology developed by InSPECT as laid out in Knight 2008a, and found it appropriate to the task. The methodology was further elaborated after the testing was underway, to incorporate the FBS design method, but it was not possible in the time available to reconcile this with the testing undertaken; this should be addressed in future work. The results of the testing are reported under Implementation above for each object type. In each case, assessors were able to determine a minimum set of Significant Properties. Audio files had additional properties that were significant when BEXT metadata was present, and additional significant properties are required when preserving an email thread as distinct from individual messages. It would be useful if other object types – most notably, those in the sibling projects funded by JISC but completed before InSPECT – could use the same framework for testing SPs in e-learning objects, moving images, software and vector images as representative of the breadth of forms of digital objects. In this way, it may be possible over time to develop standard object 'profiles' for different types of digital objects.

Grace and Knight gave presentations and workshops on SPs and the InSPECT approach at various events; the workshops especially helped to clarify thinking.

- JISC/ British Library/ Digital Preservation Coalition workshop *What to preserve? Significant properties of digital objects* (London, 7 April 2008)
- Delos Summer School (Tirrenia, Italy, 9-13 June 2008)
- Tutorial at IS&T Archiving 2008 conference (Bern, Switzerland, 24 June 2008) jointly with Neil Grindley (JISC) and Grant Young (TASI)
- iPRES 2008, 5th International Conference on Preservation of Digital Objects (London, 29 September 2008) jointly with Maureen Pennock (UKOLN)
- DCC Digital Curation 101 workshop (Edinburgh, 7 October 2008)

Knight and Pennock's paper from iPRES 2008 was subsequently selected for publication in the International Journal of Digital Curation in 2009 (Vol 4 Part 1). Grace convened a panel debate on SPs with Sheila Anderson (King's College London) and Christopher A Lee (University of North Carolina) at DigCCurr 2009: digital curation practice, promise, and prospects – a symposium on education for digital curation held in Chapel Hill, North Carolina on 1-3 April 2009. Montague gave presentations on InSPECT's work at various PLANETS internal meetings, to keep the latter project informed of developments. Wherever possible, presentations are made available through the project website at www.significantproperties.org.uk along with formal outputs.

Outcomes

Expand and articulate the concept of 'significant properties'

Wilson's Significant Properties Report (2007) articulates the concept of significant properties and offers a definition that is precise and comprehensive. After the project work concluded a detailed paper by Dappert and Farquhar (2009) recommended that the term "significant characteristics" should be used in preference to "significant properties", and the revision of the OAIS Reference Model introduced the term "transformation information property". It would be useful to explore this alternative terminology, and definitions, with a wider group of those engaged with digital preservation in order to seek a consensus. The InSPECT project helps to delineate significant properties from more general concerns of Representation Information, by specifying that the object in the definition above is the Information Object as defined in the OAIS reference model.

Determine sets of significant properties for a specified group of digital object types

Significant properties were identified for all four object types. The raster image case studies helps to delineate SPs from Representation Information (Repinfo). Colour space, while an important piece of Repinfo for those needing to know how to recreate the image on a rendering device (display screen), is not significant if the digital object has been migrated from RGB to CMYK, for instance.

Evaluate methods for measuring these properties for a sample of relevant representation formats

Investigate and test the mapping and comparison of these properties between different representation formats

The framework for assessing significant properties, and the methodology for doing so, was successfully applied to all four object types considered by InSPECT. Assessors were able to identify significant properties in audio files, emails, raster images and structured texts. By performing multiple transformations (migrating from one file format to another) it was possible to compare different file formats to see whether any were better at preserving the properties identified as significant. Such actions may suggest preferred file formats for preservation planning approaches. Similarly, the Comparator tool of the PLANETS project "measures differences between digital objects before and after treatment", and it would be useful to see what agreement there is between the results of the two projects. It was notable, though, that the limited number of characterisation tools used were not always able to identify the significant properties in identified in analysis phases. Further development of characterisation tools is required.

The combination of technological and epistemological approaches was crucial to the project. Functional analysis allowed an assessor to consider the purpose of each property both as technical characteristics (Object analysis) and for a designated community (Stakeholder requirements analysis). This analysis can be revisited in the Reformulation stage. Designated Communities are central to OAIS thinking, and InSPECT ensures the needs of the identified users are taken into account when determining SPs. This requires digital curators to engage with Designated Communities in identifying and recording SPs; such a task may be difficult to undertake for reasons of time, proximity to users, etc. It will help if content creators (such as in academic research) help curators in the task of documenting SPs for digital objects they produce.

Identify issues which will require further research

The Implications and Recommendations sections suggest some areas where further work is required to validate the InSPECT approach, and to integrate it to related work.

Conclusions

Significant properties offer a way of articulating requirements for digital objects through their life, despite the changes that might be enacted on them or the changing audiences for them. The initial identification of SPs may be undertaken by a digital curator on assuming care for the objects, and seek to represent the intention of the creator (producer), but further assessments may be required as the requirements of the Designated Community change, or a new community becomes the focus. In this way, Significant Properties can act as a bridge across time to ensure the persistence of what is important in digital objects through any required transformations. Representation Information can document the knowledge needed by a Designated Community to interpret the encoded Data Object,

while Significant Properties can document the prior and post states of any transformations to assure users that the Information Object's meaning is preserved. The same point was eloquently expressed by Dappert and Farquhar (2009): "significant characteristics capture constraints on characteristics across time – before and after a preservation action".

Implications

Significant Properties needs to be incorporated into routine workflows of digital curators. Curators need access to appropriate tools such as JHOVE to identify properties for a given object type and format, and metadata to record such information with the object. More work is needed on characterisation and metadata extraction tools, to ensure they are appropriate for the task. The data dictionary needs to find a home in an appropriate existing metadata schema, such as PREMIS, or be developed as a new schema for storing structured information about Significant Properties. Standard object profiles could be developed through digital curators sharing their locally-developed profiles (expressed in some standard form) and agreeing on best practice.

Significant Properties In the Laboratory (SPIL) is a short follow-on project at CeRch funded by JISC to consider applying the InSPECT methodology to scientific data in the Randall Division of Biophysical Research at King's College London. This will offer a first step in showing how SPs can be incorporated into curation for a specific use case, in this case in various biophysical sub-disciplines of biomedical science, but more case studies applying InSPECT in different settings should be encouraged. These case studies would approach the task in determining SPs for different object types, file formats and institutional settings. Over time, a body of SPs 'profiles' could be developed which detail standard determinations of Significant Properties which different archives could adopt or adapt as appropriate.

Recommendations

Each of the four testing reports makes specific recommendation for further work, but overall InSPECT makes the following recommendations:

- Recommend that further experimentation is made using other representation formats and types of digital object
- Recommend that version 2 of the InSPECT framework is applied to the four JISC-funded SPs projects on vector images, moving images, software and e-learning objects and to the four object types undertaken by InSPECT evaluators to enable more granular analysis
- Recommend that further experimentation is made using other migration and characterisation tools
- Recommend that specific augmentations (to enable the capture of more significant properties) are made to JHOVE and other characterisation tools
- Recommend that larger sample sets are used to validate this and subsequent work. In particular, the Planets Testbed may be used as a space to perform experiments where appropriate formats are included, and the results shared with the Planets project.
- Recommend that use cases for significant properties in specific Designated Communities be developed. *Note that the Centre for e-Research will conduct such a task in the Significant Properties in the Laboratory (SPIL) project, recently funded by JISC as a Significant Properties Demonstrator*[1]*.*
- Recommend that formal consultation is held with appropriate bodies on using the draft Data Dictionary with PREMIS, UDFR (as the planned successor to PRONOM) and JHOVE

[1] http://www.kcl.ac.uk/iss/cerch/projects/portfolio/spil.html

References

ANSI/NISO Z39.87 - Data Dictionary - Technical Metadata for Digital Still Images. NISO, 2006
http://www.niso.org/kst/reports/standards?step=2&gid%3Austring%3Aiso-8859-1=&project_key%3Austring%3Aiso-8859-1=b897b0cf3e2ee526252d9f830207b3cc9f3b6c2c

Ashley, K. Davis, R & Pinsent, E. 2008. Significant Properties of e-Learning Objects (SPeLOs), v1.0.
http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/2008sigprops.aspx

Audio Engineering Society. 2009. AES-X098B: Administrative and structural metadata for audio objects [Draft Standard]. AES, 2009.

Bearman, D. & Trant, J. 1998. Authenticity of Digital Resources: Towards a Statement of Requirements in the Research Process, *D-Lib Magazine*, June 1998.
http://www.dlib.org/dlib/june98/06bearman.html

Brown, A. 2008. White Paper: Representation Information Registries
http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf

CAMiLEON project. n.d. Creative Archiving at Michigan & Leeds: Emulating the Old on the New.
http://www.si.umich.edu/CAMILEON/

Casey, M. and Gordon, B., 2007. Sound directions: best practices for audio preservation.
http://www.dlib.indiana.edu/projects/sounddirections/papersPresent/index.shtml

CASPAR Project. n.d. Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval. http://www.casparpreserves.eu/

CDP Digital Audio Working Group. 2006. Digital audio best practices v2.1. CDP, 2006.
http://www.bcr.org/dps/cdp/best/digital-audio-bp.pdf

Cedars Project. 2002. Cedars Guide To: Digital Collection Management.
http://www.leeds.ac.uk/cedars/guideto/collmanagement/

Council on Library and Information Resources and Library of Congress. 2006. Capturing analog sound for digital preservation. CLIR Pub 137.
http://www.clir.org/pubs/abstract/pub137abst.html

Coyne, M et al. 2007. The Significant Properties of Vector Images.
http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/2008sigprops.aspx

Coyne, M. & Stapleton, M. 2008. The Significant Properties of Moving Images.
http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/2008sigprops.aspx

Consultative Committee for Space Data Systems. 2002. Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book
http://public.ccsds.org/publications/archive/650x0b1.pdf

Dappert, A. & Farquhar, A. 2009. Significance is in the Eye of the Stakeholder. ECDL 2009, LNCS 5714, pp 297-308.
http://www.planets-project.eu/docs/papers/Dappert_Significant_Characteristics_ECDL2009.pdf

Digital Curation Centre. 2008. The DCC Curation Life Cycle Model.
http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf

Digital Preservation Europe (2007). Research roadmap.
http://www.digitalpreservationeurope.eu/publications/dpe_research_roadmap_D72.pdf

Digital Preservation Testbed. 2003. From digital volatility to digital permanence: Preserving text documents.
http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=185

Dorst, K. & Vermaas, P.E. 2005, 'John Gero's Function-Behaviour-Structure Model of Designing: A Critical Analysis.', Research In Engineering Design, vol. 16, no. 1-2, pp. 17-26.
http://www.springerlink.com/content/v684602542mp8070/

Fallis, D. 2006. Social epistemology and information science. ARIST 40, pp 475-519.

Gero J.S. 1990. Design Prototypes: A Knowledge Representation Schema for Design. AI Magazine 11(4): 26–36.
http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/854

Harvard University Library (2004). Administrative metadata for digital audio files.
http://preserve.harvard.edu/resources/audiometadata.pdf

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

Jones, S. Ross, S. & Ruusalepp, R. 2009. Data Audit Framework Methodology. Version 1.8
http://www.data-audit.eu/methodology.html

Knight, G. 2008a. Framework for the Definition of Significant Properties.
http://www.significantproperties.org.uk/outputs.html

Knight, G. 2008b.. Significant Properties Data Dictionary.
http://www.significantproperties.org.uk/outputs.html

Knight, G. 2008c.. Deciding factors: Issues that influence decision-making on significant properties
http://www.significantproperties.org.uk/outputs.html

Knight, G. 2009a. Significant Properties Testing Report: Digital Audio Recordings
http://www.significantproperties.org.uk/outputs.html

Knight, G. 2009b,. Significant Properties Testing Report: Electronic Mail
http://www.significantproperties.org.uk/outputs.html

Knight, G. 2009c. InSPECT framework for Significant Properties v2
http://www.significantproperties.org.uk/outputs.html

Knight, G and McHugh, J. 2005. Preservation handbook: digital audio. AHDS, 2005
http://ahds.ac.uk/preservation/audio-preservation-handbook.pdf

Lavoie, B.F. 2004. Technology Watch Report: The Open Archival Information System Reference Model: Introductory Guide.
http://www.dpconline.org/docs/lavoie_OAIS.pdf

Library of Congress, 2003. AudioMD data dictionary.
http://www.loc.gov/rr/mopic/avprot/DD_AMD.html

MacNeil, H. et al. (2000). Authenticity Task Force Report (2000).
http://www.interpares.org/display_file.cfm?doc=ip1_atf_report.pdf

Matthews, B. et al. 2008. The Significant Properties of Software: A Study, Version 5.7.
http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/2008sigprops.aspx

Montague, L. 2009a. Significant Properties Testing Report: Raster Images
http://www.significantproperties.org.uk/outputs.html

Montague, L. 2009b. Significant Properties Testing Report: Structured Text
http://www.significantproperties.org.uk/outputs.html

The National archives, n.d. Seamless Flow.
http://www.nationalarchives.gov.uk/electronicrecords/seamless_flow/default.htm

PBCore, 2006. User guide to PBCore Public Broadcasting Metadata Dictionary.
http://www.pbcore.org/PBCore/UserGuide.html

PLANETS Project. n.d. Preservation and Long-term Access through NETworked Services.
http://www.planets-project.eu/

PREMIS Editorial Committee (2008). PREMIS Data Dictionary for Preservation Metadata. Version 2.0. http://www.loc.gov/standards/premis/

Rauch, C. Strodl, S. & Rauber, A. 2005. Deliverable 6.4.1: A Framework for Documenting the Behaviour and Functionality of Digital Objects and Preservation Strategies.
http://www.dpc.delos.info/private/output/DELOS_WP6_d641_final__vienna.pdf

RLG. 2002. Trusted Digital Repositories: Attributes and Responsibilities.
http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf

Rothenberg, J. & Bikson, 1999. T. Carrying Authentic, Understandable and Usable Digital Records Through Time: Report To the Dutch National Archives And Ministry of the Interior. RAND Europe, 1999.
http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf

Rusbridge, C. 2006. Excuse Me... Some Digital Preservation Fallacies?. Ariadne Issue 46.
http://www.ariadne.ac.uk/issue46/rusbridge/intro.html

Stalker, R. 2002. A function-behaviour-structure framework for the lifecycle of an artefact.
http://www.ruthstalkerfirth.com/pdf/CC02.pdf

Takeda, H. et al. n.d.. Analysis of Design Processes by Function, Behavior and Structure – Preliminary Reports.
http://www-kasm.nii.ac.jp/papers/takeda/95/DPW-paper.pdf

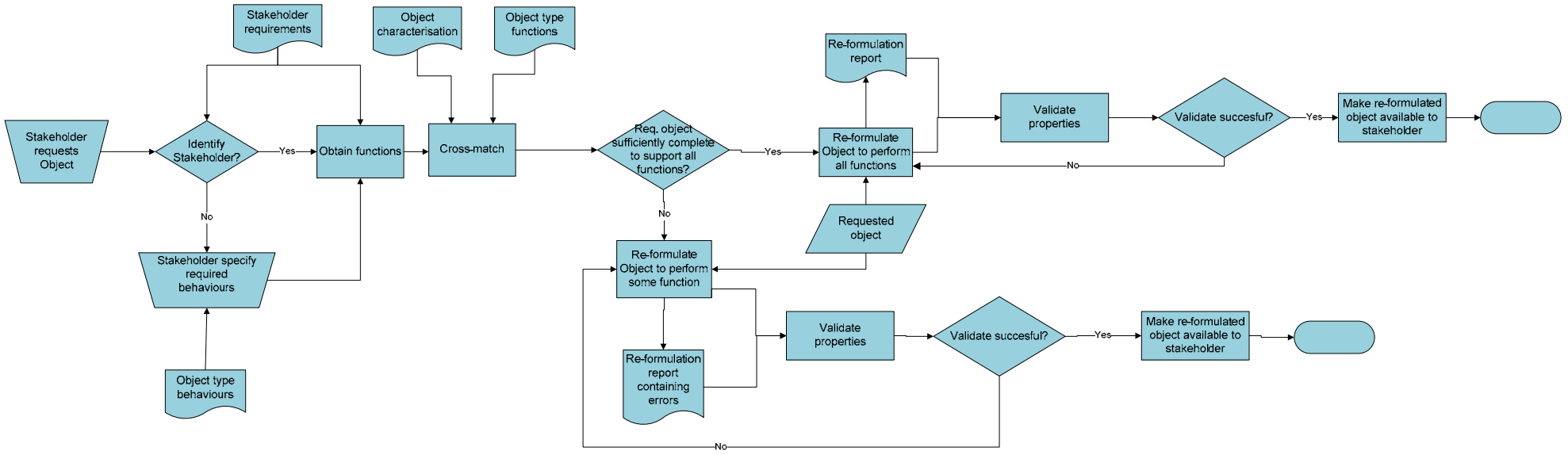
Thibodeau, K. (2002). Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years
http://www.clir.org/pubs/reports/pub107/thibodeau.html

Trustworthy Repositories Audit & Certification: Criteria and Checklist. OCLC/CRL/NARA, 2007
http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

University at Cologne (2008). Final XCDL Specification.
http://www.planets-project.eu/docs/reports/Planets_PC2-D7_FinalXCDLSpec_Ext.pdf

Wilson, A. 2007. Significant Properties Report.
http://www.significantproperties.org.uk/documents/wp22_significant_properties.pdf

Appendix A: Workflow for determining Significant Properties



This flowchart indicates the workflow to establish the set of properties that are significant in a particular context, using the InSPECT framework