Project Number	IST-2006-033789
Project Title	Planets
Sub-Project Title	Preservation Planning
Work Package Title	Preservation Policy and Strategy Models
Work Package Identifier	PP2
Deliverable Title	Report on policy and strategy models for libraries, archives and data centres
Deliverable Identifier	PP2-D2
Dissemination Level	PP
Deliverable Type	External Deliverable
Contractual Delivery Date	31 May 2008
Actual Delivery Date	24 June 2008
Author(s)	Angela Dappert, Bart Ballaux, Michaela Mayr, Sara van Bussel

Abstract

Abstract

Digital preservation activities can only succeed if they consider the strategy, policy, goals, and constraints of the institution that undertakes them. Furthermore, because organizations differ in many ways, a one-size-fits-all approach cannot be appropriate.

For digital preservation solutions to succeed, it is essential to go beyond the technical properties of the digital objects to be preserved, and to understand the cultural and institutional framework in which data, documents and records are created, managed, and preserved. Fortunately, organizations involved in digital preservation have created documents describing their policies, strategies, workflows, plans, and goals to provide guidance. They also have skilled staff who are aware of sometimes unwritten considerations.

We have analyzed preservation guiding documents and interviewed staff from libraries, archives, and data centres that are actively engaged in digital preservation. This paper introduces a conceptual model for expressing the core concepts and requirements that appear in preservation guiding documents. It defines a specific vocabulary that institutions can reuse for expressing their own policies and strategies. In addition to providing a conceptual framework, the model and vocabulary support automated preservation planning tools through an XML representation.

To perform the analysis, we used a combination of top-down and bottom-up methods. We examined the scientific literature to create a top-down model from first principles. To complement this, we analyzed actual preservation guiding documents for their content and interviewed decision makers to determine factors that influence their preservation choices.

This document represents the first iteration of this work. A second iteration is planned for May 2009.

Keyword List

Keywords	
Preservation policy	Preservation object
Preservation strategy	Preservation risk
Preservation planning	Preservation opportunity
Conceptual model	Preservation action
Machine interpretable model	Preservation requirement
Characteristic	

Contributors

Person	Role	Partner	Contribution
Angela Dappert	Work-package leader PP2	BL	Author
Bart Ballaux	PP2 partner	NANETH	Co-Author
Michaela Mayr	PP2 partner	ÖNB	Co-Author
Sara van Bussel	PP2 partner	KB	Co-Author

Document Approval

Person	Role	Partner
Adam Farquhar	Project Leader	BL
Hans Hofman	Sub-Project Leader	NANETH
Robert Sharpe	Scientific Board Member	Tessella

Distribution

Person	Role	Partner

Revision History

Issue	Author	Date	Description
V 0.1	Angela Dappert	5 June 2008	1st draft
V 0.2	Angela Dappert	24 June 2008	2nd draft after feedback from Adam Farquhar, Hans Hofman and Robert Sharpe

References

Ref.	Document	Date	Details and Version

Executive Summary

Digital preservation activities can only succeed if they consider the strategy, policy, goals, and constraints of the institution that undertakes them. Furthermore, because organizations differ in many ways, a one-size-fits-all approach cannot be appropriate.

For digital preservation solutions to succeed, it is essential to go beyond the technical properties of the digital objects to be preserved, and to understand the cultural and institutional framework in which data, documents and records are created, managed, and preserved. Fortunately, organizations involved in digital preservation have created documents describing their policies, strategies, workflows, plans, and goals to provide guidance. They also have skilled staff who are aware of sometimes unwritten considerations.

We have analyzed preservation guiding documents and interviewed staff from libraries, archives, and data centres that are actively engaged in digital preservation. This paper introduces a conceptual model for expressing the core concepts and requirements that appear in preservation guiding documents. It defines a specific vocabulary that institutions can reuse for expressing their own policies and strategies. In addition to providing a conceptual framework, the model and vocabulary support automated preservation planning tools through an XML representation.

To perform the analysis, we used a combination of top-down and bottom-up methods. We examined the scientific literature to create a top-down model from first principles. To complement this, we analyzed actual preservation guiding documents for their content and interviewed decision makers to determine factors that influence their preservation choices.

The resulting conceptual model presents a simple yet expressive representation of the preservation planning domain. It views preservation planning as a process that identifies and mitigates risks to current and future access to digital objects. It accommodates a full range of preservation planning processes such as monitoring, characterization, comparison of characteristics, and evaluation of candidate preservation actions. It allows processes to be associated with a full range of entities from institutions, and collections, down to byte-streams.

The vocabulary can be shared and exchanged by software applications. It also offers a starting point for creating individualized models for an institution; this holds true even if the institution does not require a machine-interpretable specification.

Key findings from the analysis of preservation guiding documents are:

- Data carrier refresh has become an urgent priority as institutions discover failures at rates well above earlier predictions.
- There is a lack of consensus on the use of digital preservation terms, the variety of preservation planning goals, and uncertainty as to how digital preservation should be implemented in practice.
- Preservation policy documents set a general framework for digital preservation, but do not provide specific practical guidance.
- Some existing preservation policies may not accurately reflect the institution’s actual preservation goals.
- Some institutions mandate a specific “technical preservation strategy,” such as migration, regardless of and sometimes in conflict with lower level technical requirements. This demonstrates the need to integrate institutional and data object considerations in the conceptual model.
- Non-technical aspects, such as the regulatory framework, need to be more detailed than they currently are to support automated preservation planning tools.
- Most current preservation polices specify preventive actions during ingest, including format normalization, format validation, error correction, as well as standards for both file formats and metadata.
- Most institutions currently hold fairly homogeneous digital collections that they characterize by data carrier types or file format. This simplifies the choice of tools that they use.

Table of Contents

1. INTRODUCTION 1

1.1 Goal 1

1.2 Scope 1
1.2.1 Applicable institutional types 1
1.2.2 Applicable business processes 1
1.2.3 Applicable documents 2
1.2.4 Applicable concepts 2
1.2.5 Applicable preservation actions 3
1.2.6 Applicable model 3
1.2.7 Structure of this document 3

2. TERMINOLOGY 4

3. METHODOLOGY 7

4. PRESERVATION GUIDING REQUIREMENTS AS SEEN IN LITERATURE ANALYSIS, DOCUMENT ANALYSIS AND INTERVIEWS 9

4.1 Introduction 9

4.2...... Literature analysis 9
4.2.1 What should a preservation policy or strategy document contain? 9
4.2.2 Reference models 13
4.2.3 Conclusions 13

4.3 Analysis of preservation guiding documents 13
4.3.1 Previous studies on current use of digital preservation policies in organisations 13
4.3.2 Current use of digital preservation policies in organisations 16
4.3.3 Conclusions 18

4.4 Interviews 19
4.4.1 Key findings 19

4.5 Example extraction of concepts and requirements for the conceptual model 22
4.5.1 Requirements base 23

4.6 Differences between institutional types 26

5. A CONCEPTUAL MODEL AND SPECIFIC VOCABULARY 29

5.1 Introduction 29

5.2 Use of the model and vocabulary 29
5.2.1 A worked example 29
5.2.2 Use 39

5.3 Modelling the context of preservation planning 39

5.4 Preservation Guiding Document 42

5.4.1 Stratml 42
5.4.2 Preservation guiding document 44

5.5 Preservation Object 47
5.5.1 Collection 49
5.5.2 Deliverable unit 49
5.5.3 Expression 49
5.5.4 Component 50
5.5.5 Manifestation 51
5.5.6 Manifestation file 51
5.5.7 Bytestream 51

5.6 Environment 53

5.7 Environment Component 54

5.8 Preservation Risk 57

5.9 Preservation Action and Preservation Workflow 59

5.10 Characteristic 64
5.10.1 Property 64
5.10.2 Value 68

5.11 Requirement 70
5.11.1 Risk specifying requirement 72
5.11.2 Preservation guiding and action defining requirement 72
5.11.3 Preservation process guiding requirement 73
5.11.4 Non preservation requirement 73

5.12 Shared data types 77

5.13 Using the data model for preservation planning 77

5.14 Conclusion 79

6. SUMMARY 81

7. APPENDICES 83

7.1 Modelling approach 83
7.1.1 UML class diagrams 83
7.1.2 OCL – The Object Constraint Language 84
7.1.3 XML 85

7.2 Reports on interviews 86
7.2.1 Interview help sheet 86
7.2.2 National library 1 89
7.2.3 Data centre 1 92
7.2.4 National archives 1 97

7.3 Vocabulary for specifying Properties 99

7.4 Citations 110

Table of Tables

Table 1 PP2 Terminology 6
Table 2 Overview of issues listed in Digital Preservation Policies, compared to TRAC's audit checklist, part A 18
Table 3 Preservation Action Classes 62
Table 4 Example property 68

Table of Figures

Figure 1 Overview over the PP2 deliverables 30
Figure 2 Preservation Planning Conceptual Model 40
Figure 3 An example snippet from http://xml.gov/stratml/BSAStratPlan.xml 43
Figure 4 Preservation Guiding Documents 44
Figure 5 Vocabulary for Preservation Object Types 47
Figure 6 Environment and Environment Component 54
Figure 7 Top-level vocabulary for Environment Component types 55
Figure 8 Vocabulary for peripheral types for Environment Components 55
Figure 9 Vocabulary for application and administrative software 56
Figure 10 Vocabulary for internal influences for Environment Components 56
Figure 11 Vocabulary for external influences for Environment Components 56
Figure 12 Vocabulary for Preservation Risk Types 58
Figure 13 Preservation Action and Preservation Workflow 59
Figure 14 Example Preservation Actions (simplified representation) 60
Figure 15 Vocabulary for Preservation Action Types 61
Figure 16 Preservation Actions 62
Figure 17 Property Types 65
Figure 18 Vocabulary for Requirement Types 72
Figure 19 Example Requirement 76
Figure 20 Propagating preservation planning results up the Preservation Object hierarchy. 78

1. Introduction

Digital preservation activities can only succeed if they consider the strategy, policy, goals, and constraints of the institution that undertakes them. Furthermore, because organizations differ in many ways, a one-size-fits-all approach cannot be appropriate.

For digital preservation solutions to succeed, it is essential to go beyond the technical properties of the digital objects to be preserved, and to understand the cultural and institutional framework in which data, documents and records are created, managed, and preserved. Fortunately, organizations involved in digital preservation have created documents describing their policies, strategies, workflows, plans, and goals to provide guidance. They also have skilled staff who are aware of sometimes unwritten considerations.

1.1 Goal

The overall aim of the work-package is to produce a conceptual model of organisational digital preservation policies and strategies (preservation guiding documents), that incorporates all relevant organisational characteristics and strategic directions, that covers the full life cycle of documents and records from the moment of creation, and that supports automated digital preservation planning. This is done by analysing how institutions – implicitly or explicitly – define and materialise their commitment and effort to digital preservation.

The reasons for doing this are

- To identify common features and systematic differences in the policies and strategies of different types of organisation.
- To enable parts of the preservation planning process and decision support to be based on organisational policy and strategy requirements.
- To add to the scientific understanding of preservation.

The concrete deliverables are

- A conceptual model which can be reused amongst related work-packages.
- A specific vocabulary for the concepts in the model from which institutions can pick in order to model their own policies and strategies.
- A machine interpretable model which can be used by preservation planning tools.

This is emerging work and this document represents an initial model. We will modify and improve it over the coming year in response to integration efforts with related work, and as the Planets project tries to exploit the ideas in practice. The Methodology section explains our past and intended future approaches. An improved release of this work is planned for May 2009.

1.2 Scope

1.2.1 Applicable institutional types

The model is based on investigations of national libraries, archives, and data centres. In the second iteration of the model, the study may be extended to include universities and business sectors with strong retention needs (e.g., Pharmaceutical, Aerospace).

1.2.2 Applicable business processes

Examination of existing preservation guiding documents shows that they span the whole range of repository and preservation tasks, ranging from collection development, pre-ingest, ingest, data storage and housing, compression, data management, access and security, to administration functions. This model is restricted to supporting the preservation planning sub-tasks which are directed at diagnostic treatment and wellness of digital objects. It does not encompass other preservation aspects such as collection development and security, and may have reduced applicability for purposes such as normalisation and validation during ingest, which might be considered pro-active preservation actions.

1.2.3 Applicable documents

The analysis summarised in this report is based on the content of preservation guiding documents. This term covers documents, such as policy, strategy, or business documents, as well as applicable legislation, guidelines, rules, or even a choice of temporary runtime parameters during a preservation action. The term "document" should be understood generously to possibly include oral representations, as well as written representations in databases, source code, web sites, etc.

They specify *Requirements*, which are constraints or rules that make the institution's values or constraints explicit and influence the preservation planning process. The term goes beyond and refines the notion of "organisational policy and strategy" documents that were originally foreseen as basis for the analysis.

Preservation guiding documents are a subset of institutional documents which

- may have any institutional scope (corporate, departmental, project related, etc.),
- may have any business focus (policy, strategy, mission, process, etc.),
- are relevant to the business process of preservation planning and form an input to the preservation planning process. Preservation plans are the output of a preservation planning process and are not considered preservation guiding documents as used in this report.

Concepts that are found in our model may be found in any of the documents in this space. We are not trying to prescribe to an institution which concepts should be implemented in which sort of document. This has to remain a personal choice of the institution.

1.2.4 Applicable concepts

This report models **organisational** preservation guiding documents.

The term "organisational" means that the model should encompass all organisational aspects (legislative, financial, etc) which apply to domain objects.

The vocabulary does **not** explicitly list traditional descriptive metadata for digital objects, since they are described in great depth elsewhere[1,2,3, etc.] Rather, it provides for an extension point. The model might refer to descriptive metadata, however, in order to express a condition in a requirement. "Publisher", for example, is a typical descriptive metadata element, which might be taken from the MODS metadata framework[3]. A requirement might use this element. Equally, institutions can write their own requirements referring to their own metadata schema of choice. An example requirement might be:

> "If the publisher is Elsevier then normalise the Bytestream using the"Elsevier_Normaliser2.0" tool.
>
> The machine-interpretable representation of this requirement might use the MODS concept "publisher":
>
> If MODS:publisher = "Elsevier" Then PreservationActionTool= "Elsevier_Normaliser2.0."

The term "organisational" does not mean that the model is limited to concepts which model only the organisation as a whole, but rather we include concepts that describe the parts of the organisation at any level, such as dynamic and static collections, deliverable units, expressions, manifestations, components, or files[4], if they affect the preservation planning process and would be expected to be expressed in preservation guiding documents. It is, for example, necessary to refer to characteristics at a lower level to represent requirements at a higher level. For example, in order to specify "collections which contains files that exceed 1 GB", you need to be able to specify the file property "file size".

[1] Library of Congress, Network Development and MARC Standards Office, http://www.loc.gov/marc/

[2] The Dublin Core Metadata Initiative, http://dublincore.org/

[3] MODS Metadata Object Description Schema. http://www.loc.gov/standards/mods/

[4] Definitions may be found in the Terminology section of this report. See [Core] for a motivation of these concepts.

Even though they are not the focus, the technical aspects of a digital preservation policy or strategy, as well as the state of technology on the basis of which high level constraints can be derived, need to be part of the research scope of this work. Some institutions appear to mandate a particular “technical preservation strategy” (migration, for example) at the preservation policy level, regardless of the lower level technical requirements. This demonstrates the need to integrate institutional and data object considerations in the conceptual model.

1.2.5 Applicable preservation actions

While Planets focuses on preservation actions related to software (e.g. migration, emulation, file repair, etc.), and does not address hardware related preservation actions (e.g. data carrier replacement or hardware replacement/reconstruction/repair), this model is general enough to support all kinds of preservation actions.

1.2.6 Applicable model

Since we are modelling preservation guiding documents, we restrict ourselves to creating a structural model of the domain. Behavioural and interactive models of the preservation planning process are created in other work-packages within the preservation planning sub-project.

1.2.7 Structure of this document

The introductory part of this document describes the basis of our work, the goals we wanted to achieve and the scope.

Sections 2 and 3 describe the terminology and the methodology used.

Section 4 describes our work modelling preservation guiding documents from definitions and domain concepts found in the literature, in actual preservation policy and strategy documents, from interviews with decision makers, and from first principles.

Section 5 proposes a conceptual model and vocabulary for preservation guiding documents. It motivates the results with a worked example that gives an overview of how the model and vocabulary in this report can be used. Following it, each section introduces a new concept with its relationships, followed by a description of its vocabulary. Our model is depicted using UML class diagrams.

The appendices include background on our modelling approaches, vocabulary for properties, detailed reports on some interviews and a selection of resources.

2. Terminology

Some key terms are defined here. Currently many of these terms are used inconsistently across various preservation research efforts. These definitions will be adapted as this work will be further co-ordinated with other efforts.

The source of the definition is given in square brackets. Terms marked as [Planets] have been taken from the Planets Wiki in May 2008. Terms marked as [Core] have been taken from the Planets Core Conceptual Data Model, distributed by email (Robert Sharpe), 01/04/2008.

Many examples for and explanations of the terms are contained in the section on the conceptual model.

Term	Definition
Preservation	[ALA[5]] Digital preservation combines policies, strategies and actions that ensure access to digital content over time. For a more detailed definition please follow the link in the footnote.
Preservation policy	[PP2 based on InterPARES2[6]] A formal statement of direction or guidance as to how an organization will carry out its preservation mandate, functions or activities, motivated by determined interests or programs.
Preservation strategy	[Planets] The strategy is a procedure of preservation actions to preserve a collection of digital objects. It treats only technical aspects. The preservation strategy thus contains a detailed description of the preservation action(s) to be taken, including used hardware and software, parameter settings for used tools and actions, input and output formats, and available metadata about the action(s). In a preservation strategy, different tools and parameter settings can be defined for different file formats. Appropriate characterization tools allow even different tools and parameter setting for the same file format with different characteristics.
Preservation guiding document	[PP2] Documents, such as policy, strategy, or business documents, as well as applicable legislation, guidelines, rules, or even a choice of temporary runtime parameters during a preservation action[7]. They specify *Requirements*, which are constraints or rules that make the institution's values or constraints explicit and influence the preservation planning process.
Preservation plan	[Planets] A preservation plan defines a series of preservation actions to be taken by a responsible institution due to an identified risk for a given set of digital objects or records (called collection). The Preservation Plan takes into account the preservation policies, legal obligations, organisational and technical constraints, user requirements and preservation goals and describes the preservation context, the evaluated preservation strategies and the resulting decision for one strategy, including the reasoning for the decision.

[5] Association for Library Collections & Technical Services of the American Library Association, Definitions of Digital Preservation http://www.ala.org/ala/alcts/newslinks/digipres/PARSdigdef0408.pdf

[6] See the InterPARES2 glossary at: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_glossary.pdf&CFID=243105&CFTOKEN=70677126, p. 20 (accessed: 23 May 2008). A similar definition can be found in R. Pearce-Moses, *A glossary of archival & records terminology*. Chicago, 2005, p. 300: "An official expression of principles that direct an organization's operations."

[7] The term "document" should be understood generously to possibly include oral representations, as well as written representations in databases, source code, web sites, etc..

Term	Definition
	It also specifies a series of steps or actions (called preservation action plan) along with responsibilities and rules and conditions for execution on the collection. Provided that the actions and their deployment as well as the technical environment allow it, this action plan is an executable workflow definition.
Characteristic	[PP2] A characteristic of a preservation object is the concrete value which this preservation object has for an abstract property in a defined context (a concrete property/value pair). In the model it is the characteristic of an environment component which belongs to a preservation object or a preservation action.
Property	[PP2] An abstract attribute, trait or peculiarity suitable for describing an environment component.
Value	[PP2] Every characteristic has a value which can either be assigned or be inherent in the object. The value can be looked up if it is stored explicitly or measured with an associated measuring tool, or deducted with a given logic if it is implicit in the object.
Significant property	[Andrew Wilson, National Archives of Australia] 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.
Preservation requirement	[PP2] A constraint which limits the space of allowable preservation planning activities. It is expressed through one or more property/value constraint specifications on environment component types. They are limited to specified preservation object types or preservation action types, and may include pre- or post-conditions.
Preservation risk	[PP2] A preservation risk arises when a characteristic of an environment component of a preservation object conflicts with the institution’s risk specifying requirements.
Preservation action	[PP2] The execution of an action to ensure the continued accessibility of a digital object across time and changing environments and the preservation of its critical significant properties that transforms the digital object itself, the environment required to support access to the object, or a combination thereof.
Preservation workflow	[PP2] A preservation workflow connects preservation actions together and may include conditional branches and other control-flow constructs. Planets uses the Business Process Execution Language (BPEL) to describe workflows.
Preservation Object	[PP2] Any object that is directly or indirectly at risk and needs to be preserved.
Institution	[PP2] The institution whose preservation guiding document is being modelled.
Collection	[PP2] A grouping of deliverable units to be processed or kept together.
Deliverable Unit	[PP2] A deliverable unit is a distinct intellectual or artistic creation.
Expression	[PP2] An expression is the specific intellectual or artistic form that a deliverable unit takes as it is “realized”. It is, however, a conceptual, not a physical realization.
Component	[PP2] A part of the whole of an expression (or of a deliverable unit, if expressions are omitted) for which values for characteristics can be measured.

Term	Definition
Manifestation	[PP2] The physical embodiment of an expression (or of a deliverable unit, if expressions are omitted) or component.
Bytestream	[Core] An ordered sequence of bytes.
Environment	[PP2] The set of factors which constrain a preservation object and that are necessary to interpret it.
Environment Component	[PP2] A factor which constrains a preservation object and that is necessary to interpret it.

Table 1 PP2 Terminology

3. Methodology

Our approach to date was based on a combination of methodologies.

- Top-down approaches:
  - Create a model from first principles.
  - Analyze the literature for abstract definitions of preservation policies and preservation strategies.
- Bottom-up approaches:
  - Analyze actual preservation policy and strategy documents for their content.
  - Interview decision makers to determine factors that influence their preservation decisions.
  - Analyse the requirements base.

In the two top-down approaches we investigated what the scope, context, and tasks represented in preservation guiding documents should be, and what concepts should be present to support these tasks. We created a preliminary model from first principles, and expressed it in UML.

For the literature analysis, we compiled a list of relevant literature investigating abstract definitions of preservation policies and preservation strategies, reviewed a representative subset of publications, and extracted concepts which were considered relevant for preservation guiding documents. We used the results to validate and improve the preliminary model.

In the bottom-up approaches, we compiled a list of the relevant sources from potential institutions drawn from various institution types, reviewed documents and interviewed decision-makers of a representative subset of the sources to determine the factors that influence their preservation decisions, extracted concepts and example requirements from the information gained, refined our preliminary model with the newly found concepts, and compiled a list of example requirements found.

We then analysed the requirements found to see if they could be expressed with the concepts and vocabulary in our model. We used the insights gained to improve on the model and vocabulary. In addition we analysed the requirements for commonalities and categorized them depending on their intended use and their dependencies on model concepts. We also analysed the requirements' complexity, investigated candidate modelling languages with sufficient expressiveness, and started an effort to represent some sample requirements in the chosen modelling language (OCL) using the model concepts and vocabulary.

Future approaches for refining the model in the next iteration of the work package include

- Continued clean-up and expansion of the vocabulary
- Co-ordination activities
- Case studies
- Design a corresponding appropriate machine-interpretable model (e.g. XML schema for requirements)

Coordination activities will discuss potential issues and results with other work packages that are dependent or have a close relationship with these activities. For example, we will assist in the use of the model in preservation planning tools, and continue our effort to integrate with the Core Planets model[8].

We will use the co-ordination activities to validate the model's concepts and vocabulary. In order to align the model with other work packages, we will make changes to our model to accommodate other work packages and influence changes in other work packages if this seems helpful.

Optionally, we will also assist in the deployment of the model for real world case studies to evaluate the usefulness of the machine interpretable model and the vocabulary, including sample requirements.

[8] Planets Core Conceptual Data Model, distributed by email (Robert Sharpe), 01/04/2008.

For a brief explanation of our modelling tools, UML, OCL, and XML, please refer to the appendix 7.1.

4. Preservation guiding requirements as seen in literature analysis, document analysis and interviews

4.1 Introduction

In a top-down approach, we studied how other research efforts are defining important conceptual elements for an ideal preservation policy or strategy, and found that different work has looked at the domain at different levels.

In a complementary bottom-up approach, we analysed existing policy documents to determine which conceptual elements have actually been used. We again found that the organisations emphasized very different aspects in their policies. It became apparent that there is not yet a shared understanding of which concepts a preservation policy document should contain. It appears also, that the lack of understanding of how a preservation policy might practically be used has so far resulted in high-level non-specific policy documents.

In addition, we interviewed persons involved in digital preservation at several institutions in order to discover new concepts and requirements present in their practice, but not mentioned in preservation guiding documents.

We extracted and combine the described concepts into a formal conceptual model and added our own concepts where gaps became apparent.

This section describes the general results and impressions from those approaches. Section 4.2 describes the literature analysis, Section 4.3 describes the document analysis, and Section 4.4 describes the interviews. Section 4.5 illustrates how we used the results from the research phase to extract concepts, vocabulary and requirements for our model. Section 4.6 discusses the differences which we found between the various institutional styles.

Work done as described in Sections 4.2 and 4.3 is considered as closed. The added value of scrutinising more literature or preservation guiding documents appears minimal. If necessary, additional interviews will be conducted.

4.2 Literature analysis

Digital preservation literature of the past decade has explored what digital preservation policies and strategies should contain and why it is important to have these documents.

Reasons identified as important for creating institutional policies are (ERPANET research, 2003):

- Planning coherent digital preservation programs
- Ensuring accountability
- Allocation of funds
- Ensuring digital materials are available for current and future use
- Defining significant properties that need to be preserved for particular types of resources
- Providing a comprehensive statement on digital preservation
- Providing security measures

4.2.1 What should a preservation policy or strategy document contain?

There is no consistent distinction drawn between what constitutes a preservation ‘policy’ versus a ‘strategy’. The terms are used variously and the delineation between them varies in different institutions. We have introduced a more general term, preserving guiding documents, to cover policies, strategies, and a variety of other documents that give guidance to preservation planning and other key preservation processes.

ERPANET

ERPANET was a European Commission funded research project (2202 to 2004). One of the main lines of research was an analysis of existing digital preservation policies and their implementation in several economic sectors and organisations. The Digital Preservation Policy Tool – "erpaguidance"[9] gives an overview of what should be included in a digital preservation policy, and provides a list of existing digital preservation policies at the moment of publication, i.e. 2003. Although the document does not include a definition of a digital preservation policy, it includes a general outline of what a policy should do.

- "a policy needs to convey the very philosophy of an organization concerning digital preservation; it should induce a common understanding of the objectives, of whether each Collection item should be preserved with maximum effort possibly applying multiple preservation paths, or whether a certain pragmatism should be pursued;
- a digital policy should facilitate the sustainability of an institution's present and future digital holdings;
- a digital preservation policy has to demonstrate its benefits, its effectiveness;
- a digital policy should be connected and integrated with a risk assessment document;
- every policy should be practicable, not definitive, capable of being put into practice by institutions with varying resources and needs, and, especially, flexible to adapt itself to changing administrative and technological circumstances;
- any policy should be characterized by clarity, adequacy, transparency, efficiency, effectiveness and logical organization of contents;
- a digital preservation policy should be written in a simple and suitable language, without redundancies and, at the same time, without lowering the level of quality contained in its contents;
- once a digital preservation policy is operative, it should be re-though[t], reviewed or newly conceived on a regular basis to take into account changes in the organizational, legal and technical environment and to make rules and guidelines more precise and explicit where there is any ambiguity about implementation;
- a digital policy should offer achievable solutions, provide for the management training and, finally, be maintained through time."

SOLINET

The **So**utheastern **Li**brary **Net**work (SOLINET)[10] is a not-for-profit membership organisation serving library, information, and cultural organisations in the American south-east. Their study identifies elements that should be covered by a digital preservation policy:

- "Introduction to the plan: the plan's purpose, author, organization, and update schedule.
- Institutionwide Collection priorities: a list of digital Collections to be preserved, in order of their priority to the institution. This list will be helpful in establishing budgetary guidelines. This list also makes it clear what the institution takes responsibility for in terms of digital preservation.
- Supported file formats: detailed information concerning what types of file formats will be created and supported as part of a long-term preservation strategy. Separate these formats by media type — text, images, and sound files.
- Risk assessment: an analysis explaining why certain file formats were chosen over others, and a technology forecast of how long the formats can reasonably be

[9] Digital Preservation Policy Tool http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf, pp. 3-4 (accessed: 23 May 2008).

[10] http://www.solinet.net/preservation/preservation_templ.cfm?doc_id=3678.

expected to be supported.

- Storage strategy: specified types of storage media, the number and types of copies to be held, and the locations of these copies. Include the average life expectancy of the storage media. This section should also state standards for protective enclosures and environmental controls for the storage of media such as compact discs and magnetic tapes.
- Media management: a maintenance schedule for storage media, specifying how long between checks for media readability and integrity, as well as schedules for regular media migration and/or refreshing. Include the job title of the persons responsible for these tasks."

InterPARES2

The InterPARES2 project is an international research project with experts of various disciplines, including the archival and IT field, aimed to develop and articulate concepts and principles that can ensure creation, maintenance and long-term preservation of records. It defined 'policy' as[11]

> "a formal statement of direction or guidance as to how an organization will carry out its mandate, functions or activities, motivated by determined interests or programs'"

Thus, a digital preservation policy should indicate –in general terms– which direction an organisation will follow in its digital preservation program.

A more elaborate and different definition of preservation policy in the InterPARES2 project is[12]:

> "The authoritative set of coherent policies, standards, guidelines, methods, and criteria for maintaining and preserving records, their aggregates and their related metadata as well as their constituent digital components, as long as required according to the retention policy. These policies and standards include guidelines and criteria for maintaining digital components, and for reconstituting and reproducing records in authentic form. The policy is taking into account the evaluation of the recordkeeping framework, performance information, the (intellectual and technical) Characteristics of the electronic records, usage requirements, prior preservation policies and the state of technology."

According to the InterPARES2 glossary[13], a strategy is

> "the complex of practical means formally articulated by an entity for reaching a specific purpose, that is, a plan or a road map for implementing policies".

The American Library Association

The Association for Library Collections & Technical Services of the American Library Association [14] defines digital preservation strategies according to the digital object's lifecycle stage as following:

> "Digital preservation strategies and actions address content creation, integrity and maintenance.

[11] See the InterPARES2 glossary at: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_glossary.pdf&CFID=243105&CFTOKEN=70677126, p. 20 (accessed: 23 May 2008). A similar definition can be found in R. Pearce-Moses, *A glossary of archival & records terminology.* Chicago, 2005, p. 300: "An official expression of principles that direct an organization's operations."

[12] http://www.interpares.org/display_file.cfm?doc=ip2_BDR_model(consultation_draft_20070730).pdf (p. 18 [definitions of arrows]).

[13] http://www.interpares.org/ip2/display_file.cfm?doc=ip2_glossary.pdf&CFID=243105&CFTOKEN=70677126, p. 21 (accessed 23 May 2008).

[14] Association for Library Collections & Technical Services of the American Library Association, Definitions of Digital Preservation http://www.ala.org/ala/alcts/newslinks/digipres/PARSdigdef0408.pdf

Content creation includes:

- Clear and complete technical specifications
- Production of reliable master files
- Sufficient descriptive, administrative and structural metadata to ensure future access
- Detailed quality control of processes

Content integrity includes:

- Documentation of all policies, strategies and procedures
- Use of persistent identifiers
- Recorded provenance and change history for all objects
- Verification mechanisms
- Attention to security requirements
- Routine audits

Content maintenance includes:

- A robust computing and networking infrastructure
- Storage and synchronization of files at multiple sites
- Continuous monitoring and management of files
- Programs for refreshing, migration and emulation
- Creation and testing of disaster prevention and recovery plans

Periodic review and updating of policies and procedures"

It defines digital preservation policies as follows:

"Digital preservation policies document an organization's commitment to preserve digital content for future use; specify file formats to be preserved and the level of preservation to be provided; and ensure compliance with standards and best practices for responsible stewardship of digital information."

The Technical Advisory Service for Images

TASI, the Technical Advisory Service for Images in the UK (a JISC funded service), divides the notion of a preservation strategy in a technical and an organisational part. The technical preservation strategy refers to solutions like emulation or migration, while the organisational strategy includes aspects such as budget, personnel – training, and management, elements that are also covered in the above mentioned policy documents.

JISC

In other contexts, preservation strategy or digital preservation strategy has a more narrow meaning. It refers to the existing techniques for preserving digital documents, i.e. emulation, migration, encapsulation, Universal Virtual Computer (UVC), and others.[15] This narrower definition of digital preservation strategy excludes the organisational part of the TASI terminology.

TRAC

The TRAC audit checklist[16] was published by OCLC in 2007 to provide a checklist that should be met by any certified repository. Although not intended as a conceptual model for preservation planning, it provides a list of useful organisational characteristics which we incorporated into our model..

[15] See for instance JISC, http://www.jisc.ac.uk/publications/publications/pub_digipreservationbp.aspx, PADI http://www.nla.gov.au/padi/topics/18.html, or
Cornell University, http://www.library.cornell.edu/iris/tutorial/dpm/terminology/strategies.html (accessed 23 May 2008).
Technical preservation strategy is synonym of preservation method or technique as defined in Pearce-Moses, *A glossary*, pp. 306-307.

[16] Trustworthy Repositories Audit & Certification: Criteria and Checklist. http://www.crl.edu/PDF/trac.pdf (accessed 23 May 2008).

4.2.2 Reference models

An effort has been made to align the model with

- The Open Archival Information System (OAIS) reference model which provides a basis for the consistent description of archives and repositories. Specifically OAIS identifies preservation planning as an important task, and identifies preservation metadata concepts and environment components for data objects which inform our modelling.
- Management frameworks: To be effective, preservation planning must be embedded in a broader management framework for a digital repository. The RLG/OCLC report on 'Trusted digital repositories' describes a management framework for repositories. The ISO 15489:2001 records management standard describes one for archives and records management. Factors listed have informed our vocabulary.
- The Planets Core Conceptual Data Model[8]: The Planets core conceptual model enables inter-operability between the Planets components and workflow steps. For example, the output of characterisation tools can be used as the input to create a preservation plan for the characterised deliverable units. Similarly, this preservation plan must be understood by the preservation action tools if they are to carry out the plan correctly. Interoperability is facilitated by defining the meaning and role of the various shared entities used and/or created by these Planets functions.
- Strategy document markup language: stratML provides a basic conceptual model for describing the essential contents of a strategy document. It is envisioned as an ISO standardized XML schema and vocabulary for US Federal agency strategic plans.

More effort will go into completing the alignment with those reference models in the next iteration of our work.

4.2.3 Conclusions

Literature review shows a wide range of features expected in preservation guiding documents. A comparison of ERPANET and SOLINET features, for example, makes clear that both documents differ widely in scope and detail. ERPANET deals with higher level policy concepts such as the aims of the preservation, benefits, and sustainability. SOLINET appears to address concepts on a lower level, considering details of implementation. Both might be useful and, in fact, complementary.

These differences should not come as a surprise. There is not yet a clear agreement on the use of digital preservation terms in the community, and different organisations have justifiably different preservation planning goals, resulting in the use of terms with various meanings or scope.

4.3 Analysis of preservation guiding documents

4.3.1 Previous studies on current use of digital preservation policies in organisations

As a starting point for further and more in-depth research into institutional requirements, a study of previous studies about policies was necessary. This section presents a general overview of how policies and strategies have been interpreted and identified by previous research.

Ayre and Muir at Loughborough University conducted a key survey in 2004[17] to evaluate preservation policies in the UK. The following quotation summarises the key results of their work:

> "The library questionnaire asked respondents whether or not they have a digital preservation

[17] C. Ayre and A. Muir, *Right to Preserve? Copyright and licensing for digital preservation project. Final Report*, Loughborough, 2004. Online available at http://www.lboro.ac.uk/departments/dis/disresearch/CLDP/DOCUMENTS/Final%20report.doc (accessed: 23 May 2008).

policy. Of the 69 libraries that responded to this question, only four currently have a policy. This may indicate that only four of the 122 libraries with digital Collections have a preservation policy. Even more worryingly, only four of the 51 libraries taking responsibility for the preservation of their digital resources have a policy to help them do this.

The questions also asked about respondents' intentions to develop a preservation policy in the next twelve months. Of the 85 respondents to this question, only 27.1% are planning to develop a policy with the rest uncertain or with no plans.

Interviewees were also asked about their preservation plans. One librarian who said that his library sees itself as having responsibility for preservation explained that:

saying that we have a responsibility doesn't necessarily mean we have a systematic policy.

Another librarian interviewed does not have a formal policy for preservation, but has a policy for backing up digital material. He views this as contributing to preservation:

That's partly preservation, because a lot of the content will be static, only really needs to be backed up once.

This will only help until the format in which the content is stored becomes obsolete. The publisher questionnaire asked whether publishers have formal long-term digital preservation policies. Almost 70% of those responding to this question do not have a formal strategy at the moment, but a larger proportion of publisher respondents than library respondents do have a formal policy. Thirteen publisher respondents already have policies, and a further six respondents have plans to develop a policy in the next 12 months.

The publishers interviewed were also asked about their preservation plans. One was unsure whether they had a plan and one had 'no specific plans'. One was not doing anything 'systematically' and another thought that preservation should be addressed on an 'as needs' basis. One publisher admitted that his company is 'not very good at even archiving our print material', and said that they were 'looking to start something' like this. Another felt that 'this is an issue we do need to just make sure we've thought through'. Two publishers stated that they see it as their customers' responsibility to keep the digital materials they have purchased accessible. One of these said that backing up was done for the company's benefit since users 'have got the product already, so they don't need anything else'. Despite this, most interviewees said that their intention was not to lose any of their digital materials. One publisher stated that 'our policy is that we try not to lose anything. That's about it, really', while another said: 'we have absolutely no intention of getting rid of anything that we've published'.

The author questionnaire asked whether authors take publishers' preservation policies into account when deciding where to publish their work. Of the ten respondents who replied to this question, equal numbers replied 'yes' and 'no'. One explained: 'I am first concerned with dissemination'. Another agreed that other factors are more important:

Most work in my area is published in journals. These vary considerably in specialisation and in status and breadth of dissemination. These have to be the deciding factors, given career needs.

A third respondent, who had answered 'yes' to this question added: 'BUT I have never actually seen a publisher publish a policy on preservation'.

The one organisation interviewed that does have a formal preservation policy is the British Library. This is to be expected, since the British Library is closely involved with digital preservation and is likely to play an even greater role in it once legal deposit has been fully extended to digital publications. Deborah Woodyard, the British Library's Digital Preservation Coordinator, explained that:

I'm trying to make sure that the whole of the library has a consistent approach to the preservation of digital materials, and also that the preservation of the digital material that we're collecting is taken into account in areas where it may not have been considered.

In keeping with its role as an archive of the nation's output, the British Library is planning to preserve its digital holdings, 'for the long term / indefinitely'.

The British Library's preservation policy, which has just been published [as of 2004], is very general:

It basically contains very broad, high-level statements, say, 'we intend to preserve the digital materials in the Collection'. And a couple of basic principles, like 'we will always keep an original copy'.. however it came in, although to preserve it we may need to migrate it to another format'.

The British Library is taking a 'whole life-cycle approach' to digital preservation, so it views preservation as including collecting material, producing metadata and making storage decisions. It is aware of the range of preservation strategies currently available, and is planning to use different strategies as appropriate, rather than using the same strategy for everything. Its policy therefore:

lists a whole range of different preservation strategies that are possible, and says that we will continue to examine them and use whichever one is appropriate as we see fit.

However, the policy does state that there are some strategies that will not be used:

We do say that as a strategy, 'do nothing' is not acceptable. And technology preservation is not suitable, either. So trying to keep all the different machines is not really an option.

This plan therefore reflects the fact that digital preservation strategies are still being developed.

"

It is important to note, that the situation has advanced significantly in the interim. In 2008, for example, the British Library has an updated preservation policy, an extensive risk assessment of digital material, and develops specific preservation plans for specific classes of digital content.

A 2003 ERPANET survey[18] on ‘Legislation, rules and policies for the Preservation of digital resources’ including 21 institutions, indicates that only 51% of the institutions already had a preservation policy (and these institutions were all large institutions).
The ERPANET report states on p. 27 that

‘the fact that there is not an explicit obligation, at the regulatory level, mandating to draft a policy on digital materials preservation, makes the policy tool entirely optional and therefore scarcely used.’ Moreover, it is clear ‘that even the institutions that are mandated to manage and preserve the community's cultural and scientific heritage do not always view as an essential requisite the need to design and systematically apply clear and well defined guidelines and procedures aimed at preservation.’

Finally, the ERPANET report indicates on p. 27

‘that the term used (policy) is ambiguous and that the questionnaire was not accompanied by a glossary unambiguously explaining some terms and components that may be too idiosyncratic and linked to very specific sectoral and juridical elements.’

Furthermore, Mind the Gap[19] a DPC assessment of digital preservation needs in the UK in 2005 found that in only 18% of the study participants had a preservation policy in place.

In the MLA[20] regional survey only 10 out of 23 organisations which had a corporate planning document referred in it to digital preservation, and only 6 out of 26 organisations surveyed had a digital preservation policy. However, others indicated that digital preservation was not yet embedded in policy documents, but was managed at an operational level; their policy development was imminent, they made and followed policy on a project by project basis, they

[18] M. Guercio, L. Lograno, A. Battistelli and F. Marini, *Legislation, Rules and Policies for the Preservation of Digital Resources, a survey. Draft*. Online available at http://eprints.erpanet.org/65/01/Dossier1_English_version_Full.pdf (accessed: 23 May 2008).
[19] DPC: Mind the Gap: Digital Preservation Needs in the UK http://www.ariadne.ac.uk/issue48/semple-jones/ and http://www.dpconline.org/graphics/reports/mindthegap.html (full report)
[20] Museums, Libraries and Archives Council, MLA regional study www.mla.gov.uk/resources/assets//M/mla_dpc_survey_pdf_6636.pdf

had some form of alternative, such as maintaining a 'file to paper' policy, or they adopted documentation devised by others.

They also cautioned that

> "Some organisations have corporate plans, others have business plans, others have a range of policy or strategy documents under a number of different titles. So it is not easy to be sure that like is being compared with like."

Finally, interviews conducted for this work package (see section 2.4) revealed that not all institutions had a digital preservation policy and interviewees indicated that this was not a priority.

4.3.2 Current use of digital preservation policies in organisations

This section analyses two sample preservation policies drawn from the set available online.

They were randomly chosen for analysis in order to investigate which concepts are included in this type of documents:

- Digital Preservation Policy of the State Library of Victoria | SLV (Australia)
- Digital Preservation Policy of the Digital Archives of Georgia | DAG (USA)

4.3.2.1 General analysis

Both documents have a similar length: three A4 pages, but have a different structure and scope. Both documents include a purpose statement that indicates the different role of the policy document in the organisations. While the SLV document intends to define what and how digital objects will be preserved and who will be responsible, the purpose of the DAG document is to minimize risks.

Central in the DAG document are:

- commitment to lifecycle management
- collaboration with other organisations
- use and implementation of standards and best practices, e.g. OAIS
- adherence to principles of reliable digital repositories (storage facilitation)

Central in the SLV document are:

- rules for image capture
- rules for storage and media preference
- description/cataloguing of digital objects
- copyright issues
- collaboration with other organisations

The two organisations attempt to accomplish very different goals with their policies.

4.3.2.2 Comparing concrete preservation guiding documents and abstract checklists

Because these digital preservation policies are high level documents, we chose the TRAC audit checklist[21] as a first benchmark against which to compare. The TRAC audit checklist was published by OCLC in 2007 to provide a checklist that should be met by any certified repository. Although not intended as a conceptual model for preservation planning, it provides a list of useful organisational characteristics.

		SLV	DAG
A	***Organizational infrastructure***		

[21] Trustworthy Repositories Audit & Certification: Criteria and Checklist. http://www.crl.edu/PDF/trac.pdf (accessed 23 May 2008).

A1	**Governance & organizational viability**		
A1.1	Repository has a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information.	x	x
A1.2	Repository has an appropriate, formal succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope.	-	-
A2	**Organizational structure & staffing**		
A2.1	Repository has identified and established the duties that it needs to perform and has appointed staff with adequate skills and experience to fulfil these duties	-	-
A2.2	Repository has the appropriate number of staff to support all functions and services	-	-
A2.3	Repository has an active professional development program in place that provides staff with skills and expertise development opportunities	-	-
A3	**Procedural accountability & policy framework**		
A3.1	Repository has defined its designated community(ies) and associated knowledge base(s) and has publicly accessible definitions and policies in place to dictate how its preservation service requirements will be met	x	-
A3.2	Repository has procedures and policies in place, and mechanisms for their review, update, and development as the repository grows and as technology and community practice evolve	x	x
A3.3	Repository maintains written policies that specify the nature of any legal permissions required to preserve digital content over time, and repository can demonstrate that these permissions have been acquired when needed	-	x
A3.4	Repository is committed to formal, periodic review and assessment to ensure responsiveness to technological developments and evolving requirements	x	x
A3.5	Repository has policies and procedures to ensure that feedback from producers and users is sought and addressed over time	-	-
A3.6	Repository has a documented history of the changes to its operations, procedures, software, and hardware that, where appropriate, is linked to relevant preservation strategies and describes potential effects on preserving digital content	-	-
A3.7	Repository commits to transparency and accountability in all actions supporting the operation and management of the repository, especially those that affect the preservation of digital content over time	-	-
A3.8	Repository commits to defining, collecting, tracking, and providing, on demand, its information integrity measurements	-	-
A3.9	Repository commits to a regular schedule of self-assessment and certification and, if certified, commits to notifying certifying bodies of operational changes that will change or nullify its certification status	x	x
A4	**Financial sustainability**		
A4.1	Repository has short- and long-term business planning processes in place to sustain the repository over time	-	-
A4.2	Repository has in place processes to review and adjust business plans at least annually	-	-
A4.3	Repository's financial practices and procedures are transparent, compliant with relevant accounting standards and practices, and audited by third parties in accordance with territorial legal requirements	-	-
A4.4	Repository has ongoing commitment to analyze and report on risk, benefit, investment, and expenditure (including assets, licenses, and liabilities)	-	-
A4.5	Repository commits to monitoring for and bridging gaps in funding	-	-

Legend: x: issue mentioned in Digital Preservation Policy

-: issue not mentioned in Digital Preservation Policy

Table 2 Overview of issues listed in Digital Preservation Policies, compared to TRAC's audit checklist, part A

As is shown in Table 2, the two digital preservation policies, due to their purpose, cover less than half of the issues that are related to organisation's structure and policy, as identified in the TRAC audit checklist. However, many of the central, relevant issues of the two digital preservation policies as identified above are not covered in any part of the TRAC checklist. This illustrates that this checklist alone is not a suitable benchmark to apply for preservation guiding documents, but rather can serve as a further source for relevant policy and strategy concepts contained in these documents, to be integrated into the developing conceptual model of PP2, which then can serve as a benchmark for other documents.

As indicated in the ERPANET research documents, a policy document should cover issues like costs and staffing that are clearly not included in the two example digital preservation policies. This might be controversial, however, as a policy document might be a statement of intent that is used to influence an institutions costs and staffing. A policy should stay the same even if there was a short-term downturn in available funding. To move discussions of this nature into a different forum, we avoid the issue of which concepts should be contained in which class of document, and rather develop a comprehensive overview of relevant concepts in preservation guiding documents.

4.3.3 Conclusions

- Only few institutions have developed and implemented well defined digital preservation policies and strategies.[22]
- There is no clear definition, nor a common understanding of what elements a digital preservation policy should include. Various elements, with differing levels of detail and scope, are put forward. In general, policies are very general documents that set the framework for digital preservation. Specific requirements should not be expected from these documents, as the notion of digital preservation has not been clearly delineated yet and institutions are still exploring how digital preservation should be implemented in practice.
- Existing policies do not provide sufficient detail to guide preservation planning activities.
- Most requirements at institutional and collection level are by nature not machine interpretable.
- It is not clear how accurate or representative these policies are of what their institutions want to achieve. Given that digital preservation is still an evolving concept, it would not be surprising to discover that existing preservation policies are in fact not particularly useful, fit for purpose, or accurate in reflecting what an institution is trying to achieve with preservation.
- Available digital preservation policies vary widely in length, breadth and depth, and show a wide range of features.[23]
- A digital preservation policy or strategy can be divided into technical and organisational aspects.
- Even though they are not the focus, the technical aspects of a digital preservation policy or strategy, as well as the state of technology on the basis of which high level constraints can be derived, need to be part of the research scope of PP2. Some institutions appear to mandate a particular "technical preservation strategy" (migration, for example) at the preservation policy level, regardless of the lower level

[22] Cf. the citations in the appendix and the list in the ERPANET document, pp. 9-12. http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf, pp. 3-4 (accessed: 23 May 2008).

[23] See also the conclusions of ERPANET training seminar: 'Policies for digital preservation represent an issue that still needs a lot of attention. Little practical experience yet exists and most of the ideas are still rather theoretical.' http://www.erpanet.org/events/2003/paris/ERPAtraining-Paris_Report.pdf (accessed 23 May 2008).

technical requirements. This demonstrates the need to integrate institutional and data object considerations in the conceptual model.

- The non-technical part (including aspects such as legislation and the regulatory framework) of a digital preservation strategy should be detailed enough to allow the automatic execution of preservation planning tools, but may in the near future just consist of general directions.
- It seems that for some institutions the problem is not just articulating preservation policy, but also receiving guidance as to what the particular policy should be.

4.4 Interviews

In a further bottom-up analysis, interviews with digital preservation decision makers were conducted to determine which factors influence their digital preservation decisions. Three institution types were considered: libraries, archives and data centres. The interviewed institutions were:

- The British Library
- The UK Data Archive
- The Austrian National Library
- Data Archiving and Networked Services (Netherlands)
- The Koninklijke Bibliotheek (Netherlands)
- The National Archives of Australia

The general goal of the interview was to answer the following questions:

- How do you decide which digital preservation actions to take? (as recorded in policy documents, strategy documents, business rules, informal decision processes, runtime parameters)
- What sort of things are subject to digital preservation?
- What guides you in your decision making process?
- Who has which functions in digital preservation? (decision making and execution)

followed up with requests for as much detail as possible (Why, why not, can you give an example, can you give more granularity?).

All interviews were conducted in person. The interviews were unstructured, but guided by a choice of questions from the interview help sheet which can be found in the appendix 7.2.1. The questions were not supposed to be followed rigorously, but were rather meant to inspire a stalled conversation or to provide more depth to the discussion, if necessary. The general structure followed the *Environment Component Types* hierarchy from Section 5.7. Some questions were borrowed from Erpanet studies. Where the interviewee gave permission, the interview transcript is included in the appendix.

4.4.1 Key findings

Some of the key findings that came out of the interviews are described in the following.

- Most current preservation policies and strategies "hard-wire" the choice of preservation action. No on-the-fly preservation planning is needed. These generally fall into 2 categories:
    - Preventive preservation actions, such as accepting only limited numbers of formats or performing format normalisation upon ingest (to focus on a small set of supported formats), diligence during ingest (e.g. rigorously validate and repair errors during ingest) and the use of standards (e.g. in file formats and metadata), avoid difficult preservation situations later on. While this might avoid difficult preservation issues, it must be noted that there is a risk of losing essential characteristics of the originals, especially with regard to look and feel, and especially if the normalisation happens before submission to

the institution, so that the preservation of essential characteristics is out of the control of the institution.

  - Data carrier refresh, a re-active preservation action. Now, that the first generations of data carriers are deteriorating at an alarming rate, these replacements are considered urgent and take priority over migration or emulation efforts, which can at this moment safely be postponed. We found the opinions that

    - Digital preservation is mainly a technical issue.
    - Technical issues are issues that need a solution in the near future.
    - Digital preservation is mainly dictated by the longevity of technical solutions.

  Reactive, non-hardware preservation solutions, such as migration and emulation, which require on-the-fly preservation planning are currently avoided, if possible, or not considered necessary or high-priority yet.

- One of the institutions had a preference to perform all digital preservation through emulation, although there was no emulator that met their specific needs and they felt that there are financial and legal uncertainties surrounding emulation.
- Collections are considered to be well-identified by either their file format types or by the data carrier types of the material. Most institutions have not yet accrued large mixed-format collections which would require automated discovery of existing preservation risks. This is partly because of a policy of normalisation, and partly because many digital collections are still rather small at this point.
- The choice of migration tools is generally considered straight-forward. There are few alternatives to chose from and they are perceived to have clearly identifiable advantages for the given situation.
- Most institutions use digital preservation watch to keep up to date with current developments concerning file formats, existing preservation action tools and general danger to digital objects in repositories.
- Most institutions felt that there are few factors that limit them in their preservation decision making. It was, for example, felt that
  - The legal framework does not practically limit the choice of hardware, software, data carriers, preservation actions, or formats used. It determines **why**, but **not how** things are done.
  - Budget, legislation, personnel, structure of organisation, etc. are in some ways important for digital preservation, but are by no means drivers or decisive factors for digital preservation.
  - Some institutions consider the cost of preservation and storage a minor issue. If there is a legal mandate to preserve, then ways have to be found to finance preservation.
- With respect to the producer and consumer communities we found the following:
  - Access policies are based on internal policy as well as on agreements with content providers.
  - The interviewed institutions are not always able to demand compliance with standardised input formats. They often have to accept whatever they get and in whatever format they receive it. Quality control can be an issue.
  - The requirements of the (future) consumer must be kept in mind when making preservation decisions.
  - Agreements with content providers influence preservation policy.
- Things that were consciously omitted from preservation guiding documents were the following:

  - Costs of storage
  - Cost of preservation actions
  - Hardware is not part of digital preservation planning within some organisations; this is handled by IT departments.
- Tools that were considered desirable for preservation support were the following:
  - New preservation action tools, such as on-demand migration tools for old versions of statistics packages to newer versions. Some institutions develop their own tools for digital preservation.
  - A preservation policy checklist which assists in writing preservation policy documents.
  - A checklist which assists in collecting all necessary facts and documents when ingesting new material based on underlying registries (e.g. to registered license agreements with all universities).
  - Characterisation tools.
  - Automated preservation risk checking routine of random samples of digital objects.
  - Tools which analyse the file format composition of a collection (collection profiling tools).
  - Tools which automate or semi-automate processes that are executed manually at the moment (such as risk analysis, preservation execution, quality assurance).

Additional interviews may be conducted with other institution types, such as museums and audio-visual archives, since their digital collections differ from those of the previously interviewed institutions.

4.5 Example extraction of concepts and requirements for the conceptual model

Textual analysis of frameworks, preservation guiding documents, and the interviews enabled us to identify key concepts and include them in the conceptual model. Some examples on how this approach was implemented are given on the basis of sections of preservation guiding documents. The concepts which are referred to in the example are only defined later in the document. For now this section should only illustrate the methodology.

UK Data Archive Preservation Policy

p. 11: "The UKDA has chosen to implement a preservation strategy based upon open and available file formats, data migration and media refreshment."

Example derived concepts are:

- Preservation Object: bytestream, institution = UKDA
- Preservation Action Types: migration, media refresh
- Environment Components: format, data carrier
- Properties / Values: format availability, format openness

Example derived requirements are:

- File format must be open.
- File format specifications must be available.
- Preservation actions must be migration or media refresh.

Digital Preservation Policy, State Library of Victoria

"**Storage**. Born-digital objects published on disk (CD-R or DVD) are considered the archival copy and will be stored appropriately. When needed and authority granted, the physical format data may be copied to another storage carrier in order to preserve its contents. The master TIFF files shall be stored appropriately in a secure location on the Library's LAN, and back-ups made in accordance with TSD policy."

Example derived concepts are:

- Preservation Object: (file) collection (on disk), bytestream
- Preservation Action Types: conservation storage, copy, back up
- Environment Components: data carrier, store, TSD policy, format
- Properties / Values: storage location / LAN, origin / born digital, data carrier technology / disk (CD-R or DVD), file format designation / TIFF

Example derived requirements are:

- CD-R and DVD should be stored in appropriate conditions.
- Preservation action may be backup.
- Preservation action must comply with TSD policy.
- Master Tiff files must be stored on LAN.

The requirements can be broadened to keep the datasets, documentation and metadata in conditions suitable for long-term archival storage and to define what "appropriate conditions" means.

Requirement can be redefined into several more specific requirements, based on other preservation guiding documents.

An approach to the preservation of digital records, National Archives of Australia

p. 14: “The digital preservation program must be able to preserve any digital record that is brought into National Archives’ custody regardless of the application or system it is from or data format it is stored in.”

Example derived concepts are:

- Preservation Object: bytestream, institution = National Archives of Australia
- Preservation Action Types: ingest, preservation
- Environment Components: application software, hardware, format, content/self

Example derived requirement is:

- All records that are ingested should be preserved.

Corollaries might be:

- There should be a plan for every file format in custody.
- Do not accept custody for any formats that do not have a plan.

ISO/TR 18492:2005: Long-term preservation of electronic document-based information

p. 12: “Migration to standard formats. Storage repositories should consider migrating electronic document-based information from the wide variety of formats used by creators or recipients to a smaller number of “standardized” formats upon their transfer to the custody of the repository. “Standardized” formats could be a consensus on formats that are widely used and are likely to cover a majority of a particular class of electronic document-based information. Proprietary file formats should be avoided. Among the technology neutral formats that merit consideration are PDF/A-1, XML, TIFF and JPEG.”

Example derived concepts are:

- Preservation Object: bytestream
- Preservation Action Types: migration
- Environment Components: format
- Properties / Values: format designation / PDF/A-1, XML, TIFF and JPEG.

Example derived requirements are:

- Migrate ingested files with non-standard formats.
- Preferred migration format for text is PDF/A-1 or XML.
- Preferred migration format for images is TIFF or JPEG.

4.5.1 Requirements base

In the next iteration of this work, we will attempt to represent the requirements in the requirements base as OCL expressions. The initial list of requirements, which we extracted during literature analysis, document analysis and interviews, are expressed in natural language and are listed in this section. Many of these requirements are by nature not machine-interpretable. In order to translate the ones that are machine-interpretable to OCL

- The conceptual model needs to be refined and extended to be able to express all concepts found within the requirements.
- The requirements need to be expressed with more precision. Crisp, measurable definitions are needed that permit evaluation tools to determine whether the constraints are satisfied.

Note:

There are sets of requirements which can be referenced summarily from one requirement by virtually pointing into a registry. For example, the requirement “Ensure that *Preservation Actions* are in compliance with requirements of the <*funding agency*> “ refers to the requirements defined by a *funding agency*. A particular institution can instantiate the value for <*funding agency*> in the requirement, and refer to the individual requirements collected in the requirements registry of funding agencies.

Note:

Significant Properties are not contained in this requirements base since much work is going into modelling them in other work.

Preservation Guiding Requirement

(dependent on *Characteristic*s of input and output *Preservation Object*s and *Preservation Actions*)

1.01. Eliminate software dependence by sacrificing structure.
1.02. Sacrifice usability for authenticity (or vice versa).
1.03. Provide authentic, reliable versions of data collections to the designated user community
1.04. Maintain the integrity and quality of the data collections.
1.05. Preservation actions must be compatible with the medium most appropriate for the task the digital resource performs.
1.06. Preservation output formats must be chosen with specific reference to the "data types" under consideration.
1.07. Always use migration to an open format as preservation action.
1.08. Textual data must be migrated to XML or RTF (or PDF...) formats.
1.09. Digital video data must be migrated to MPEG2, etc..
1.10. Emails must be migrated to XML.
1.11. Similarly for other content-types (table of original to target formats).
1.12. Don't migrate jpeg files to another format, unless lossless compression is possible.
1.13. Migrate image files to PNG format, with the exception of jpeg images.
1.14. Migrate text documents to ODT format.
1.15. Objects owned by x may only be preserved as a single digital object (you cannot have more than one copy, not several Manifestations).
1.16. Objects owned by NLW must be preserved according to NLW preservation policy (in shared repository).
1.17. The cost of a Preservation Action may not exceed the value of the object to be preserved.
1.18. The cost of executing the Preservation *Action* may not exceed the preservation budget(by more than x %) .
1.19. Don't produce output manifestations/files that are larger than x Bytes.
1.20. Prefer output manifestations whose file formats are supported by existing HW and SW .
1.21. Must preserve colour information.
1.22. The font type may not be changed unless the font type is related to proprietary software.
1.23. Preservation Actions must comply with all legal (national, regional, archival,...) requirements (refers to registry of legal requirements).
1.24. Ensure that preservation actions are in compliance with requirements of the funding agency.
1.25. Ensure that all documents are preserved as required by the funding agency.
1.26. Ensure that applicable international and national standards are observed.
1.27. Ensure that the preservation process is in accordance with the ISO quality standards.
1.28. Private correspondence in government agency email repositories may not be preserved (Belgium).
1.29. The staff cost of supporting new output environments must follow rules in document x.
1.30. Prefer open formats for migration.
1.31. If applicable, use lossless compression techniques.
1.32. Only the following file formats are accepted: <list of formats>.
1.33. Don't accept file formats that are licensed by <x>.
1.34. Accept that pagination after migration is not identical.

Action Defining Requirement

(dependent on *Characteristic*s of *Preservation Actions*)

2.01. Target file format specifications must be available

2.02. Target file formats must be open
2.03. Preservation Actions must be one of migration or data refresh (or...)
2.04. Chose Preservation Actions which improve the speed and efficiency with which information is preserved and retrieved
2.05. Prefer preservation actions which are more stable
2.06. Prefer preservation actions which are better supported (compares several output formats and refers to the file format registry)
2.07. Chose preservation actions which create platform independent objects
2.08. Chose preservation actions which optimise the use of space for storage purposes
2.09. The preservation actions undertaken are uniform regardless of the perceived value of any dataset
2.10. Original data carriers will not be preserved
2.11. Preservation output formats must be information-rich
2.12. The preservation process must be OAIS compliant
2.13. Prefer preservation actions which use software under existing licenses
2.14. Prefer preservation actions which produce target outputs which satisfy the main user needs
2.15. Prefer preservation actions for which there is expertise
2.16. Prefer preservation tools that were developed in-house
2.17. With every preservation action produce a print-quality (300dpi) PDF for print-on-demand customers.
2.18. Don't archive derivative copies which can be derived from others)
2.19. Never delete original copies.

Risk Specifying Requirement

3.1. If HW/SW becomes obsolete then perform preservation actions
3.2. Ensure timely upgrades in both hardware and software.

Preservation Object Selecting Requirement

4.01. Preserve digital resource subsets for which random sampling shows more than 0.5% corruption.
4.02. Preserve digital resource subsets where enough objects of a given value are preserved to amortize a certain Preservation Action.
4.03. Preserve digital objects for which we do not have printed backup.

Preservation Process Guiding Requirement

(independent of *Characteristic*s)

5.01. The Institution must migrate objects at least every 5 years
5.02. Depositors must be notified of ingest and Preservation Actions taken before release to the community
5.03. If a new HW/SW version becomes available then evaluate preservation need
5.04. Every preservation action needs to be documented with the target object, so that changes can be traced
5.05. Every preservation action needs to be validated for authenticity of the substantive content
5.06. A preservation plan has to ensure best use of resources
5.07. Ensure that all data Collections are protected and kept secure during preservation
5.08. Follow good practice in active preservation management
5.09. Develop and maintain systems of low-cost storage, with appropriate location and with regular review
5.10. Optimise the use of space for storage purposes
5.11. Keep the datasets, documentation and metadata in conditions suitable for long-term archival storage
5.12. With every preservation action old derived files are retained / discarded
5.13. With every preservation action original files are retained
5.14. Preservation action output files have to follow a consistent directory structure for storage
5.15. Preservation action output files must have standardised file extensions with a single extension allowable for each type of file
5.16. Preservation Actions must be accompanied with formal documentation specified in the preservation strategy procedures

5.17. If metadata is supplied it must be preserved
5.18. If metadata is not supplied then don't produce it
5.19. If the budget for digital preservation has been consumed for a fiscal year, then postpone preservation actions or chose a cheaper, but less effective and less reliable preservation approach

Preservation Infrastructure Requirement

This kind of *Requirement* is most frequently found in *Preservation Guiding Documents*. Many more examples were found than are listed here. Unfortunately they do not lend themselves to supporting automatic preservation planning.

6.01. To provide an adequate level of redundancy the preservation system must consist of on-site, near-site and off-site storage.
6.02. Mirror versions of on-site systems must be provided
6.03. Digitised materials need at least 2 digital copies
6.04. Digitally born materials need at least 3 digital copies
6.05. Different operating systems must be installed across the systems
6.06. Adequate storage capacity for all holdings must be maintained
6.07. Unlimited capacity from external media is to be provided at all times
6.08. Must use secure networking and communications equipment
6.09. Must provide adequate connectivity
6.10. Must provide the ability to restrict to valid Mac addresses
6.11. Must provide a facility to segment the network for switched separated firewall connectivity
6.12. All servers must be protected by power surge protection systems
6.13. Disaster recovery procedures must be in place
6.14. Always preserve the original bit stream and preserve one migrated version on two separate systems
6.15. Before ingesting documents in the preservation system, do a virus and malware check
6.16. Before ingesting documents in the preservation system, put the received documents in quarantine for 28 days
6.17. Always preserve all documents (originals and normalised/migrated versions) on two separate systems coming from different hard- and software developers
6.18. Constantly do integrity checking by screening checksums of preserved files
6.19. Ensure that tools are written in standardised programming syntax that is documented and easily understood (by persons not involved in the programming process)
6.19. Ensure that tools for digital preservation are built and written by employees of the organisation
6.19. Ensure that tools for digital preservation are open source, so that other organisations can help to refine and upgrade them (collaboration)
6.19. Ensure that the two physical systems, on which the documents are stored, are in two different buildings
6.19. If Characteristics are lost during the migration process, re-migrate the originally received documents after fixing the migration issues
6.19. Ensure that digital preservation programming is done according programming best practices
6.19. Ensure that physical carriers are stored in a physical environment that is suitable and secured for long-term preservation

4.6 Differences between institutional types

This section reflects on how the conceptual model for preservation guiding documents varies amongst the three types of institutions studied: libraries, archives, and data centres Our initial expectation was that there would be substantial differences in preservation policies across the types of institutions that we studied. Our analysis, however, shows that all institutions studied used very similar concepts.

Our confidence in this finding is tempered, however, because

- We have a very small sample size, so it is not possible to draw statistically significant conclusions;
- The documents studied are mostly based on theoretical considerations and may lack the essential details which might differentiate institutional types. Growing practical experience might produce insights into differences which we do not have at the moment;
- Institutions sometimes take on a role for a certain collection which is different from the role that is suggested in their title. If an archive, for example, is responsible for old census data then it acts like a data centre for this collection. If a corporation preserves data of historically seminal technology then it functions as an archive for that data. This change of role confuses potentially existent inherent differences.
- Institutions almost always have multiple roles for their repositories which might be conflicting. In that case the multiple roles confuse the issue about which concept is relevant for a given role. Conflicting roles that might be found in the same place are
  - enabling access versus supporting preservation.
  - contemporary usability of data objects versus reflecting the authenticity of the original.

This finding, however, caused us to make a substantial change in the design of our subsequent work. Rather than developing separate models for each institutional type, and then integrating them upon their completion, concepts for all institution types were incrementally integrated into one model. Each concept was annotated with the institution types from which it was derived, in case key difference were to arise later on.

Although the key concepts and requirements do not appear to be substantially different, we did informally observe some differences in emphasis.

National libraries

National libraries have a goal to preserve as much a possible and emphasise both presentation and long-term access to the material. This also means that it is important to preserve as much dynamic behaviour as possible. Some deposit libraries might be legally required to care for certain objects. Most libraries have little to no control over acceptable ingest formats. They would rather not refuse any object because of the file format it is in. For libraries, batch preservation is essential due to the large size of their collections.

Archives

Within archives, there is less perceived need to preserve the interactive behaviour of content. However, authenticity and integrity are extremely important. The value is the same for all the objects. Most archives are legally required to appraise, preserve and provide access to authentic (government) records to anyone who has any interest in them. Preservation issues should be considered from the point of creation of the records within government agencies; they have, however, not been included in this work.

Corporate repositories

Corporate repositories archive material that is produced by the company. This is usually not for public access. One motivator is to enable the company to protect its intellectual property. Since in this case the main goal of the repository is to produce documentary evidence, authenticity is very important. Another motivator is to enable the company to protect its intellectual capital for business processes, such as the safekeeping of product data for maintenance and adaptation. In this case authenticity is not the primary goal, but rather the usability of the data by future design and manufacturing tools. The timescale is dictated by the product lifetime or regulatory requirements, while libraries and archives tend to work with indefinite timescales.

Data centres

The goal for data centres is to preserve the data as authentic as possible; but usability is also very important, and one may sacrifice authenticity for usability if, for example, technological advances allow for improved analysis. This needs to documented in detail in event logs.

Complying with current legislation is in many cases the responsibility of the content provider, not of the data centre. Privacy legislation is important if personal information is included in the data. There is no legal obligation for data centres to care for any material. They are free to select what is deposited and thus have much control over the formats that are ingested. Preservation is not done in batches, but on a case by case basis – data centres typically hold a relatively small number of relatively large homogeneous data sets.

5. A conceptual model and specific vocabulary

5.1 Introduction

This section proposes a conceptual model and vocabulary for preservation guiding documents. It can be shared across all of the work-packages within preservation planning and can also be useful for communicating about preservation planning with other work packages within the Planets project, as well as outside of the project.

The core of a preservation planning model are the requirements which are expressed in preservation guiding documents and all preservation objects and characteristics to which those requirements refer. Besides these requirements, however, there are some general aspects which should be contained in preservation guiding documents. We borrow some basics from a model called stratML.

The model also draws from the OAIS model and from the Planets core conceptual model[8], which defines key concepts related to digital objects, and should eventually also be synchronised with the partial models in use within other individual work-packages.

The description of the model and vocabulary are combined in this section. In each section, a new concept with its elements and relationships is introduced, followed by a description of its vocabulary.

5.2 Use of the model and vocabulary

The key components of our approach are an abstract conceptual model, the specific vocabulary accompanying it and an example machine-interpretable model. This section will illustrate how they can be used by an institution to create a machine-interpretable model which corresponds to the (machine-interpretable parts) of this institution's preservation guiding documents.

Obviously, a smaller percentage of requirements is machine-interpretable for higher-level objects, such as *Collections*, than for lower-level objects, such as *Bytestreams*; the non-machine-interpretable ones tend to be described in more abstract preservation guiding documents, such as policy documents. The percentage of machine-interpretable requirements increases as preservation guiding documents become more and more concrete, moving from strategy documents to runtime parameters. Even if requirements cannot be expressed in a machine-interpretable way, the vocabulary offers a starting point for creating individualized models for an institution. In addition, it is advantageous to incorporate the ones which are machine-interpretable at any level uniformly into the preservation planning process.

5.2.1 A worked example

Figure 1 gives an overview of how the models in this report can be used. The numbering in the following refers to components in the figure. Numbering including the letter "a" describes components in the general model, which are described in this report. Numbering including the letter "b" describes components in an instantiated model, which an institution might create from the general model..

(1a) The conceptual model gives a very concise description of the basic concepts which are needed in the domain of organisational preservation policies and strategies as they apply to preservation planning. It also specifies the relationships between them. They comprise *Preservation Objects*, *Environments*, *Environment Components*, *Characteristics*, *Preservation Actions* and *Requirements*, expressed as a UML domain model. This is an extension of the Planets core conceptual model[8] and currently not yet fully aligned. It can be reused by other work packages across the project.

(2a) The specific vocabulary describes which subtypes of the basic concepts exist. It also describes which properties exist for the *Environment Components* of all types of *Preservation*

*Object*s. It describes which values *Properties* can take[24]. It is a representative (i.e. not exhaustive) specific vocabulary expressed as a UML domain model.

(3a) The requirements base describes classes of organisational requirements which may be contained in preservation guiding documents. They are expressed solely in terms of the concepts and attributes of our conceptual model and of the specific vocabulary. They may be parameterised so that they can be instantiated to a specific institution's conditions. We are planning to represent requirements in OCL.

Significant Properties are a special class of requirements that are not contained in the requirements base since much effort is going into modelling them in other work. If an institution should chose to, it may, however, express them consistent with this model, so that they can be integrated into a holistic planning process for the institution.

(4a) The elements in the conceptual model, the specific vocabulary, and the requirements base can be translated into several implementation specific machine interpretable representations. An example machine-interpretable model, expressed as XML schema, and its documentation, for both the conceptual model and the vocabulary should be contained in the final deliverable of this work-package.

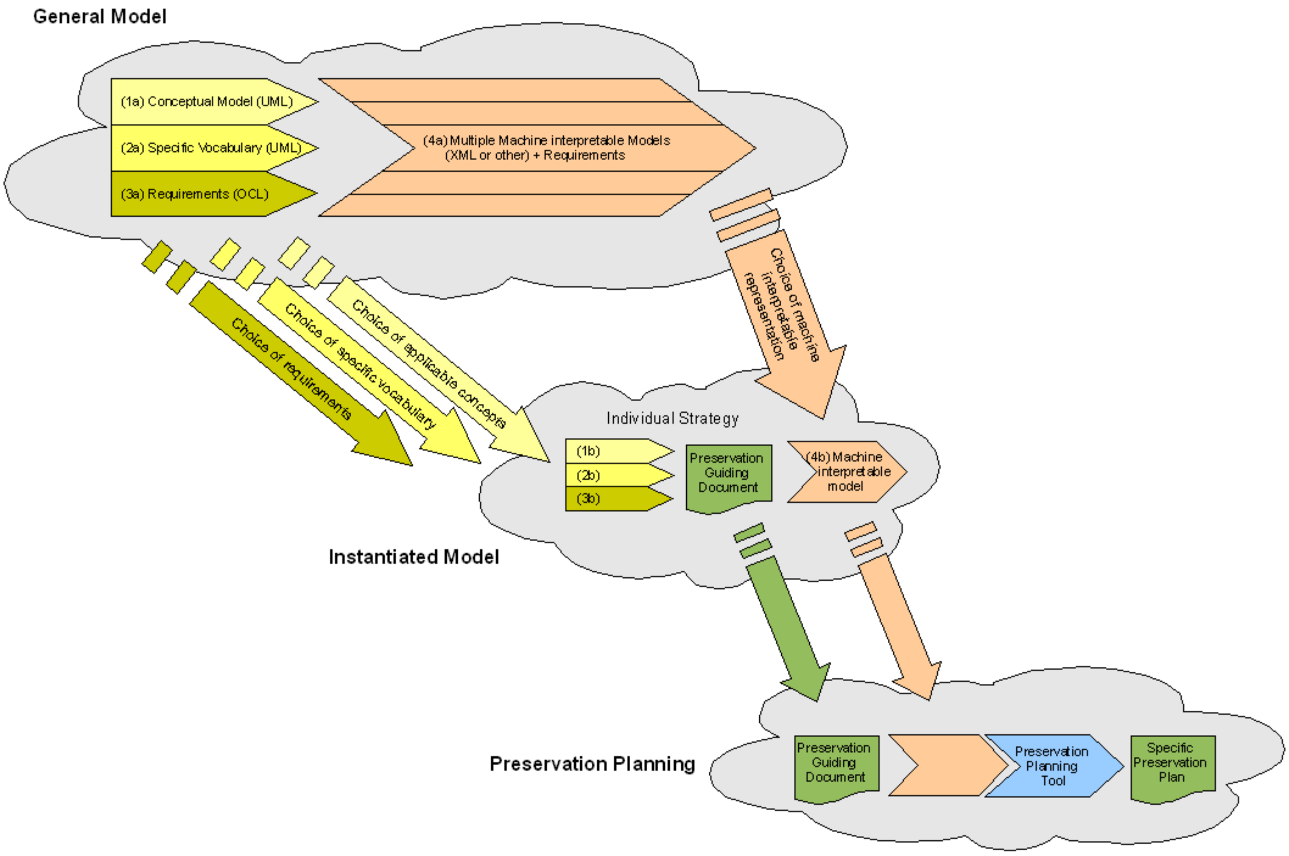

General Model
(1a) Conceptual Model (UML)
(2a) Specific Vocabulary (UML)
(3a) Requirements (OCL)
(4a) Multiple Machine interpretable Models (XML or other) + Requirements
Choice of requirements
Choice of specific vocabulary
Choice of applicable concepts
Choice of machine interpretable representation
Individual Strategy
(1b)
(2b)
(3b)
Preservation Guiding Document
(4b) Machine interpretable model
Instantiated Model
Preservation Planning
Preservation Guiding Document
Preservation Planning Tool
Specific Preservation Plan

Figure 1 Overview over the PP2 deliverables

(1b) The institution chooses which of these concepts are supported in its setting and are needed by its preservation planning service.

Since the conceptual model is very concise, in most cases all of the concepts would be expected to be used.

(2b) The institution chooses which specific vocabulary applies to it. The institution also assigns values to the *Characteristics* which describe its preservation *Environment* if these values will not be measured automatically, or otherwise specifies the method of obtaining measurements or derivations.

[24] This is not contained in this report.

(3b) The institution chooses which *Requirements* in the *Requirements* base apply and instantiates them, so that they are now un-parameterised.

The outputs of steps (1b), (2b) and (3b) form the core part of a preservation guiding document.

(4b) From the choices of steps (1b), (2b) and (3b), and the choice of machine-interpretable model results an instantiated machine-interpretable description of the institutional *Requirements* which can serve as a basis for automated preservation planning.

The planning tool now matches the *Requirements* in the machine-interpretable version of the *Preservation Guiding Document* (4b) against the state of the institution to see which *Preservation Actions* can best satisfy the *Requirements* under the given state.

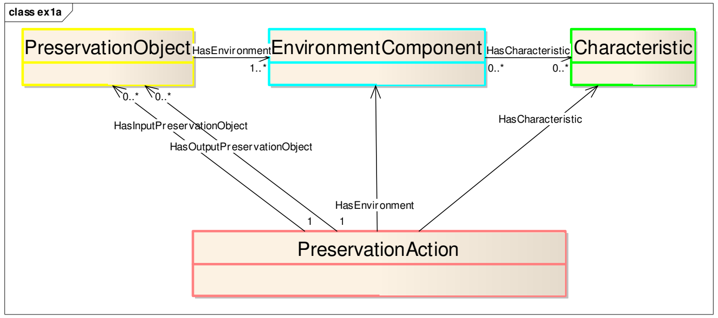
Example:

The following table gives an example illustrating the issues raised in the preceding figure and shows how the models in this report can be used. It uses concepts which will be defined only later in the document. For the moment, this will have to serve as an intuitive example. The example is simplified compared to the actual model to aid the illustration. Also, not all features mentioned in the example are yet fully implemented in this iteration of work. The left column illustrates the deliverables of PP2. The right column illustrates how these results may be used by preservation planning tools for a given institution.

(1a)

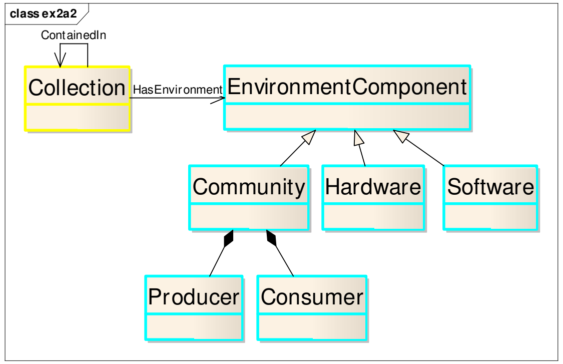
Abstract model: The abstract model gives a very concise description of the basic concepts which are needed in the domain and the relationships between them.

E.g. “PreservationObject”, “*Environment*Component”, “Characteristic”



(2a) The specific vocabulary describes which subtypes of the generic concepts exist.

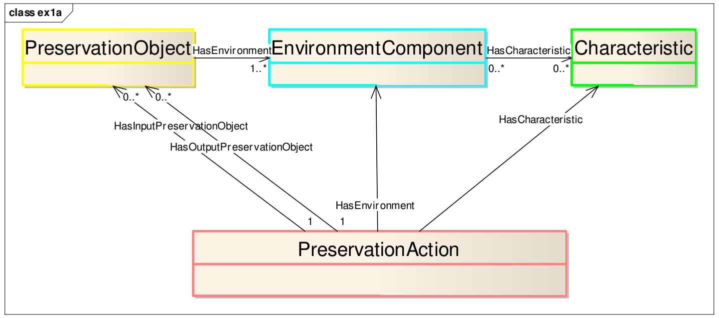
E.g. for *Environment Component*s of the collection “Producer” is a subconcept of “Community”



(1b)

The institution chooses which of these concepts are supported in its setting and are needed by its preservation planning tool.

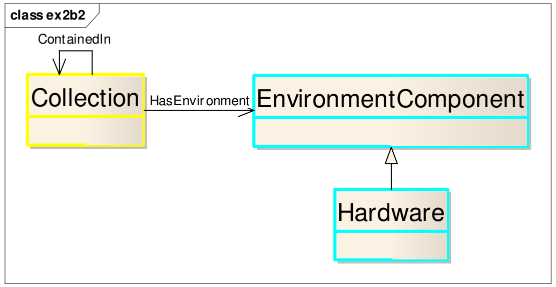
Since the conceptual model is very concise, in most cases, all of the concepts would be expected to be used



(2b)

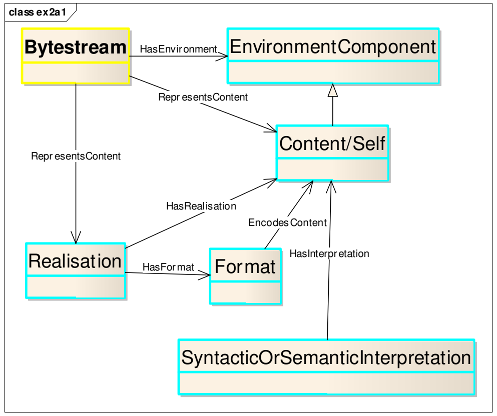
The institution chooses which specific vocabulary applies to it.

E.g. for *Environment Component*s of the collection, the institution only wishes to express requirements about its hardware

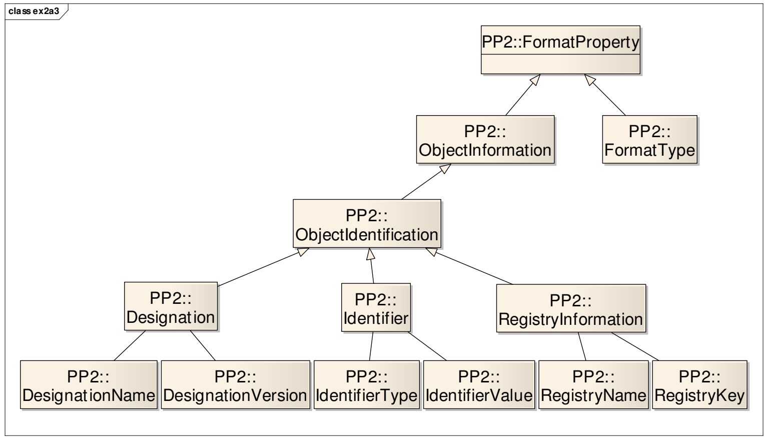


(2a contd.)

E.g. for *Environment Components* of Bytestream objects “Format” is a subconcept of the concept “Content/Self”

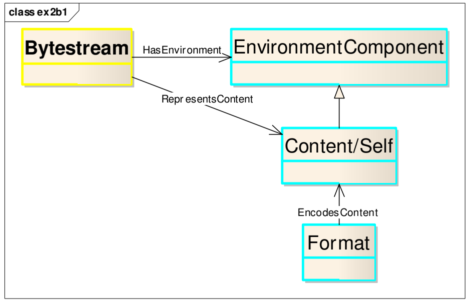


The specific vocabulary also describes which properties exist E.g. for *Properties* of formats “DesignationName” is a subconcept of the concept “FormatProperty”.

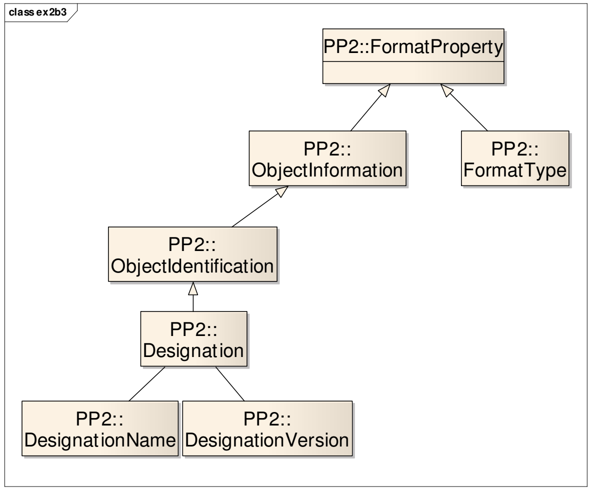


(2b contd.)

E.g. for *Environment Component*s of Bytestream objects the institution only wishes to express requirements about their content and their format



E.g. for *Properties* of formats the institution only wishes to express designation information and the format type.



(2a contd.)

The specific vocabulary also describes which values attributes can take.

E.g. for “FormatType” of *Environment Component* “Format” the allowable values are “textFormat”, “imageFormat”, “applicationFormat”, “audioFormat”

For “DesignationName” of *Environment Component* “Format” the allowable values are “jpg”, “jpeg2000” , “gif” , “tiff”, ……

(2b contd.)

E.g. The institution only allows format types “textFormat” or “imageFormat” for *Environment Component* “Format”.

The institution also assigns values to the *Characteristic*s which describe its preservation *Environment* if these values will not be measured automatically, or otherwise it specifies the method of obtaining measurements or derivations.

E.g. The institution assigns the following values

- DesignationName of a ByteStream object will be set on ingest to the value of the <format> element of Jhove, Release 1.1 output
- DesignationVersion of a ByteStream object will be set upon ingest to the value of the <version> element of Jhove, Verion Release 1.1 output.
- Hardware *Characteristic*s of the institution will be set in its hardware registry (e.g. model, manufacturer, cost, usability, training needs, etc)
- Characteristics describing the specific hardware held by the institution will be set in its hardware inventory (e.g. number of licenses held, age, capacity, access rights, repair record, etc.)

These values can be accessed by the preservation planning tool.

(3a)

The requirements base contains classes of organisational requirements which may be contained in preservation guiding documents.

A generic English language requirement might state the following

"If an input file is of type "textFormat" then the *Preservation Action* must produce an output file of format *preferredFormat*."

To create a generic requirement we introduced a variable ***preferredFormat*** which can be instantiated differently by each institution.

This requirement falls into the category of "preservation guiding requirement".

The requirement is called **NormaliseTextInput (preferredFormat)**.

It refers solely to elements of our conceptual model:

Objects: "PreservationAction", "Bytestream",

Relationships: "HasInputObject" and "HasOutputObject",

Properties: "FormatType", "DesignationName" and "DesignationVersion" for *Environment Component* "Format"

We represent requirement classes in OCL (Object Constraint Language).

An abstract OCL representation of this requirement might look as follows:

```
        Context PreservationAction
        Pre: InputBytestream.FormatType=textFile
        Post:
        OutputBytestream.DesignationName=preferredFormat.DesignationName and
        OutputBytestream.DesignationVersion=preferredFormat.DesignationVersion
```

(3b)

The institution can then choose applicable requirements and instantiate them according to its needs.

Its strategy document might contain an instantiation of this requirement, corresponding to the English language requirement

"If an input file is of type "text file" then the *Preservation Action* must produce an output file of format *PDF/A*".

These abstract requirements can be instantiated by each institution as they wish.

E.g.

The organisation can instantiate the above requirement with a call that might look similar to:

myPreferredFormat.DesignationName:= "PDF/A"

myPreferredFormat.DesignationVersion := "1a"

new Requirement NormaliseTextInput (myPreferredFormat)

thereby declaring that their preferred output format for text files is ***PDF/A-1a***.

(3b contd.)

The outputs of steps (1b), (2b) and (3b) form the core part of a preservation guiding document.

(4a)

The elements in the conceptual model, the specific vocabulary, and the requirements base can be translated into several implementation specific machine interpretable representations. They may be represented in an XML schema.

```
 <?xml version="1.0" encoding="UTF-8" ?>
- <xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">
- <xsd:element name="Institution" >
- <xsd:complexType>
- <xsd:sequence>
 <xsd:element ref="Name" minOccurs="0" />
 <xsd:element ref="Acronym" minOccurs="0" />
     </xsd:sequence>
     </xsd:complexType>
     </xsd:element>
- <xsd:element name="Requirement">
- <xsd:annotation>
 <xsd:documentation>Requirements are measurable subsets of goals expressed in
         Characteristics, values and units.</xsd:documentation>
     </xsd:annotation>
- <xsd:complexType>
- <xsd:sequence>
 <xsd:element ref="SequenceIndicator" minOccurs="0" />
 <xsd:element ref="Name" minOccurs="0" />
 <xsd:element ref="Description"/>
 <xsd:element ref="Stakeholder" minOccurs="0" maxOccurs="unbounded" />
 <xsd:element ref="OCL Definition" minOccurs="0" />
       …
     </xsd:sequence>
     </xsd:complexType>
     </xsd:element>
       …
```

(4b)

The institution specifies the machine interpretable model of its choice.

From the choices of steps (1b), (2b) and (3b), and the choice of machine-interpretable model results an instantiated, machine-interpretable description of the institutional requirements which can serve as a basis for automated preservation planning.

This may be represented in an XML document.

E.g.

```
 <?xml version="1.0" encoding="UTF-8" ?>
- <PreservationStrategy>
 - <Institution>
        <Name>The British Library</Name>
        <Acronym>BL</Acronym>
     </Institution>
 - <Requirement>
        <SequenceIndicator>1.1</SequenceIndicator>
        <Name>NormaliseTextInput</Name>
        <Description>This requirement specifies the default output format to which
             ingested text files are migrated.</Description>
        <Stakeholder> Digital Preservation Team</Stakeholder>
     .....

     </Objective>

       ……
```

The planning tool now matches the requirements in the machine-interpretable version of the preservation guiding document (4b) against the *Values* of the *Characteristics* which describe the institution at that time to see which *Preservation Action*s can best satisfy the *Requirements* at the given time.

5.2.2 Use

- The model and vocabulary are intended to be input to current and future preservation planning tools. It is therefore advantageous to have a general model and vocabulary which can be used by different approaches. If the model is valid and the systems which use it are well-defined, it should be fairly easy to do a mapping from the abstract model into desired machine-interpretable representations.
- The same reasoning applies to the requirements base. OCL is a part of UML. It is an abstract modelling language which can be translated into many machine-interpretable representations. Tools exist already to translate OCL models into languages like Java. We can translate it into a suitable XML representation. Translating it for the use of the Plato tool being developed in PP/4 will be a first practical validation of the model which should allow us to improve our model and vocabulary.
- The model is supposed to be consistent, as complete as possible and necessary, and useful. Therefore the model's vocabulary exceeds the needs of most institutions. Not every concept *Type*, *Property*, permissible *Value*, or *Preservation Requirement* applies to every institution. Not every system which uses the model has to make use of all its features. Instead, a choice of elements of this model can be composed to describe a unique institution's preservation *Environment* and *Requirement*s.
- The model's vocabulary, of course, cannot cover the needs of all institutions. Most institutions will have concept *Types*, *Properties*, permissible *Values*, or *Preservation Requirement*s that are unique. Therefore, it is important that the vocabulary can be extended by concepts which are specific to an institution.
- We do not provide an instantiated model for ready use. Every institution will need to chose and instantiate their applicable *Environment Component*s, *Properties*, permissible *Values*, or *Preservation Requirement*s.
- Preservation experts can use the model to build up a *Requirement*s base that can be instantiated for many organisations.
- The complexity of the model can be limited to describe simpler systems with lesser complexity, e.g. by reducing the expressiveness of the constraint language for *Requirement*s.

5.3 Modelling the context of preservation planning

Preservation planning is a process which identifies and mitigates risks to current and future access to digital objects. Preservation planning involves information about an institution's policies and goals, its infrastructure, its user community, and the external *Environment* in addition to information about the digital objects held within a *Collection*.

Preservation planning goals are to

- Identify which parts of the *Collection* present the greatest risks (- risk analysis and assesment)
  (or alternatively: Identify which parts of the *Collection* present the greatest opportunities for improvement)
- Identify candidate *Preservation Actions* (alternatives) that could be taken to mitigate the risks (- determine candidate solutions)
- Evaluate the candidate *Preservation Actions* to determine their potential costs and benefits, (- cost/benefit analysis of candidate solution)
- Weigh the cost/benefit of candidate *Preservation Actions*. The cost may comprise the cost of executing the action, the cost of needed infrastructure for sustaining preservation output, the cost of essential *Characteristic*s lost in the *Preservation Action* (i.e. loss of authenticity) etc.. The benefit of the preservation action is the benefit of mitigating the risk in terms of

the value of the object, the severity of the risk, etc.. Obviously these costs and benefits are not necessarily monetary.

- Provide justified recommendations for which actions to execute on which *Collections*.

The result of the preservation planning process is a set of justified prioritised recommendations for actions that mitigate the risks presented to aspects of a Collection. These recommendations are defined as workflows that, in some cases, can be executed automatically by a preservation plan execution engine.

An essential aspect of this preservation planning model is that it takes into account the goals and limitations of the institution, features of its user community, and the environment in which its users access digital content. Thus, the scope of preservation planning extends beyond merely considering file formats and preserving *Characteristics* of individual digital objects.

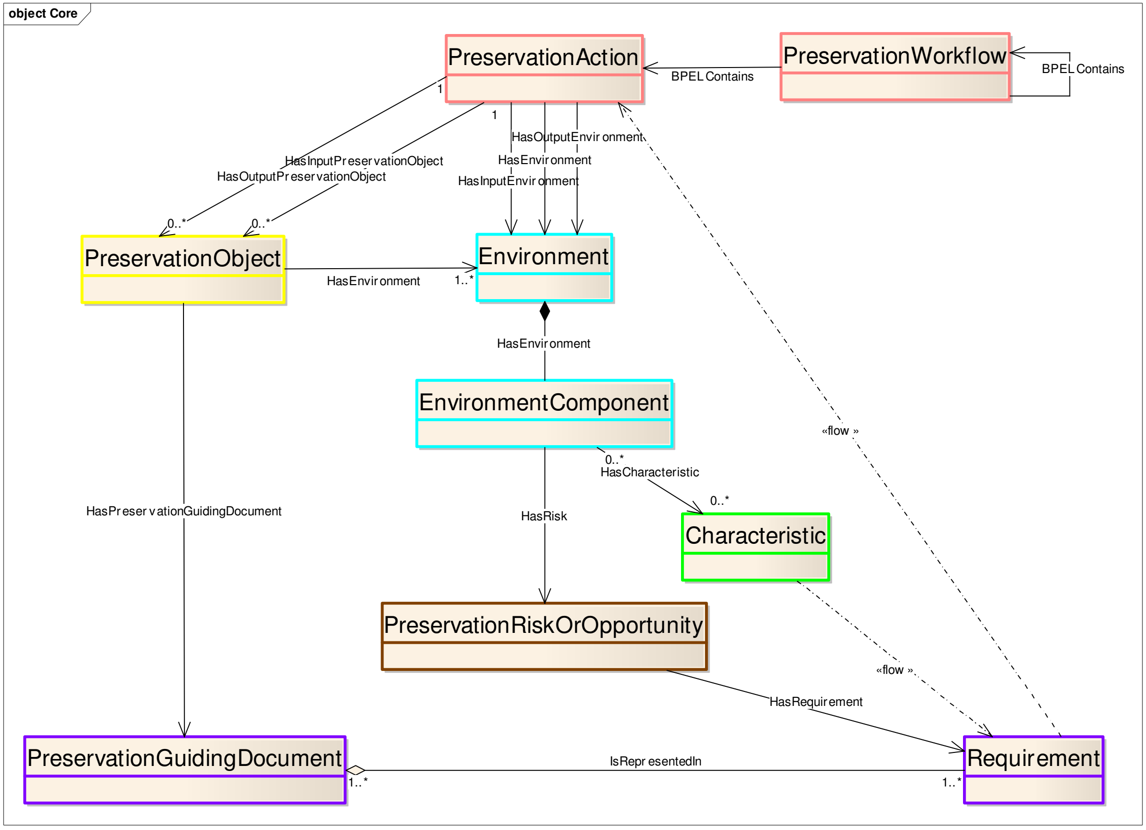


Figure 2 Preservation Planning Conceptual Model

The key conceptual data model in the context of preservation planning is summarised in Figure 2. It shows the concepts and relationships which are explained in detail in the following sections.
In summary:

Any *Preservation Object*, such as a *Collection*, down to an individual *Bytestream* has one or more *Environments*.

Every *Environment* in which the *Preservation Object* is embedded consists of a number of *Environment Components*, such as hardware and software components, the legal system, and other internal and external factors.

Whenever changes occur to an *Environment Component*, such as obsolescence of hardware or software components, decay of data carriers, or changes to the legal framework, this may introduce a *Preservation Risk*.

Preservation Risks are specified in *Risk Specifying Requirements.* Whenever *Characteristics* of a *Preservation Object's Environment Component* take on certain values which are specified in the *Requirement* then the *Preservation Object* is considered at risk.

Once a *Risk Specifying Requirement* is violated, a preservation monitoring process should trigger the preservation planning process. It, in turn, determines the optimal *Preservation Workflow* which should mitigate this *Preservation Risk*. This preservation monitoring process is outside the scope of our model.

A *Preservation Workflow* connects *Preservation Actions* together and may include conditional branches and other control-flow constructs. Planets uses the Business Process Execution Language (BPEL) to describe workflows. How this is done is outside the scope of our model.

When a *Preservation Action* is applied to a *Preservation Object* and its *Environment* then a new copy of the *Preservation Object* and/or a new *Environment* is created in which the *Preservation Risk* is mitigated. Every *Preservation Action*, therefore, does not only have an *Input Preservation Object* and an *Input Environment*, but also an *Output Preservation Object* and an *Output Environment*. For example, if a Microsoft Word *Bytestream* is migrated to an Adobe PDF *Bytestream*, not only do we create a new *Preservation Object*, which might have slightly new *Characteristics*, but we also need to embed it in a new *Environment* in which it can be used – in this case the platform needs to at least contain an Adobe PDF viewer. This approach works equally for migration, emulation, and hardware solutions.

For any given *Preservation Object* and its *Environment* there are multiple possible *Preservation Actions* which might mitigate the *Preservation Risk*. Which of these *Preservation Actions* is the most suitable for the *Preservation Object* can be derived from the information in the *Requirements*. These *Requirements* define

- acceptable *Characteristics* of the *Preservation Action* itself (such as that PDF may, for a given institution, not be an acceptable preservation output format of a *Preservation Action*)
- acceptable output *Characteristics* of the *Preservation Object*, which may be dependent on input *Characteristics* (such as that the size of the *Preservation Action*'s output *Preservation Object* should not exceed a maximal size set by the institution)

Preservation Guiding Documents also contain *Requirement*s which

- describe the preservation process itself independent of the *Characteristics* of the *Preservation Object* as well as of those of the *Preservation Action* (such as that a preservation planning process should be executed for every data object at least every 5 years, independent of the *Preservation Risks* that are established for this data object).
- do not describe the preservation process itself. They are contained in *Non Preservation Requirements*.

5.4 Preservation Guiding Document

5.4.1 Stratml

Strategy Markup Language (StratML) is a basic conceptual model for describing the essential contents of a strategy document. It is envisioned as an ISO standardized XML schema and vocabulary for US Federal agency strategic plans that is aligned with the Federal Enterprise Architecture, government policy, and leverages existing standards. (based on http://www.xml.gov/presentations/gpo/stratml20060118.ppt). We are borrowing most of stratML's basic elements to describe the non-requirement parts of preservation guiding documents.

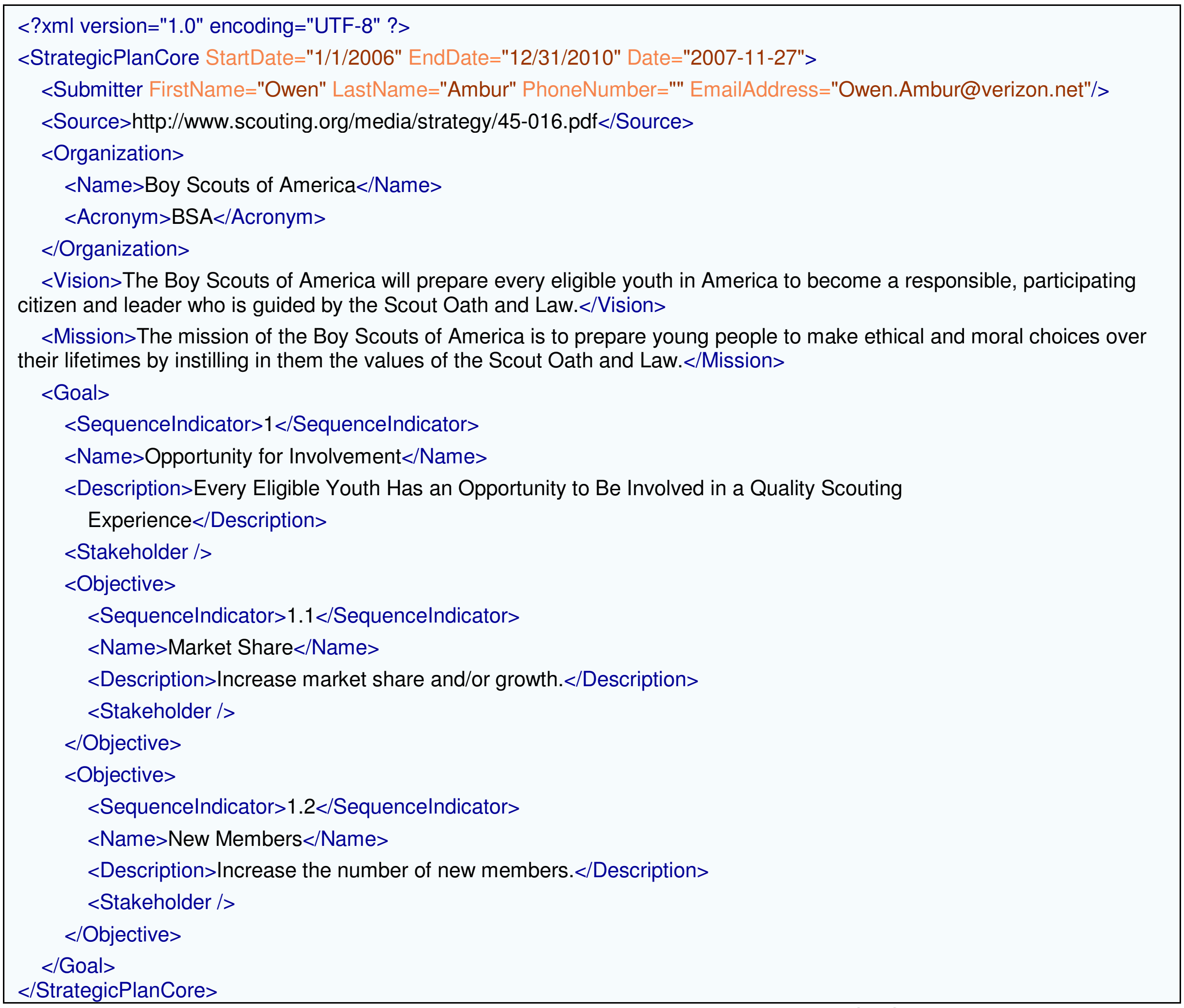
The top-level elements in stratML are as follows:

- *Submitter*: The person submitting the plan. With sub-elements
- *Source*: The Web address (URL) for the authoritative source of this document
- *Organization*: The legal or logical entity to which the report applies.
- *Vision*: Vision statements are distinguished from goals in that they are the focus of constant pursuit but can never be satisfied in the sense of being met or completed.
  A concise and inspirational description of a state the organization will strive to approach over a relatively long span of years but which can ultimately never be fully achieved.
- *Mission*: Mission Statement. A brief description of the basic purpose of the organization. An agency's goals should flow from the mission statement.
- *Value*: A principle that is important and helps to define the essential character of the organization.
- *Goal*: General Goal
  A relatively broad statement of intended results to be achieved over more than one resource allocation and performance measurement cycle.
  Goals define a purpose and direction and take all stakeholders and perceived present and future needs into account. Goals must be capable of being effectively pursued with measurable results over more than one budgetary execution cycle but within the reasonably foreseeable future. Goals should be objective, quantifiable, measurable, and defined at the level to be achieved by a program activity.
  Supports Mission
- *Objective*: Performance Goal.
  A target level of results expressed in units against which achievement is to be measured within a single resource allocation and performance execution cycle.
  Supports Goal.
  Objectives are measurable subsets of goals to be achieved within a given time period with available resources. Objectives provide the day-to-day support for achieving goals.
  Submitter, source, organization, vision, mission and value to be used as in stratML. They can be directly used for automatic preservation planning as they are described in stratML.

The schema definition can be found in http://www.xml.gov/stratml/StrategicPlanCore-ns.xsd.

Within our model, these concepts are used in the following way:

- *stratML:organization* is called Has *Institution*.
- *stratML:Value*, which expresses an (ethical) value of an institution, is different from the "*Planets:Value*", which expresses the *Value* of a *Characteristic* (= assigned or derived value).
- A *stratML:objective* is roughly equivalent to a *Requirement* in Planets. In stratML an objective is represented as a string. In order to support automated preservation planning, however, a refined, machine-interpretable definition of the objective / requirement is needed. This will be developed in the next few sections. In order for the other stratML elements to be used in preservation planning, since they in general express assignments of values, they can be simply looked up and used by preservation planning tools.



```xml
<?xml version="1.0" encoding="UTF-8" ?>
<StrategicPlanCore StartDate="1/1/2006" EndDate="12/31/2010" Date="2007-11-27">
  <Submitter FirstName="Owen" LastName="Ambur" PhoneNumber="" EmailAddress="Owen.Ambur@verizon.net"/>
  <Source>http://www.scouting.org/media/strategy/45-016.pdf</Source>
  <Organization>
    <Name>Boy Scouts of America</Name>
    <Acronym>BSA</Acronym>
  </Organization>
  <Vision>The Boy Scouts of America will prepare every eligible youth in America to become a responsible, participating citizen and leader who is guided by the Scout Oath and Law.</Vision>
  <Mission>The mission of the Boy Scouts of America is to prepare young people to make ethical and moral choices over their lifetimes by instilling in them the values of the Scout Oath and Law.</Mission>
  <Goal>
    <SequenceIndicator>1</SequenceIndicator>
    <Name>Opportunity for Involvement</Name>
    <Description>Every Eligible Youth Has an Opportunity to Be Involved in a Quality Scouting
      Experience</Description>
    <Stakeholder />
    <Objective>
      <SequenceIndicator>1.1</SequenceIndicator>
      <Name>Market Share</Name>
      <Description>Increase market share and/or growth.</Description>
      <Stakeholder />
    </Objective>
    <Objective>
      <SequenceIndicator>1.2</SequenceIndicator>
      <Name>New Members</Name>
      <Description>Increase the number of new members.</Description>
      <Stakeholder />
    </Objective>
  </Goal>
</StrategicPlanCore>
```

Figure 3 An example snippet from http://xml.gov/stratml/BSAStratPlan.xml

5.4.2 Preservation guiding document

Definition of Preservation Guiding Document

Documents, such as policy, strategy, or business documents, as well as applicable legislation, guidelines, rules, or even a choice of temporary runtime parameters during a preservation action. The term “document” should be understood generously to possibly include oral representations, as well as written representations in databases, source code, web sites, etc..

They specify *Requirements*, which are constraints or rules that make the institution’s values or constraints explicit and influence the preservation planning process.

The term goes beyond and refines the notion of “organisational policy and strategy” documents that were originally foreseen as basis for the analysis.

Preservation guiding documents are a subset of institutional documents which

- may have any institutional scope (corporate, departmental, project related, etc.),
- may have any business focus (policy, strategy, mission, process, etc.),
- are relevant to the business process of preservation planning and form an input to the preservation planning process. Preservation plans are the output of a preservation planning process and are not considered preservation guiding documents as used in this report.

Concepts that are found in our model may be found in any of the documents in this space. We are not trying to prescribe to an institution which concepts should be implemented in which sort of document. This has to remain a personal choice of the institution.

Figure 4 explores (un-exhaustively) the space of institutional documents.

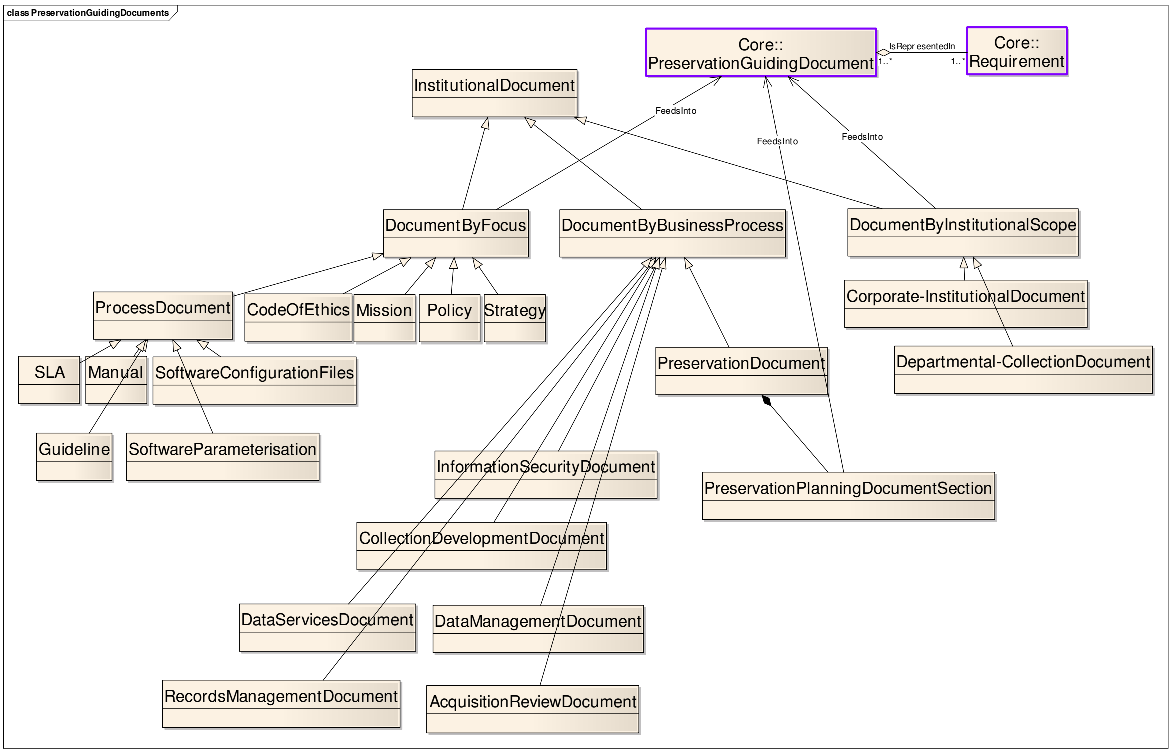

class PreservationGuidingDocuments
Core::
PreservationGuidingDocument
IsRepresentedIn
1..*
1..*
Core::
Requirement
InstitutionalDocument
FeedsInto
FeedsInto
FeedsInto
DocumentByFocus
DocumentByBusinessProcess
DocumentByInstitutionalScope
ProcessDocument
CodeOfEthics
Mission
Policy
Strategy
Corporate-InstitutionalDocument
SLA
Manual
SoftwareConfigurationFiles
PreservationDocument
Department-CollectionDocument
Guideline
SoftwareParameterisation
InformationSecurityDocument
PreservationPlanningDocumentSection
CollectionDevelopmentDocument
DataServicesDocument
DataManagementDocument
RecordsManagementDocument
AcquisitionReviewDocument

Figure 4 Preservation Guiding Documents

Elements of Preservation Guiding Document

- *Document Identification* (mandatory, non-repeatable):
  - *Document Identifier* (mandatory, non-repeatable): a unique identifier of the *Preservation Guiding Document* (data constraint: none)
  - *Document Name* (optional, repeatable): a human readable meaningful descriptor for the *Document* (data constraint: string)

  - *Document Version* (optional, non-repeatable): Version of the *Document* (data constraint: none)
- *Has Institution* (mandatory, non-repeatable): a unique identifier of the institution (data constraint: Institution ID)
- *Document Approval* (optional, repeatable)
  - *Status* (mandatory, non-repeatable): (data constraint: one of *proposed*, *approved*, *superseded*)
  - *Initiator* (optional, repeatable): Person who proposed, approved or withdrew the document. (data constraint: Agent[25] ID) (N.B. This subsumes the *stratML:submitter* element)
  - *Status Date* (mandatory, non-repeatable): Date on which the document was proposed, approved or withdrawn. (N.B. This subsumes the *stratML:Date* attribute)
- *Document Applicability* (mandatory, non-repeatable)
  - *Start Date* (optional, repeatable): The date the document is projected to become valid (data constraint: date)
  - *End Date* (optional, repeatable): The date the document is projected to cease, if it is not subsequently extended (data constraint: date)
- *stratML:Source* (optional, non-repeatable) The Web address (URL) for the authoritative source of this document. (data constraint: anyURI)
- *stratML:Vision* (optional, repeatable): Vision statements are distinguished from goals in that they are the focus of constant pursuit but can never be satisfied in the sense of being met or completed. A concise and inspirational description of a state the organization will strive to approach over a relatively long span of years but which can ultimately never be fully achieved. (data constraint: string)
- *stratML:Mission* (optional, repeatable): Mission Statement. A brief description of the basic purpose of the organization. An agency's goals should flow from the mission statement. (data constraint: string)
- *stratML:Value* (optional, repeatable) A principle that is important and helps to define the essential character of the organization.
  - *stratML:Name*
  - *stratML:Description* (optional, repeatable)
- *stratML:Goal* (mandatory, repeatable)
  - *stratML:SequenceIndicator* (optional, non-repeatable)
  - *stratML:Name* (optional, non-repeatable)
  - *stratML:Description* (mandatory, non-repeatable) (data constraint: Description)
  - *stratML:Stakeholder* (optional, repeatable) (data constraint: Agent ID)
  - ***Has Requirement*** (optional, repeatable): a unique identifier of the *Requirements* included in this document (data constraint: *Requirement* ID)
  - *stratML:OtherInformation* (optional, non-repeatable)
- *Has References* (optional, repeatable)
  - *Has Collection* (optional, repeatable): unique identifiers for each of the institution*'s Collections* (see Section 5.5.1 below) to which the preservation guiding document refers. (data constraint: Collection ID)
  - *Has Registry References* (optional, repeatable): unique identifiers for each of the registries and inventories (see Section 5.10.1 below) to which the preservation guiding document refers (data constraint: Registry ID)

[25] The agent concept is defined in [core]

- *Has Predecessor Document* (optional, repeatable): unique identifiers for each of the predecessor document(s) of the preservation guiding document
- *Has Related Document* (optional, repeatable): unique identifiers for each of other related document(s)

5.5 Preservation Object

Definition of Preservation Object

A *Preservation Object* is any object that is directly or indirectly at risk and needs to be digitally preserved.

Vocabulary for Preservation Object Types

A *Bytestream* is the primary *Preservation Object.* If it is at risk of decay or obsolescence it becomes the object of preservation. We create and execute preservation plans to preserve it.

A *Bytestream* is, however, embedded in a larger context, as illustrated in Figure 5. Since higher-level objects (such as the *Manifestation* that includes the affected *Bytestream*, and the *Collection* in which this *Manifestation* is held) are indirectly affected by its preservation need, they also need to be considered during preservation planning and are, therefore, indirectly *Preservation Objects*. Conversely, an institution can not consider the preservation of each individual data object in isolation. Institutions need to take a global look at all their *Collection*s and resources in order to prioritise their *Preservation Action*s and co-ordinate preservation activity. In order to facilitate this we are devising a model for *Preservation Guiding Documents* as a basis for preservation planning, which goes well beyond planning for the individual data object.

These *Preservation Object* Types are introduced in te Planets Core model [8] and will be completely aligned with it in the next iteration of this work.

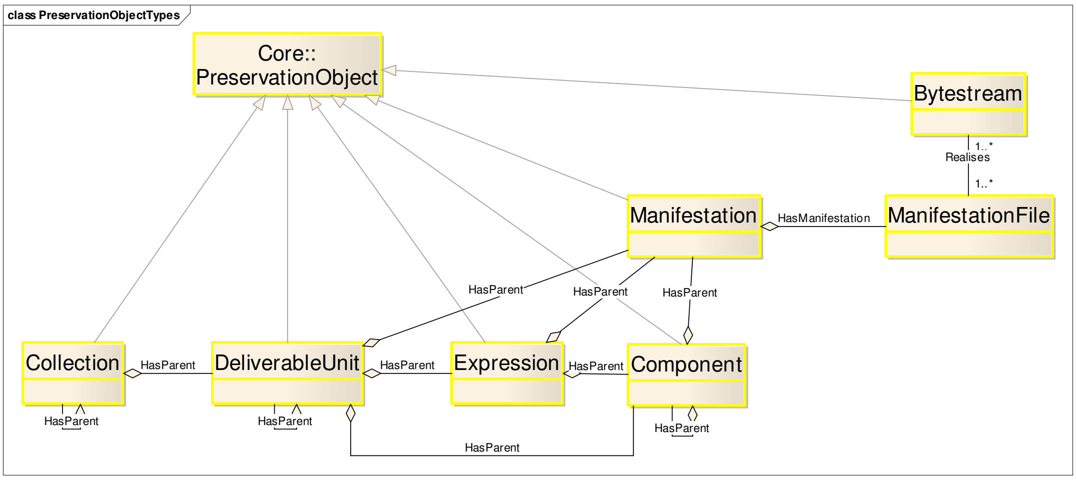


Figure 5 Vocabulary for Preservation Object Types

<u>Vocabulary for *Preservation Object* Types</u>:

Preservation Object Types are *Collection, Deliverable Unit, Expression, Component, Manifestation, Bytestream.*

<u>Examples</u>:

- A digital file (*Bytestream*) is part of its *Manifestation* (e.g. a MPEG-4 video *Bytestream* is part of an HTML *Manifestation* of an article).
- This *Manifestation* represents an *Expression* of this article which contains a video stream. Other *Expressions*, such as a still image *Expression* of the article, might hold an image instead of the video stream.
- All Expressions of this article make up the *Deliverable Unit*. The *Deliverable Unit* is the abstract concept representing the distinct intellectual creation, which is the article. There might be several *Expressions* with several *Manifestation*s of the same article (e.g. an HTML, a PDF, an XML, a publisher specific format).
- The article is part of another *Deliverable Unit*, the issue. (Hence the recursive link in Figure 5)
- And the issue is part of the *Deliverable Unit* journal, which is the abstract concept describing all issues of the same title.

- The journal belongs to a *Collection*. The *Collection* might be static for the institution, such as the Science Collection, or it might be determined dynamically, such as the *Collection* of all articles that contain TIFF3.0 files. *Collection* s may contain digital and non-digital objects.
- *Collection* s may be recursively contained in larger *Collections*.
- Finally, all *Collections* are part of the whole institution, which is modelled as the top-level *Collection*.

In addition there are *Component*s of a *Deliverable Unit* which have *Characteristic*s of their own.

Examples:

A "text string" *Component* or a "title" *Component* of a journal article.

► Most of these *Preservation Object Types* are related to others in an aggregate relationship. Since they are all affected by *Preservation Actions*, each of them is considered a *Preservation Object Type.*

► *Collection*, *Deliverable Unit*, *Expression*, and *Component* are abstract descriptions of logical objects.

Manifestation and *Manifestation File* are abstract descriptions of physical objects.

Bytestreams are physical objects.

Elements of Preservation Object

- *Preservation Object Identifier* (mandatory, non-repeatable): a unique identifier of the object (data constraint: Preservation Object ID)
- *Preservation Object Name* (optional, repeatable): a human readable meaningful descriptor for the *Preservation Object* (data constraint: string)
- *Preservation Object Type* (mandatory, non-repeatable): a type specification of the *Preservation Risk* (data constraint: Preservation Object Type)
- *Preservation Object Description* (optional, repeatable): a human readable meaningful description for the *Preservation Object* (data constraint: Description)
- *Has Constraint* (optional, repeatable): *Characteristic*(*s*) of *Environment Components* to constrain the *Preservation Object* or to specify aggregates (e.g. a dynamic *Collection* consisting of all PDF *Bytestreams*, a dynamic *Collection* consisting of all static *Collections* containing *Bytestreams* older than 20 years, a *Deliverable Unit* consisting of the *Expressions* that contain audio, etc.) (data constraint: logical constraint)
- *Has Parent* (mandatory, non-repeatable (this implies a tree structure)): a unique identifier of the parent object (data constraint: Preservation Object ID)
- *Has Environment* (optional, repeatable): unique identifiers to each of the *Preservation Object*'s *Environment*s (data constraint: *Environment* ID)
- *Has Preservation Guiding Document* (optional, repeatable): unique identifiers to each of the *Preservation Object*'s *Preservation Guiding Documents* (data constraint: Document ID)
- *Has Institution* (optional, repeatable): unique identifiers to each of the *Preservation Object*'s institutions (data constraint: Institution ID)
- *Has Rights* (optional, repeatable): unique identifiers to each of the *Preservation Object*'s *Rights* objects (data constraint: Rights ID[26])
- *Has Event* (optional, repeatable): unique identifiers to each of the *Preservation Object*'s *Event* objects (data constraint: Event ID) [27]

Other relationships of Preservation Object

[26] Rights and Event concepts are defined in the Planets Core Conceptual model [Core]. They are not repeated in this model.

[27] For all events the following holds: Whether recording a certain event is mandatory, and which event to record is a business requirement of the institution. It is not made mandatory by the data model.

- *Preservation Action* has a *Has Input Preservation Object* and a *Has Output Preservation Object* relationship with *Preservation Object.*
- *Collection*, *Deliverable Unit*, *Expression*, *Component*, *Manifestation* and *Bytestream* have a subclass relationship with *Preservation Object*.

5.5.1 Collection

Definition of Collection

A grouping of *Deliverable Units* to be processed or kept together.

► A *Collection* can be statically defined by the institution or dynamically defined at a given time by a conditional description (e.g. all files older than x, all objects larger than y, all objects on a given data carrier type) using the *Has Constraint* element. It can be technically homogenous (e.g. one file-format), or consist of different types of objects or file formats.

Elements of Collection

- Elements inherited from *Preservation Object*
  - The *Has Parent* relationship is a reference to a parent *Collection* or may be nil, since *Collection* is the top-level *Preservation Object* (data constraint: Collection ID)
- Collection description elements as defined by Planets PP/3 and PP/6

Other relationships with Collection

- *Collection* is a *subclass* to *Preservation Object*.
- *Deliverable Unit* has a *Has Parent* relationship with *Collection.*

5.5.2 Deliverable unit

Definition of Deliverable Unit

A *Deliverable Unit* is a distinct intellectual creation.

This definition is meant to include artistic creations.

A *Deliverable Unit* is an abstract concept, which does not prescribe its physical realization, and may have many *Expressions*.

► In general, a *Deliverable Unit*, as the central logical object of preservation, will be a (set of) content item(s) held by the institution (e.g. a web site, a web page, a journal title, a journal issue, an article)[28]. It may however also be a derived intellectual object which the institution would like to have preserved (e.g. metadata, schemas, full-text indices which need to be preserved together with the content, a description of the software and hardware *Environment* on which to access the content).

Elements of Deliverable Unit

- Elements inherited from *Preservation Object*
  - The *Has Parent* relationship is a reference either to a parent *Deliverable Unit* or a parent *Collection.* (data constraint: Collection or Deliverable Unit ID)

Other relationships with Deliverable Unit

- *Deliverable Unit* is a *subclass* to *Preservation Object*.
- *Component* has a *Has Parent* relationship with *Deliverable Unit.*
- Expression has a *Has Parent* relationship with *Deliverable Unit.*
- Manifestation has an *aggregate* relationship with *Deliverable Unit.*

5.5.3 Expression

A *Deliverable Unit*, such as an article, may have several *Expressions*. An HTML *Manifestation* of the article for example might include a video stream. This video stream could not be present in

[28] Its metadata is kept in the set of *Characteristics* associated with the *Deliverable Unit*.

static *Manifestation*s, such as a PDF. It might be replaced by an image in those *Manifestations*. This article *Deliverable Unit* would therefore have two *Expressions*, one with video stream and one with an image.

Definition of Expression

An *Expression* is the specific intellectual or artistic form that a *Deliverable Unit* takes as it is realized. It is, however, a conceptual, not a physical realization.

► Expression encompasses, for example, realization in different languages, realization as a still or a moving image, the particular phrasing resulting from the realization of a musical work. The term *Expression* is here borrowed from FRBR[29] but, unlike FRBR, does not necessarily exclude aspects of physical form, such as typeface and page layout, if they are considered integral to the intellectual or artistic realization of the *Deliverable Unit*.

► In the Planets model *Expressions* are optional for a given instantiation of the model. If, for a given institution, *Deliverable Units* contain several *Expressions* then the model instantiation should contain an *Expression* concept. If all *Manifestations* of a *Deliverable Unit* contain exactly the same significant *Components* then the *Expression* concept may be omitted from a given institution's model. This property is reflected in the direct *aggregate* link between *Component* and *Deliverable Unit* which skips the *Expression* concept.

Elements of Expression

- Elements inherited from *Preservation Object*
  - The *Has Parent* relationship is a reference to its parent *Deliverable Unit* (data constraint: Deliverable Unit ID)

Other relationships with Expression

- *Expression* is a *subclass* to *Preservation Object*.
- *Component* has an *aggregate* relationship with *Expression*
- *Manifestation* has an *aggregate* relationship with *Expression*

5.5.4 Component

Definition of Component

A part of the whole of an *Expression* (or of a *Deliverable Unit,* if *Expressions* are omitted) for which *Values* for *Characteristics* can be measured.

Example:

A "text string", "footnote" or "abstract" *Component* in a journal article.

Vocabulary for Component Types

Vocabulary for *Component Types* (such as header, body, footer / title, abstract, appendix / sub-string, table) is being developed in preservation characterisation research. For text-based systems the vocabulary to specify the *Component Types* can, for example, be taken from the NLM DTD[30] which uses tags for mark-up of journal article components. Other component types can be defined for other content-type specific needs, such as sound, video, etc..

Elements of Component

- Elements inherited from *Preservation Object.*

[29] IFLA Study Group on the Functional Requirements for Bibliographic Records. Functional requirements for bibliographic records : final report. München: K.G. Saur, 1998. (UBCIM publications ; new series, vol. 19). ISBN 3-598-11382-X.

[30] National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM). *Archiving and Interchange Tag Set*. http://dtd.nlm.nih.gov/

  - The *Has Parent* relationship is a reference to a parent *Expression* (or *to* a parent *Deliverable Unit,* if *Expressions* are omitted) or to a parent *Component.* (data constraint: Expression or Deliverable Unit or Component ID)
  - One *Has Event* relationship is a reference to a Component Discovery Event.

- *Component Type* (mandatory, repeatable): type specification of the *Component* (data constraint: extensible vocabulary: taken from the specific vocabulary for Component Types).

Other relationships with Component

- *Component* is a *subclass* to *Preservation Object.*
- *Manifestation* has an *aggregate* relationship with *Component.*

5.5.5 Manifestation

Definition of Manifestation

The physical embodiment of an *Expression* (or of a *Deliverable Unit,* if *Expressions* are omitted) or *Component.*

► *Expressions* (or *Deliverable Units,* if *Expressions* are omitted) or Components may have multiple *Manifestations*. For example a journal article may come both in .doc format and as an .XML document with associated files. Any set of files that allows authentic rendering of the *Expression* within its technical *Environment* is a *Manifestation* of the *Expression*.

► In the Planets model, *Deliverable Units* or *Expressions* are not contained in *Manifestations*. If, for example a CD contains several songs, then the CD as a whole may be a *Deliverable Unit*, and each of the songs may be a separate *Deliverable Unit*, which is contained in the CD *Deliverable Unit*. The CD, as well as each song, may have one or more *Manifestations* of itself. In FRBR this may be represented as the CD *Manifestation* containing several song *Expressions*. This nesting is not allowed using Planets relationship links.

Elements of Manifestation

- Elements inherited from *Preservation Object*
  - The *Has Parent* relationship is a reference either to the *Expression* (or *Deliverable Unit,* if *Expressions* are omitted) and *Component* for which the *Manifestation* serves as physical embodiment. (data constraint: Expression or Deliverable Unit or Component ID)

Other relationships with Manifestation

- *Manifestation* is a *subclass* to *Preservation Object*.
- *Manifestation File* has an *Has Manifestation* relationship with *Manifestation*

5.5.6 Manifestation file

Definition of Manifestation Files

A Digital File that is associated with a Manifestation.

Elements of Manifestation Files

- Elements taken from the Planets Core model[8]
- *Has Manifestation* (mandatory, repeatable): unique identifier to each of the *Manifestation File*'s *Manifestations* (data constraint: Manifestation ID)

Other relationships with Manifestation Files

- *Manifestation File* has an *aggregate* relationship with *Manifestation*
- *Manifestation File* has a *realises* relationship with *Bytestream*

5.5.7 Bytestream

Definition of Bytestream

An ordered sequence of bytes.

► It can be a digital file or an embedded byte-stream within a digital file.

Elements of Bytestream

- Elements inherited from *Preservation Object*
    - The *Has Parent* relationship may be set to nil*.*
- *Realises* (mandatory, non-repeatable): unique identifier to each of the *Bytestream* which is realised by the *Manifestation File* (data constraint: Bytestream ID)

Other relationships with Bytestream

- *Bytestream* is a *subclass* to *Preservation Object*.

5.6 Environment

Definition of Environment

The set of factors which constrain a *Preservation Object* and that are necessary to interpret it.

► Every *Preservation Object* has one or more *Environments* which may be fulfilling different roles. For example, a *Bytestream* or a *Manifestation* object may have creation, ingest, preservation, and access *Environments*; a *Collection* may have an internal, a physical delivery, and an online delivery *Environment*.

► *Environments* for *Preservation Object*s at a higher level also apply to *Preservation Objects* at a lower level. But lower level *Preservation Objects* may have additional *Environment Components* or *Characteristics*.

Therefore, the *Environment* for a *Bytestream*, for example, can be different from the *Environment* of the *Manifestation* to which it belongs. As long as the *Bytestream* is part of its *Manifestation*, it will live in the *Manifestation's Environment*. When it is taken out of the *Manifestation*'s *Environment*, for example to be used in a migration, then the *Bytestream*'s individual *Environment* will influence the *Environment* of its new *Manifestation*.

► It is also worth noting, that it is not necessarily possible to derive the best *Environment* from a *Bytestream*'s file format. If, for example, a *Bytestream* does not make use of the full range of features of the file format then it may be supported by an *Environment*, which in general might not support all *Bytestreams* of its file format. Institution*s* may wish to specify the *Environment* together with their intentions (necessary, recommended, acceptable…)

Elements of Environment

- *Environment Identifier* (mandatory, non-repeatable): a unique identifier of the *Environment* (data constraint: *Environment* ID)
- *Environment Role* (optional, repeatable): (data constraint: extensible vocabulary: one of *creation*, *ingest*, *preservation*, *access*, …)
- *Environment Intention* (optional, repeatable): (data constraint: extensible vocabulary: one of *necessary*, *recommended*, *acceptable*…)
- *Has Environment Component* (optional, repeatable): unique identifiers to each of the *Environment Component* objects (data constraint: *Environment Component* ID)
- *Has Preservation Object* (optional, repeatable): a unique identifier of the *Preservation Object* to which the *Environment* belongs; (data constraint: *Preservation Object* ID) (This optional relationship is also established via the *Has Environment* relationship which leads from the *Preservation Object* to the *Environment*).
- *Has Event* (optional, repeatable): unique identifiers to each of the *Environment's Event* objects (data constraint: Event ID)

Other relationships with Environment

- *Preservation Object* has a *Has Environment* relationship with *Environment.*
- *Preservation Action* has a *Has Input Environment* relationship and a *Has Output Environment* relationship with *Environment.*
- *Environment Component* has a *Has Environment* relationship with *Environment.*

5.7 Environment Component

Definition of Environment Component

A factor which constrains a *Preservation Object* and that is necessary to interpret it.

Example:

Every *Preservation Object* is embedded, in an *Environment* which consists of a number of *Environment Components*, such as data, software, hardware, and community or other internal and external *Environmenta*l factors, such as legal or budget restrictions.

► Which candidate *Preservation Action* is chosen may depend on the *Characteristics* of these *Environment Components* and the *Characteristics* which the output *Environment Component* would have if the given candidate *Preservation Action* was to be executed.

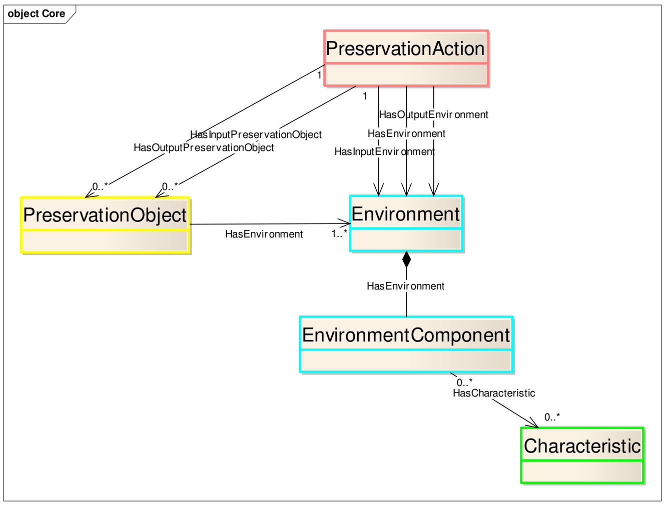


Figure 6 Environment and Environment Component

Elements of Environment Component

- *Environment Component Identifier* (mandatory, non-repeatable): a unique identifier of the *Environment Component* (data constraint: *Environment Component* ID)
- *Environment Component Type* (mandatory, non-repeatable): a type specification of the *Environment Component* (data constraint: extensible vocabulary: taken from the specific vocabulary for *Environment Component* Types).
- *Has Preservation Object Type* (mandatory, non-repeatable): a type specification of the *Preservation Object Type* for which this *Environment Component* stands (data constraint: Preservation Object Type).
- *Has Characteristic* (optional, repeatable): unique identifiers of each of the *Characteristics* of the *Environment Component* (data constraint: Characteristic ID). Every *Environment Component* has one or more *Characteristics* with associated *Values* which may influence the choice of *Preservation Action*.
- *Has Environment* (optional, repeatable): unique identifiers of each of the *Environments* to which the *Environment Component* belongs; (data constraint: *Environment* ID).
  (This relationship is also established via the *Has Environment Component relationship* which leads from the *Environment* to the *Environment Component*).
- *Has Risk* (optional, repeatable): unique identifiers of each of the *Preservation Risks* which arise as the *Environment Component*'s *Characteristics* violate a *Risk Specifying Requirement* (data constraint: Preservation Risk ID).
- *Has Event* (optional, repeatable): unique identifiers to each of the *Environment Component's Event* objects (data constraint: Event ID)

Vocabulary for Environment Component Types

The top-level vocabulary to specify the *Environment Component Type* can be taken from Figure 7. Lower-level vocabulary is specified in Figure 8, Figure 9, Figure 10, and Figure 11. They can be extended according to institution-type specific needs.

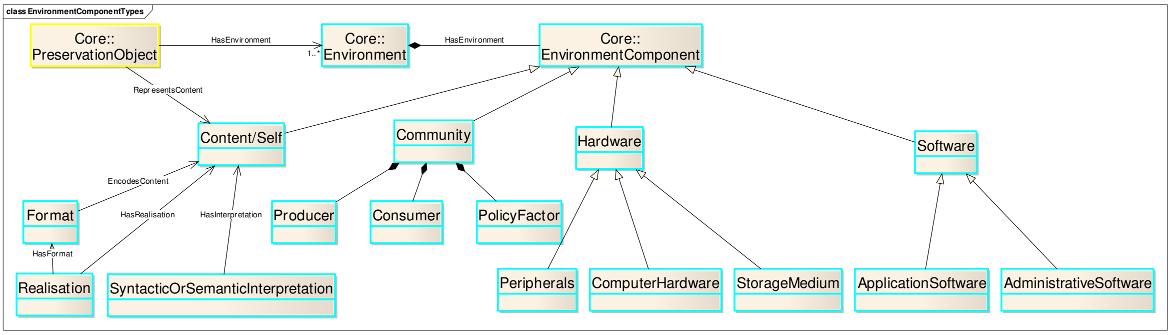


Figure 7 Top-level vocabulary for Environment Component types

The top-level vocabulary includes software, hardware, community and content/self. The latter deserves the most explanation. We have chosen to represent the factors that make up the *Preservation Object* as part of its *Environment*, rather than as part of the *Preservation Object* itself. Instead we treat *Preservation Object* as just an abstract concept. Its components, that is the intellectual content of the *Preservation Object*, the semantic and syntactic interpretation of the content which are necessary to interpret the content, the format in which the content is encoded, and the physical realisation of the content, are considered part of the object's *Environment*. They can then be treated like other *Environment Components* with their associated *Characteristics* and *Values* and be used in the preservation planning process in a uniform way.

These latter *Environment Component Types* mirror the OAIS[31] representation information concepts. We have chosen to express the OAIS "Digital object" as the duality of intellectual content "Content/Self" and its physical "Realisation".

<u>Examples for two types of *Preservation Objects*</u>:

For a *Bytestream*

- the content is the intellectual content of the Bytestream
- the semantic and syntactic interpretation specifies how to interpret the content (e.g. what language, what grammar, what units, the semantic meaning (e.g. average vs. maximum temperature))
- the format is the file format type
- the realisation is actual bit sequence.

For a *Collection*

- the content is the intellectual content of the Collection as a whole
- the semantic and syntactic interpretation would specify how to interpret the Collection (e.g. the filing / shelving / recording system for the items of a *Collection*, how the parts of the *Collection* are identified, sorted, accessed)
- the format is the way in which the Collection is physically stored
- the realisation is the physical realisation of the Collection

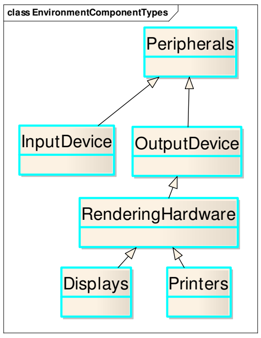


Figure 8 Vocabulary for peripheral types for Environment Components

[31] Reference Model for an Open Archival Information System (OAIS) http://public.ccsds.org/publications/archive/650x0b1.pdf

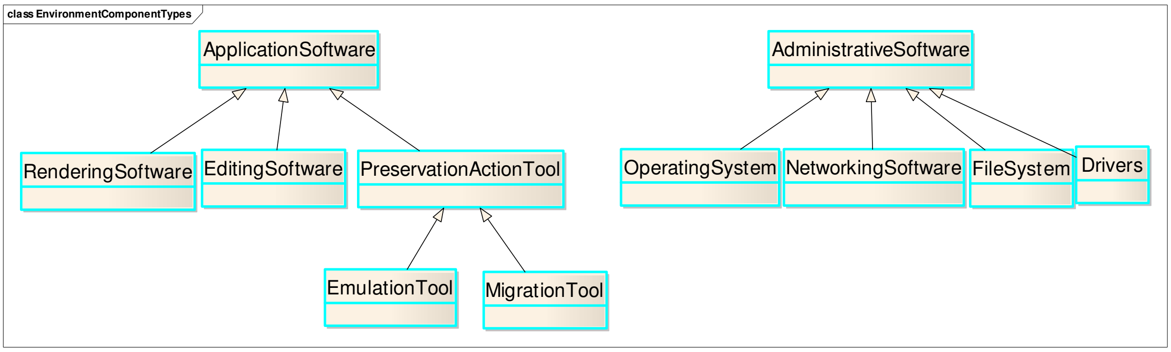


Figure 9 Vocabulary for application and administrative software

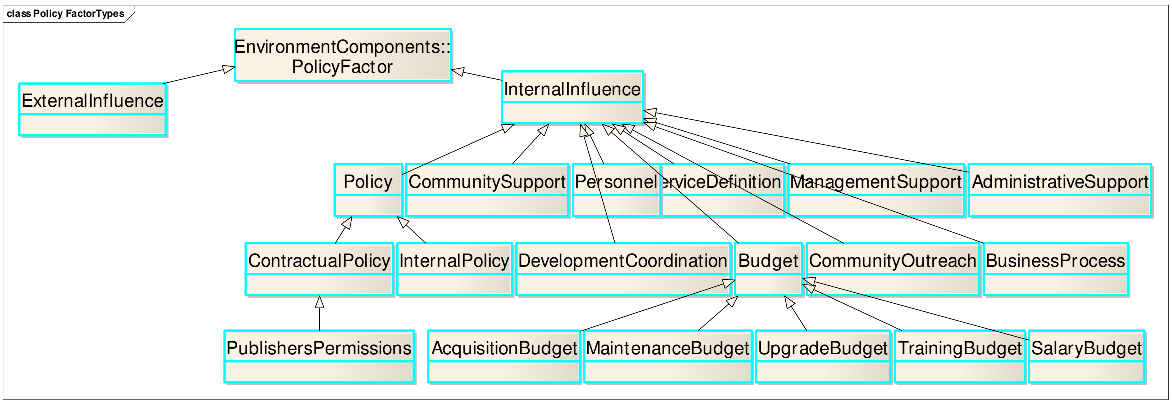


Figure 10 Vocabulary for internal influences for Environment Components

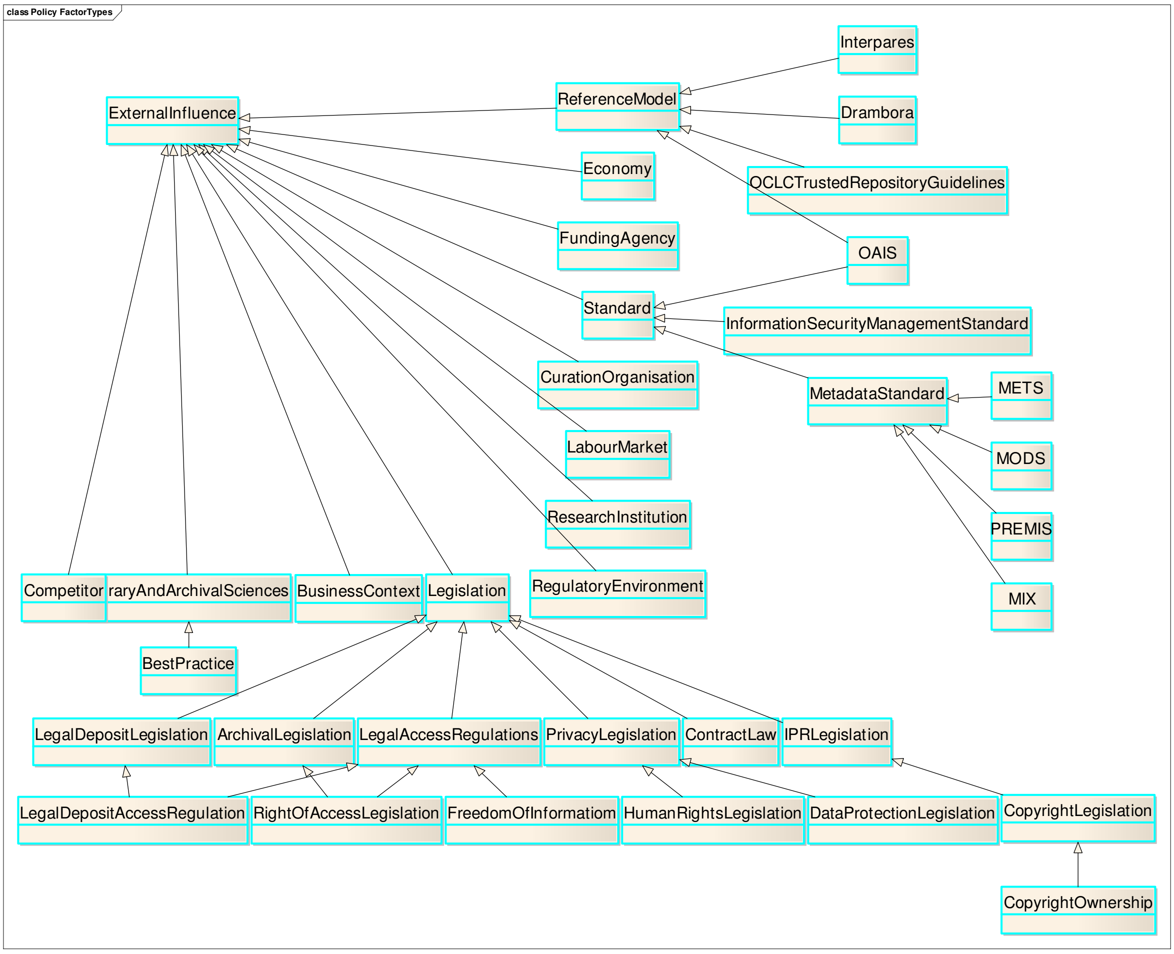


Figure 11 Vocabulary for external influences for Environment Components

5.8 Preservation Risk

► Preservation planning is about mitigating risks to access of digital objects or about taking advantage of opportunities for improvement through *Preservation Action*s.

► Specific *Preservation Risks* are associated with specific *Environment Components* of a *Preservation Object*.

Examples of *Preservation Risk* include:

- Data carriers deteriorate and cannot be read.
- The data object becomes corrupted on the carrier and the original byte stream cannot be retrieved.
- Essential hardware components are no longer supported or available.
- Software components are proprietary and the dependence is unacceptable to the institution.
- The community requires new patterns of access, such as access on a mobile phone, rather than a workstation.
- File formats (called *BytestreamFormat* in the Planets Core model[8]) become obsolete.
- The legislative framework changes and the data or access to it has to be adapted to the new regulations.

Examples of *Preservation Opportunities* include:

- Adding new features, such as interactivity, provides new usage opportunities.
- Maintaining data becomes cheaper by moving to newer formats.
- Consolidation of support structures (e.g. software or hardware *Environment*s) streamlines the maintenance of the Collection.

In the remainder of this paper when we talk about *Preservation Risks* we implicitly include *Preservation Opportunities*.

► These risks are not always inherent, but are relative to considerations such as the institution's goals and the *Characteristics* of individual *Preservation Objects*.

Examples:

- Depending on the institution's goals: One institution might find using proprietary software acceptable, another might not, and, therefore, does or does not consider it a *Preservation Risk*
- Depending on the digital object's individual *Characteristic*s: The digital object uses, or does not use macros and, therefore, is or is not subject to a *Preservation Risk*.

Each institution must therefore specify in *Risk Specifying Requirements* which state of the *Preservation Object's Environment* represents a *Preservation Risk*. We introduce parameterised *Requirement*s which can be instantiated specifically to each institution's needs.

Definition of Preservation Risk

A *Preservation Risk arises* when a *Characteristic* of an *Environment Component* of a *Preservation Object* conflicts with the institution's *Risk Specifying Requirements*.

Elements of Preservation Risk

- *Risk Identifier* (mandatory, non-repeatable): a unique identifier of the *Preservation Risk* (data constraint: Preservation Risk ID)
- *Risk Name* (optional, repeatable): a human readable meaningful descriptor for the *Preservation Risk* (data constraint: string)
- *Risk Type* (optional, repeatable): a type specification of the *Preservation Risk* (data constraint: extensible vocabulary: taken from the specific vocabulary for *Preservation Risk* Types).

- *Has Environment Component* (optional, non-repeatable): a unique identifier of the input *Environment Component that is at risk* (data constraint: *Environment Component* Type) (This optional relationship is also established via the *Has Risk* relationship which leads from the *Environment Component* to *Preservation Risk*).
- *Has Risk Specifying Requirement* (mandatory, non-repeatable): a unique identifier of the *Requirement* which is violated by the *Preservation Risk* (data constraint: Risk Specifying Requirement ID).
- *Has Event* (optional, repeatable): unique identifiers to each of the *Preservation Risk's Event* objects (data constraint: Event ID)

Other relationships with Preservation Risk

- The *Environment Component* object has a *Has Risk* association link to the *Preservation Risk* object.

Vocabulary for Preservation Risk Types

Preservation Risk Types are (see Figure 12):

NewVersion: A new version of the *Environment Component* is available. This creates a risk of future obsolescence, or a risk of having to support too many versions of this *Environment Component*.

NotSupportedOrObsoleteSupport: The *Environment Component* is no longer sufficiently supported. This creates a risk that support will cease altogether, rendering the *Environment Component* non-functional.

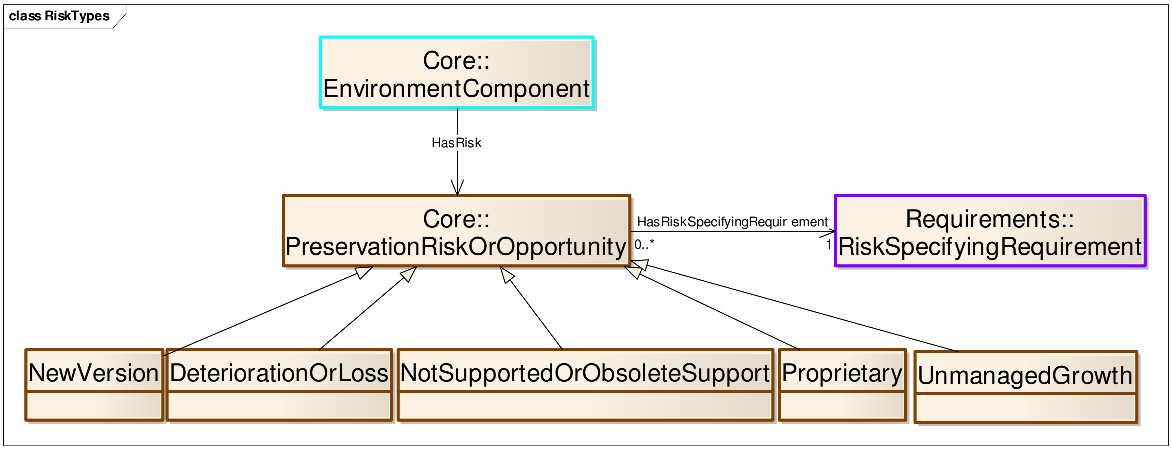


Figure 12 Vocabulary for Preservation Risk Types

DeteriorationOrLoss: The *Environment Component* is deteriorating or has been lost. Reconstruction or replacement become necessary.

Proprietary: The *Environment Component* is proprietary. There is a risk that it cannot be replaced since the specifications for it are unknown.

UnmangedGrowth: The institution's *Environment* is becoming too diverse to manage. A normalisation *Preservation Action* is needed to simplify or unify the *Environment.*

► These risk types obviously apply to technological *Environment Components*. But they also apply to community *Environment Components*. If, for example, consumers request changed services (i.e. considers existing services obsolete) then this may prompt the need for executing a *Preservation Action* which brings the services up to date.

5.9 Preservation Action and Preservation Workflow

Preservation Actions are included in the model since many *Requirements* in *Preservation Guiding Documents* refer to desired *Characteristics* of permissible *Preservation Actions*.

Definition of Preservation Action

In the Planets glossary *Preservation Action* is currently defined in the following way:

> A non-destructive action that creates new data from existing data in the archive, with the intent of preserving or increasing access to information stored in the archive

The following is an older Planets definition:

> The execution of an action to ensure the continued accessibility of a digital object across time and changing technical *Environments* and the preservation of its critical significant properties that transforms the digital object itself, the technical *Environment* required to support access to the object, or a combination thereof.

The newer definition shows a shift of focus within Planets toward *Preservation Actions* on data related actions rather than hardware related actions. The PP2 model uses the more general older approach (which encompasses the newer Planets definition).

► In order to describe any but the simplest preservation plans, actions need to be composed into workflows. (See Figure 13)

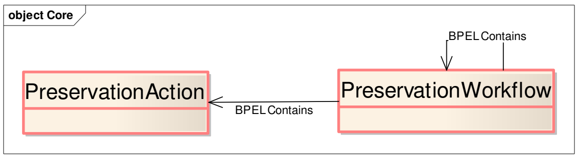


Figure 13 Preservation Action and Preservation Workflow

Definition of Preservation Workflow

> A *Preservation Workflow* connects *Preservation Actions* together and may include conditional branches and other control-flow constructs. Planets uses the Business Process Execution Language (BPEL) to describe workflows.

► BPEL provides a clear and unambiguous language to describe workflows. A BPEL execution engine also allows some workflows to be automatically executed.

► How *Preservation Actions* are composed into *Preservation Workflows* is outside the scope of this report.

► Corresponding to every *Preservation Risk Type* and the type of the affected *Environment Component* that needs to be addressed, there are appropriate *Preservation Actions* to mitigate the risk.

Examples: The risk of data carrier failure can be mitigated by a carrier refresh. The risk of file format obsolescence can be mitigated by migrating objects to an alternative format.

Figure 14 shows some elementary example *Preservation Actions*.

Most of them are self-explanatory. Some deserve some special comments:

- Modification of Content/Self might represent an action such as the reconstruction of a deteriorated file, or a file that is modified in order to satisfy new legal requirements.
- One possible *Preservation Action* is to not do anything (wait and see).
- Migration does not always imply that a different file format is chosen. One might, for example replace an XML file with another XML file. In that case the input and output file formats happen to be the same. The output *Preservation Object* might nonetheless have different *Characteristics* to the input *Preservation Object* because of the different information captured within the XML tags.
- The needs of the target community might be a deciding factor for the choice of *Preservation Actions*, and, conversely, the choice of *Preservation Actions* will shape and change the community, just as it changes the other *Environment Components*.

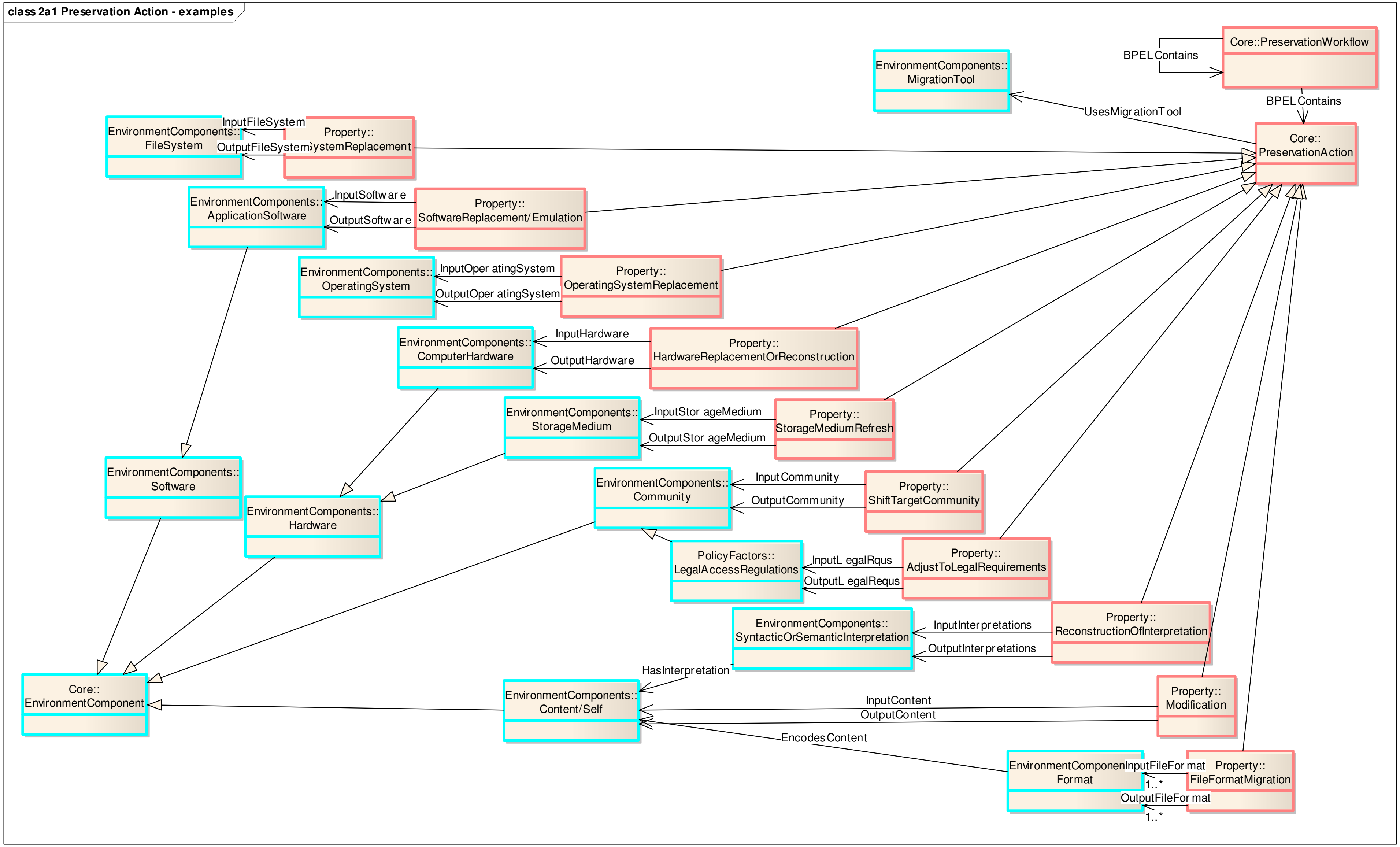


Figure 14 Example Preservation Actions (simplified representation)

- Community consists of producers and consumers. Both types are either technical (e.g. repository or IT staff, publishing staff) or content oriented (authors or readers) and will consider the digital object obsolete under different circumstances and according to their needs.
- Shifting the target community might be a somewhat unintuitive *Preservation Action*, which is parallel to all other forms of *Environment* replacement. An example might be turning a research data centre into a history-of-science repository, as the material contained in the collection seizes to live up to contemporary standards of scientific use.

Vocabulary for Preservation Action Types

A *Preservation Action* may result in the replacement or the repair or reconstruction of any of the *Environment Components* that are at risk. These are the *Preservation Action Types*.

► *Preservation Actions* can be classified into *Preservation Action Classes* by the combination of *Preservation Object Type, Environment Component Type, Preservation Risk Type*, and *Preservation Action Type*. Rather than introduce a specific vocabulary for every *Preservation Action Class*, one may use these elements to describe the class of *Preservation Action*. Preservation Action classes may suitably be described in a registry.

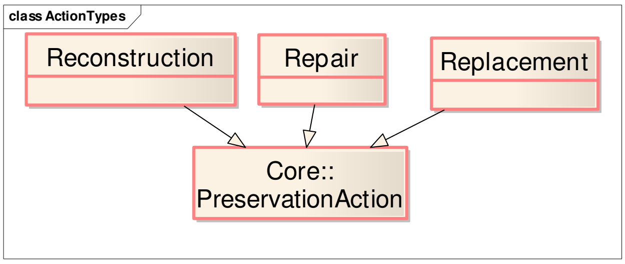


Figure 15 Vocabulary for Preservation Action Types

<u>Example</u>	*Preservation Object Type*	*Environment Component Type*	*Preservation Risk Type* (new version, not supported / obsolete, deterioration / loss, proprietary)	*Preservation Action Type* (reconstruction, repair, replacement)
Data carriers deteriorate and cannot be read	Bytestream	Data Carrier	Deterioration	Replacement
The data object becomes corrupted on the carrier and the original byte stream cannot be retrieved.	Bytestream	Realisation	Deterioration	Reconstruction
Essential hardware components are no longer supported or available	Collection	Hardware	Not supported	Replacement
Software components are proprietary and the dependence is unacceptable to the institution.	Collection	Software	Proprietary	Replacement
The community requires new patterns of access, such as access on a mobile phone, rather than a workstation	Collection	Hardware and Software	Obsolete	Replacement
File formats become obsolete.	Bytestream	Format	Obsolete	Replacement
The legislative framework changes and the data or access to it	Collection	Legislation	New Version	Replacement

has to be adapted to the new regulations				

Table 3 Preservation Action Classes

► A *Preservation Action* produces a changed version of the *Preservation Object* and/or its *Environment Component*. The model, therefore, contains an input and output *Preservation Object* and an input and output *Environment* for a candidate *Preservation Action*.

Examples:

In the case of file reconstruction there is an input and output *Bytestream* while the *Environment* may stay the same.

In the case of migration there is an input and output *Manifestation*. The input and output *Manifestations* may need different *Environments*.

In the case of data carrier refresh the input and output *Bytestreams* are the same, but the *Environment* is new.

A *Preservation Action* produces a new *Preservation Object*, if the *Preservation Object*'s *Environment Components* in its associated output *Environment* change during the *Preservation Action*, i.e. if the intellectual content of the *Preservation Object*, the semantic and syntactic interpretation of the content which are necessary to interpret the content, the format in which the content is encoded, and the physical realisation of the content change.

Example: In the case of file reconstruction there is an input and output *Bytestream* since the *Realisation* of the *Bytestream* changes. If the Bytestream is part of a *Manifestation* , then there will also be a new output *Manifestation* object, or possibly even a new *Expression* if *Characteristics* change sufficiently.

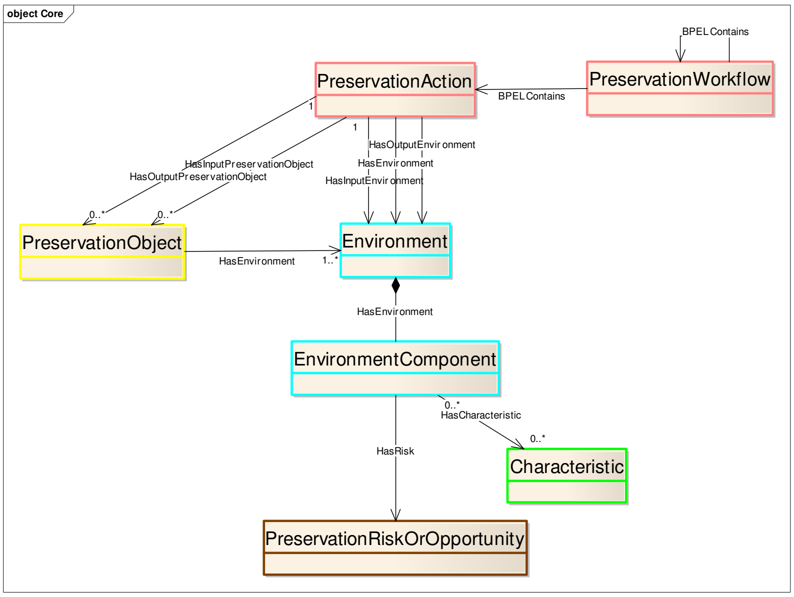


Figure 16 Preservation Actions

► In general a *Preservation Action* may result in the replacement or repair or reconstruction of a combination of *Environment Components*.

Example: Emulation can be seen as a combination of hardware, software and file format replacement, since it provides a new hardware and/or software *Environment* for the digital object, but it might also be necessary to extract data from the original digital object to feed into the emulation.

► A *Preservation Action* always applies to one input and output *Preservation Object*. This *Preservation Object*, however, may consist of several components.

Example:

- Several input components: When migrating an XML Manifestation to a PDF Manifestation, the input Manifestation consists of the XML file and its images. Migrating an Oracle database to an Access database, consumes .dbf, .ctl files, etc. and produces one .mdb file.
- Several output components: When migrating a Word Manifestation to an HTML Manifestation, the output Manifestation consists of the XML file with an accompanying CSS file. Migrating a .zip file to its expanded version leads to multiple formats.

► The *Preservation Action* has an *Environment* and *Environment Components* of its own. The migration tool, for example, is one of them.

These *Environment Components,* including the *Preservation Action's Content/Self,* have *Characteristics* of their own, such as *Accepted Input Format*, *Output Formats*, *Preservation Action Cost*. They are also used to guide preservation planning through *Action Defining Requirements*, which are discussed later in this report. *Action Defining Requirements* define which kinds of *Preservation Actions* are desirable independent of the *Characteristics* of the *Preservation Object*, but dependent on the *Characteristics* of the *Preservation Action* itself, such as that PDF may, for a given institution, not be an acceptable preservation output format of a *Preservation Action*.

► If one wanted to extend the scope of the model to business processes other than preservation planning, (whose goal is to selects *Preservation Actions*), then the concept *Preservation Action* should be replaced with a concept *Preservation Process* or *Process*. These processes can then be described in the same way as *Preservation Actions* are here, and have *Requirements* attached to them.

Elements of Preservation Action

- *Action Identifier* (mandatory, non-repeatable): a unique identifier of the concrete *Preservation Action* (data constraint: *Preservation Action* ID)
- *Action Type* (optional, repeatable): a type specification of the *Preservation Action* (data constraint: one of *replacement, reconstruction, replacement)* This is optional because the system might not implement functionality which depends on the action type of the *Preservation Action* instances.
- *Has Preservation Object Type* (optional, non-repeatable): a type specification of the *Preservation Object* (data constraint: *Preservation Object* Type)
(This relationship is also established via the *PreservationObjectType* of the input *Preservation Object*).
- *Has Preservation Risk* (optional, repeatable): a unique identifier of the concrete *Preservation Risk* which prompts the *Preservation Action* (data constraint: Preservation Risk ID) The *Preservation Risk* object contains the information about the *Preservation Risk Type* and the type of the *Environment Component* that is at risk.
- *Has Input Preservation Object* (mandatory, non-repeatable): a unique identifier of the *Preservation Object* on which the *Preservation Action* is being executed (data constraint: *Preservation Object* ID)
- *Has Output Preservation Object* (optional, non-repeatable): a unique identifier of the output *Preservation Object* which results from the execution of the *Preservation Action* (data constraint: *Preservation Object* ID)
- *Has Input Environment* (mandatory, non-repeatable): a unique identifier of the applicable *Environment* of the input *Preservation Object* (data constraint: *Environment* ID) including all *Environment Components* and their *Characteristics* which can be used to evaluate *Preservation Guiding Requirements*
- *Has Output Environment* (optional, non-repeatable): a unique identifier of the *Environment* of the output *Preservation Object* (data constraint: *Environment* ID) including all *Environment Components* and their *Characteristics* which the *Preservation Object* would have after execution of the candidate *Preservation Action*. These can be used to evaluate *Preservation Guiding Requirements*
- *Has Environment* (mandatory, non-repeatable): a unique identifier of the *Environment* of the *Preservation Action* itself (data constraint: *Environment* ID) including the tool which executes the *Preservation Action*; including all other *Environment Components* and their *Characteristics* which can be used to eval*uate Action Defining Requirements*
- *Has Event* (optional, repeatable): unique identifiers to each of the *Preservation Action's Event* objects (data constraint: Event ID)

Other relationships with Preservation Action

- The *Preservation Workflow* object has a *BPEL contains* association link to the *Preservation Action* object.

5.10 Characteristic

It is important to note that the terminology regarding characteristics, properties, values, facets, etc. throughout the preservation literature is very inconsistent. The literature refers to significant properties just as it refers to essential characteristics. The terminology in this report is internally consistent. Efforts are being made to unify the use with other work-packages.

Definition of Characteristic

> A *Characteristic* of a *Preservation Object* is the concrete *Value* which this *Preservation Object* has for an abstract *Property* in a defined context (a concrete *Property/Value* pair). In the model it is the *Characteristic* of an *Environment Component* which belongs to a *Preservation Object* or a *Preservation Action*.

The model's scope is limited to *Characteristic*s which are expected to be used in *Preservation Guiding Documents* and are expected to be useful for preservation planning.

► Natural languages can be misleading. For example, "Software Characteristic" might mean the *Characteristic* that some software package has (e.g. speed, quality of documentation, version) or it might mean which *Characteristic* something has with respect to its software (e.g. the numbers of licenses an institution holds for some given software). In the first case the software is the subject of the statement. In the second case the software is the object of the statement. It is essential to always be clear which concept is the subject of the "sentence", i.e. for which Environment Component this *Characteristic* holds.

► *Characteristics* of *Environment Components* exist for every *Preservation Object Type* (including dynamic definitions thereof) as well as for *Preservation Actions*.

5.10.1 Property

Definition of Property

> An abstract attribute, trait or peculiarity suitable for describing an *Environment Component*.

Which *Property* is applicable to an *Environment Component* depends both on the *Environment Component Type* and the Type of its *Preservation Object*.

► Unlike the other concepts introduced so far, the *Property* concept is purely abstract and defined as part of the vocabulary of the domain of preservation planning. For a final version of a data dictionary each *Property* in the domain should be described by the elements listed below. For now we restrict ourselves to illustrating them through diagrams which show their interrelationships.

► Every *Property* is valid for exactly one *Preservation Object Type* and *Environment Component Type*. A Property with the same name can be defined for other *Preservation Object Type* and *Environment Component Type* combinations, but will have a different *Property Identifier*.

Example:

The *Property* "FormatType" may exist for the "*Format" Environment Component* of "*Bytestream" Preservation Objects*., which might specify "pdf" and "doc" formats. Or it may exist for the "*Format" Environment Component* of "*Collection" Preservation Objects*, which might specify "Dewey Decimal" and "Library of Congress Classification" formats. These different definitions are distinguished by the globally unique identifier of their *Property* object.

► Every *Property* can have several data constraints. This is particularly important for preservation characterisation."bitDepth", for example, is described as one non-negative number in PNG and as three nonNegativeNumbers (one for every colour channel) in TIFF. It is important to be able to specify which data constraint is chosen and also, how this data constraint can be compared to others.

► *Properties* are modelled hierarchically. For example "maintenaceSalaryCost" is a kind of "maintenanceCost" which is a kind of "budgetCost". Their relationships have to be modelled explicitly.

► It is important to express how *Values* for this *Property* may be obtained. If it is an assigned value, for example, what are the rules that can be used to produce it? If it is a derived value, what are the algorithms and software tools that can be used to derive it.

Specifying Property Types

Figure 17 shows three *Property Types*.

Class Property: A *Property* whose value is shared by all instances of an *Environment Component* of a certain type.

E.g. For a given *Collection Preservation Object*: Every instance of a *Software Environment Component* with the same software version has the same values for its *OutputFormats*, *QualityOfDocumentation* and *ProductInformation Properties*.
Often it is advantageous to store a *Class Property* in a registry, and to refer to them via the unique identifier of their *Environment Component* used in the registry. This keeps all information about this type of *Environment Component* in one place and avoids repeating the *Property* locally. Values for these *Properties* can be set in or got from registries using OCL queries. Requirements specifying these *Properties* can be modelled using OCL invariants.

Instance Property: A *Property* of an individual *Environment Component*.

E.g. For a given *Collection Preservation Object*: The *InstallationDate Property* of a *Software Environment Component* will vary for each installation. *Values* for and *Requirements* containing these *Properties* can be modelled as described below.

Aggregate Property: A *Property* of an aggregate of *Environment Components.*

E.g. For a given *Collection Preservation Object*: The *NumberOfInstallations Property* for a *Software Environment Component* with the same software version is an aggregate information over all *Environment Components* of the same type.
Often it is advantageous to store *Aggregate Characteristics* in an inventory, to avoid repeating the *Properties* locally. Values for these *Properties* can be set or got as algorithms for computing the property (how to get/set the values) or as or invariant on the class (to express the requirements)

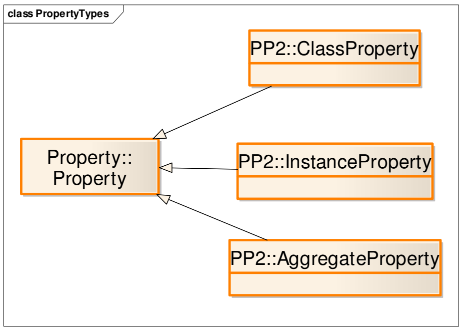


Figure 17 Property Types

► Registries may describe *Class Properties* of *Environment Components* in general, such as *Properties* of Software products Hardware products Schemata and DTDs, name spaces File formats Legal documents, including contractual agreements of the Institution Staff roles Preservation services Preservation risk types Characterisation algorithms	► Inventories may describe *Instance* or *Aggregate Properties* of *Environment Components* in use in the institution, such as *Properties* of Collections of the institution, including profiles User communities of the institution Producer communities of the institution Software products in use Hardware products in use Schemata and DTDs, name spaces in use File formats in use Legal documents in use Staffing numbers and descriptions Preservation services in use Preservation risks that apply to the Institution Characterisation algorithms in use

Elements of Property

- *Property Identifier* (mandatory, non-repeatable): a unique identifier of the *Property (*data constraint: Property ID*)*
- *Property Name* (mandatory, repeatable): a human readable meaningful descriptor for the *Property* (data constraint: string) It is repeatable in order to allow for synonyms.
- *Has Preservation Object Type* (mandatory, non-repeatable): a type specification of the *Preservation Object Type* for which this *Property* is applicable (data constraint: *Preservation Object Type*).
- *Has Environment Component Type* (mandatory, non-repeatable): a type specification of the *Environment Component Type* for which this *Property* is applicable (data constraint: *Environment Component Type*).
- *Data Constraint* (mandatory, repeatable): permissible values; a type definition for the value; possibly a URI for defined vocabulary for the *Property*
    - *Data Constraint Identifier* (mandatory, non-repeatable): a unique identifier of the *Data Constraint* (data constraint: none)
    - *Data Constraint* (mandatory, non-repeatable): permissible values; a type definition for the value; possibly a URI for defined vocabulary for the *Property* (data constraint: data type)
    - *Unit Constraint* (optional, non-repeatable): permissible Units (data constraint: taken from an extensible set of permissible units)
    - *Has Relationship To Data Constraint* (optional, repeatable): How the *Values* for the *Property* may be compared or converted from this data constraint type to another for the same *Property.* This is important for preservation characterisation and comparison.
        - *Target Data Constraint* (mandatory, non-repeatable): permissible values for the conversion target (data constraint: data type)
        - *Target Unit Constraint* (mandatory, non-repeatable): permissible units for the conversion target (data constraint: taken from an extensible set of permissible units)
        - *Conversion Technique* (optional, non-repeatable): Rule, algorithm or logic used for converting the *Value* (e.g. Anglo-American Cataloguing Rules, FFT) (data constraint: none*)*
        - *Conversion Agent* (optional, repeatable): conversion software tool and version; (data constraint: none)
- *Has Relationship* (optional, repeatable): relationship to other *Property* concepts
    - *Relationship Type* (mandatory, non-repeatable): a type specification of the relationship to an other *Property* concepts (data constraint: relationship type taken from an extensible local vocabulary, such as *GeneralizationOf, SpecializationOf, or any association name*)
    - *Related Property* (mandatory, non-repeatable): (data constraint: Property ID)
    - *Multiplicity Source* (mandatory, non-repeatable): (data constraint: one of 0, 1, 0..n, 1..n)
    - *Multiplicity Target* (mandatory, non-repeatable): (data constraint: one of 0, 1, 0..n, 1..n)
- *Has Value Options* (optional, repeatable): How the values for the *Property* may be obtained or updated if it is stored
    - *Value Option Identifier* (mandatory, non-repeatable): a unique identifier of the *Value Option* (data constraint: none)

- ○ *Value Type* (optional, non-repeatable): a type specification of the *Value* (data constraint: one of *assign, derive)*
- ○ *Value Technique* (optional, non-repeatable): Rule, algorithm or logic used for obtaining the *Value* (e.g. assigned according to Anglo-American Cataloguing Rules, FFT) (data constraint: none*)*
- ○ *Value Source Type* (optional, non-repeatable): a type specification of the original source from which the *Value* can be derived (data constraint: none)
- ○ *Creation Agent* (optional, repeatable): For measured and derived *Values*: software tool and version; For assigned *Values*: person or software tool and version (data constraint: none)
- ○ *Creation Trigger* (optional, repeatable): a Trigger for *Value* assignment: e.g. upon ingest, upon *Preservation Action*, etc. (data constraint: none)
- ○ *Property* , repeatable): unique identifiers to each of the *Property's Event* objects (data constraint: Event ID)

Example for Property

The example in Table 4 shows a definition of the *Property* "FormatType" for the "*Format" Environment Component* of "*Bytestream" Preservation Objects*.

This Property definition has 3 types of data constraints: PUIDs, MIME multipart top-level elements, and full MIME multipart format types. They all are valid alternative value systems.

For PUIDs the definition lists a conversion table and tool which will produce equivalent MIME format types.

The values may be created in two ways: They may be assigned by the ingest engine which has hard-coded information about the format types being ingested. Alternatively it may be characterised by the JHOVE file format characterisation tool

• *Property Identifier*	FormatType325
• *Property Name*	FormatType
• *Has Preservation Object Type*	Bytestream
• *Has Environment Component Type*	Format
• *Data Constraint* ○ *Data Constraint Identifier* ○ *Data Constraint* ○ *Has Relationship To Data Constraint* • *Target Data Constraint* • *Conversion Technique* • *Conversion Agent*	 DC1 PUID file://toplevel_MIME_multipart_list MappingTable325.1 PUID-MIME-Converter, Version 0.3
• *Data Constraint* ○ *Data Constraint Identifier* ○ *Data Constraint*	 DC2 file://toplevel_MIME_multipart_list
• *Data Constraint* ○ *Data Constraint Identifier* ○ *Data Constraint*	 DC3 file://full_MIME_multipart_list
• *Has Relationship* (optional, repeatable): relationship to other *Property* concepts • *Relationship Type* ○ *Related Property* ○ *Multiplicity Source*	 HasVersion FormatVersion 1..n

○ *Multiplicity Target*	0..n
• *Has Value Options*	
○ *Value Option Identifier*	CO1
○ *Value Type*	assign
○ *Value Technique*	AMHD code design specification, Version 20070913
○ *Creation Agent*	Ingest algorithm123, version 2
○ *Creation Trigger*	Ingest into the AMHD archive
• *Has Value Options*	
○ *Value Option Identifier*	CO2
○ *Value Type*	measured
○ *Value Technique*	JHOVE, Version 1.1 identification algorithm
○ *Value Source Type*	Bytestream
○ *Creation Agent*	JHOVE, Version 1.1
○ *Creation Trigger*	Ingest into the AMHD archive

Table 4 Example property

Vocabulary for specifying Properties

In the Appendix 7.3 we list an initial collection of *Property* vocabulary for a subset of the *Environment Component Type* - *Preservation Object Type* combinations*.* The goal is to have a deep vocabulary that would be generally acceptable and sharable by different institutions. The current state of the vocabulary is a first phase attempt. More work is needed to expand it, validate it, and harvest community input. For certain subsets one can refer to related work. For example, the PREMIS[32] preservation metadata defines *Properties* for *Manifestations* (≈ PREMIS Representation), and *Bytestreams* (≈ PREMIS File and Bitstream).

5.10.2 Value

Every *Characteristic* has a *Value*.

The *Value* can either be assigned explicitly or be inherent in the *Realisation* of the *Preservation Object* and extracted on demand.

Setting the *Value*:

► Assigned *Values* may be assigned manually or as a side-effect of a process. E.g. The *Budget* of an institution may be set during the execution of a *Preservation Action*: PreservationBudgetSize := PreservationBudgetSize – PreservationActionCost.

► Regular internal operations, such as ingest of digital objects, purchase of hardware and software, decommissioning of equipment, hiring, training and laying-off of staff, getting and spending money, or executing *Preservation Actions*, all change *Characteristics*. Equally, external operations, such as introducing a new file format or a new *Preservation Action* tool, change *Characteristics* which may be recorded in registries.

These changes needs to be recorded in the system to inform choices and decisions in preservation planning.

Equally changes in *Characteristics* may have to prompt an update, or an addition, or a removal of *Requirement*s.

The monitoring process to determine that a *Characteristic Value* has changed, or that a *Requirement* needs to be added, removed or changed is out-of scope for this report.

► Assigned *Value*s may be given (e.g. for every data object the content-type of an eJournal ingest system is always set to “eJournal” upon ingest), or they may be extracted from the *Preservation Object* (e.g. the *Bytestream* size or *Collection* size may be extracted).

[32] PREservation Metadata: Implementation Strategies, http://www.loc.gov /premis

► *Values* should be recorded at the *Preservation Object* granularity where they apply. Different software systems might inherit these values up or down the *Preservation Object* hierarchy in different ways, depending on their tasks. These software systems should contain the logic about the inheritance.

Getting the *Value*:

The *Value* can be looked up if it is stored explicitly. If it is inherent in the *Preservation Object*, it can be extracted and measured with an associated preservation characterisation tool according to a specified algorithm, or deduced with a given logic.

The characterisation process itself is out-of-scope for this report.

Characterisation tools are defined to work on the *Manifestation* and *Bytestream* level. But there are other tools, which characterise on a higher level, e.g. Collection profiling tools analyse *Properties* of a *Collection* at a given time and measure their *Values*.

Elements of Value

- *Value* (mandatory, non-repeatable): *Value* of the *Characteristic (*data constraint: none*)*
- *Unit* (mandatory, non-repeatable): the unit of the *Value* (data constraint: taken from an extensible set of permissible units)
- *Has Value Option Identifier* (optional, non-repeatable): a unique identifier of the *Value Option* which specifies the type of the *Value* (*assign*, *derive*) and the technique that was used to obtain the *Value*.(data constraint: none)
- *Creation Event* (optional, non-repeatable): a unique identifier of the *Event* which created the *Value*. It includes information about the dates the value was set, the creation agent and the creation source (data constraint: Event ID[33])

Elements of Characteristic

- *Characteristic Identifier* (mandatory, non-repeatable): a unique identifier of the *Characteristic (*data constraint: Characteristic ID*)*
- *Has Environment Component* (optional, non-repeatable): a unique identifier of the *Environment Component* (data constraint: *Environment Component* ID)
(This relationship is also established via the *Has Characteristic* relationship of the *Environment Component*)
- *Has Preservation Object Type* (optional, non-repeatable): a type specification of the *Preservation Object Type* to which the *Environment Component* belongs (data constraint: *Preservation Object Type*, such as *Collection*, *Deliverable Unit, Expression*, etc.).
(This relationship is also established via the *Has Environment* relationship (recorded in the *Environment Component)* which leads to the *Has Preservation Object* (recorded in the *Environment*) which leads to the *Preservation Object* and its type).
- *Has Property* (mandatory, non-repeatable): a unique identifier of the *Property* to which this Characteristic refers (data constraint: extensible vocabulary: Property).
- *Has Value* (mandatory, non-repeatable): a unique identifier of the *Value* (data constraint: Value ID)
- *Has Event* (optional, repeatable): unique identifiers to each of the *Characteristic's Event* objects (data constraint: Event ID)

Other relationships with Characteristic

- The *Environment Component* object has a *Has Characteristic* association link to the *Characteristic* object.

[33] Event is a concept from [Core]

5.11 Requirement

Definition of Requirement

A constraint which limits the space of allowable preservation planning activities.

It is expressed through one or more *Property/Value* constraint specifications on *Environment Component Types*. They are limited to specified *Preservation Object Types* or *Preservation Action Types*, and may include pre- or post-conditions.

► The *Values* of the *Characteristics* which describe the actual *Preservation Objects* at that time and the *Values* of the *Characteristics* which describe a candidate *Preservation Action* can be matched against a *Requirement* in order to determine whether it is applicable and satisfied. To do this one has to determine how the concrete *Characteristic* matches the abstract *Property* and *Value* constraint in the *Requirement*.

► Degradation to *Preservation Objects* is caused by two things:

- *Preservation Risks*
- Executing *Preservation Actions*, which might not preserve all *Characteristics* of the *Input Environment* in the newly created *Output Environment*.

Acceptable levels of either are described in *Preservation Requirements*.

► *Preservation Requirements* support

- Identifying *Risks* by explicitly stating what the perceived *Risks* for *Preservation Objects* are.
- Identifying opportunities for improving the digital Collections
- Determining the cost/benefit of a *Preservation Action* by explicitly stating the institution*'s* values. The degree to which the *Preservation Action* satisfies those *Requirements* determines its cost/benefit for the institution.

► *Preservation Requirements* are constraints that makes the institution's values explicit and influences the preservation process.

► *Preservation Requirements* are described in *Preservation Guiding Documents.*

► *Preservation Requirements* may be formulated for any *Preservation Object Type*, i.e. they may apply to *Collections*, *Deliverable Units*, *Expressions*, *Components*, *Manifestations* or *Bytestreams.*

► *Preservation Requirements* are measurable subsets of goals. They express a target level of results expressed in units against which achievement is to be measured. *Preservation Requirements* provide the day-to-day support for achieving goals. [adopted from StratML, Objectives]

► There are many sets of *Requirements* which contradict each other. An institution may only instantiate non-contradictory subsets of *Requirements*.
e.g. “With every *Preservation Action* produce a print-quality (300dpi) PDF for print-on-demand customers.”
versus “Don't archive derivative copies which can be derived from others.”

Use of Requirements for preservation planning

► For any given *Preservation Object* and its *Environment* there are multiple possible *Preservation Actions* which might mitigate a *Preservation Risk*.

How desirable it is to execute a candidate *Preservation Action* depends on

- the priorities of the institution, as described in their *Requirement*s.
  - Example: One cannot say that the perfect migration format for a PDF file is PFD/A, just because it preserves the important *Characteristic*s of the PDF file and has archival quality. This is only the case if those *Characteristic*s are actually significant to the institution. Similarly, while, for example, the video stream of an mp4 file is generally considered significant, it may not be significant to a radio station.
- the *Characteristic*s of the digital Object itself.

    - Example: If a Word file contains only text without formatting, headers and tables, etc., then a .txt output might be considered perfectly adequate, even though this would in general not be considered a target migration format for a Word file.

This is to say, that the desirability of *Characteristics* cannot depend solely on a file format and therefore the choice of *Preservation Action* cannot be purely based on the file format. A *Characteristic* is desirable if it applies under the circumstances and if it matters to an institution.

► When a *Preservation Action* is applied to a *Preservation Object* and its *Environment* then a new copy of the *Preservation Object* and/or a new *Environment* is created in which the *Preservation Risk* is mitigated. Every *Preservation Action*, therefore, does not only have an *Input Preservation Object* and an *Input Environment*, but also an *Output Preservation Object* and an *Output Environment*.

► *Requirements* define

- acceptable *Characteristics* of the *Preservation Action* itself
    - Example: "Output file formats need to be platform independent."
- acceptable output *Characteristics* of the *Preservation Object*, which may be
    - dependent on input *Characteristics*
        - compares the differences between the input and output *Characteristics* and measures to what degree this difference satisfies the required *Characteristics* (e.g. loss of *Characteristics*)
        - Example: "The loss of resolution may not exceed 20% of the original resolution".
    - independent of input *Characteristics*
        - measures to what degree the output *Characteristic* satisfies the required *Characteristic*.
        - Example: The size of the *Preservation Action*'s output *Preservation Object* should not exceed a maximal size set by the institution.

Preservation Guiding Documents also contain *Requirement*s which

- describe the preservation process itself independent of the *Characteristics* of the *Preservation Object* as well as of those of the *Preservation Action*
    - Example: a preservation planning process should be executed for every data object at least every 5 years, independent of the *Preservation Risks* that are established for this data object.
- do not describe the preservation process itself. They are contained in *Non Preservation Requirements*.

► During preservation planning one determines which of the candidate *Preservation Actions* is the most suitable for the *Preservation Object*. This can be derived by considering the *Characteristics* of the *Preservation Object* before and after the execution of a candidate *Preservation Action*, and by comparing them to the institution*'s Requirements*. This process lets us derive to what degree this *Preservation Action* would satisfy the *Requirements*. It amounts to a cost/benefit analysis of the *Preservation Action*, since the degree to which *Characteristics* are lost is a cost (not necessarily financial) to the institution. Independent of *Characteristics*, the cost/benefit analysis may amongst others also comprise the cost of executing the *Action*, and the cost of needed infrastructure for sustaining preservation output. The benefit of the *Preservation Action* is the benefit of mitigating the risk in terms of the value of the object, the severity of the risk, etc.. Obviously these costs and benefits are not necessarily monetary.

► The output *Characteristic* is not necessarily inferior to the input *Characteristic*, i.e. a preservation is not always lossy.

Examples:

- A migration from a PDF file to an XML file might render the new digital object editable, which previously was not the case. This might be desirable to an Institution.

- A file might be manually restored by a curator to the state it was presumed to have had before a corruption.

In these cases the candidate *Preservation Action* should receive an increased evaluation score.

5.11.1 Risk specifying requirement

Preservation Risks are specified in *Risk Specifying Requirements.* Whenever *Characteristics* of a *Preservation Object's Environment Component* violate certain *Values* which are specified in the *Requirement* then the *Preservation Object* is considered at risk.

Once a *Risk Specifying Requirement* is violated a preservation monitoring process should trigger the preservation planning process. It, in turn, determines the optimal *Preservation Action* which should mitigate this *Preservation Risk*.

Preservation Object Selecting Requirements are a special class of *Risk Specifying Requirements* which specifies which subset of *Preservation Objects* is at risk.

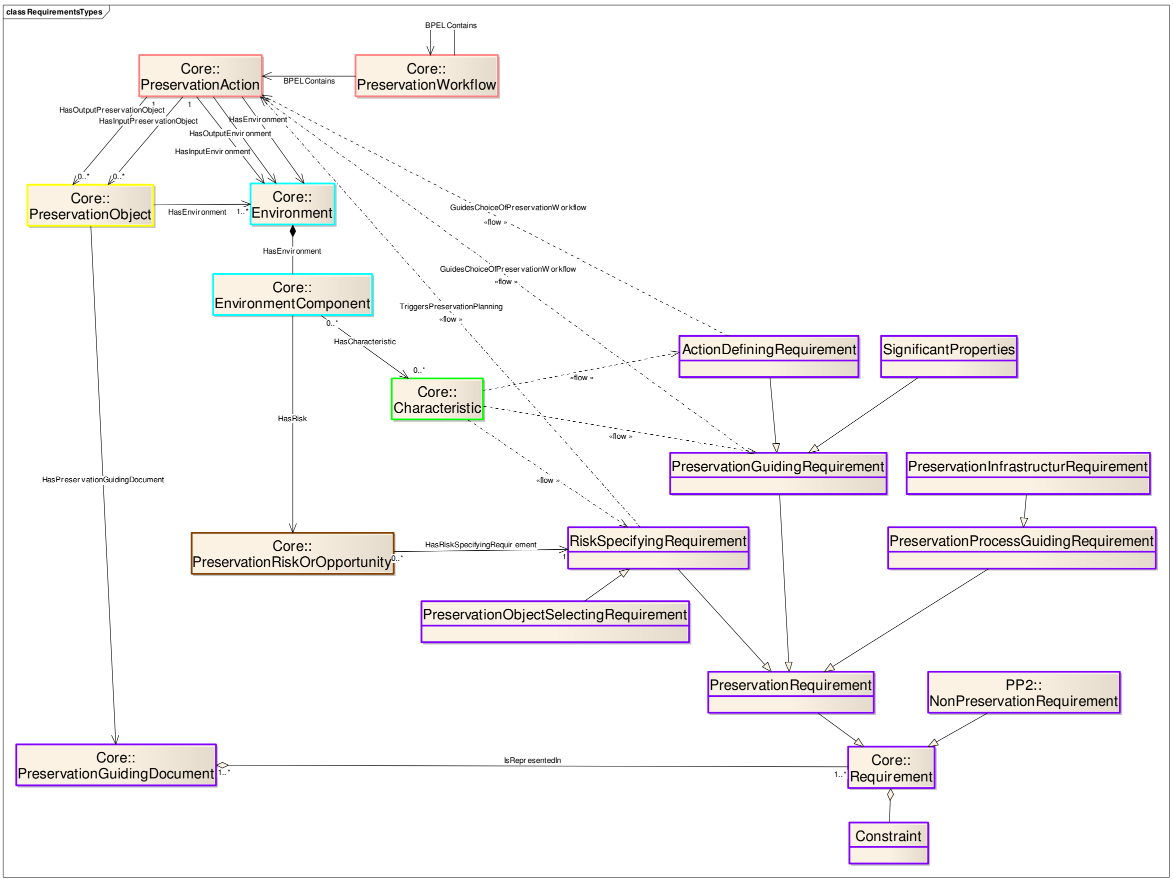

class RequirementsTypes
BPEL Contains
Core::
PreservationAction
BPEL Contains
Core::
PreservationWorkflow
HasOutputPreservationObject
HasInputPreservationObject
1
1
HasEnvironment
HasOutputEnvironment
HasInputEnvironment
0..*
0..*
Core::
PreservationObject
HasEnvironment
1..*
Core::
Environment
GuidesChoiceOfPreservationWorkflow
«flow »
HasEnvironment
GuidesChoiceOfPreservationWorkflow
«flow »
Core::
EnvironmentComponent
TriggersPreservationPlanning
«flow »
0..*
HasCharacteristic
0..*
ActionDefiningRequirement
SignificantProperties
«flow »
Core::
Characteristic
HasRisk
«flow »
PreservationGuidingRequirement
PreservationInfrastructurRequirement
HasPreservationGuidingDocument
«flow »
Core::
PreservationRiskOrOpportunity
0..*
HasRiskSpecifyingRequirement
1
RiskSpecifyingRequirement
PreservationProcessGuidingRequirement
PreservationObjectSelectingRequirement
PreservationRequirement
PP2::
NonPreservationRequirement
Core::
PreservationGuidingDocument
1..*
IsRepresentedIn
1..*
Core::
Requirement
Constraint

Figure 18 Vocabulary for Requirement Types

5.11.2 Preservation guiding and action defining requirement

► *Preservation Guiding Requirements* define which kinds of *Preservation Actions* are desirable for the *Preservation Object*, dependent on

- which input *Characteristics* of the *Preservation Object* need to be met to consider the *Preservation Action*
- which output *Characteristics* of the *Preservation Object* are permissible/ desirable (either in absolute terms or in relationship to *Characteristics* of the input *Preservation Object*, which might be a derivative or the original submitted to the institution[34].)
- which *Characteristics* of the *Preservation Action* itself are desirable

Example: The size of the *Preservation Action*'s output *Preservation Object* should not exceed a maximal size set by the institution.

[34] It is important to not accumulate errors in subsequent preservation actions, which implies that it is best to express comparative losses with respect to the original *Preservation Object.*

► *Action Defining Requirements* are a special class of *Preservation Guiding Requirements*. They define which kinds of *Preservation Actions* are desirable independent of the *Characteristics* of the *Preservation Object*, but dependent only on the *Characteristics* of the *Preservation Action* itself.

Example: PDF may, for a given institution, not be an acceptable preservation output format of a *Preservation Action*).

► Significant Properties were defined by Andrew Wilson, National Archives of Australia, to be "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."

Significant Properties are a special class of *Preservation Guiding Requirements* which is in general considered to be limited to *Characteristics* of *Bytestreams*. Often they are limited to *Characteristics* for which it is possible to evaluate the satisfaction of their *Requirements* automatically.

5.11.3 Preservation process guiding requirement

Preservation Process Guiding Requirements are independent of the *Characteristics* of the *Preservation Object* as well as of those of the *Preservation Action*. They might guide the preservation planning or preservation execution process.

Example: A preservation planning process should be executed for every data object at least every 5 years, independent of the *Preservation Risks* that are established for this data object.

Preservation Infrastructure Requirements are a special class of *Preservation Process Guiding Requirements* which specifies what *Characteristics* are required of the infrastructure.

5.11.4 Non preservation requirement

Preservation Guiding Documents also contain *Requirement*s which do not describe the preservation process itself. They are contained in *Non Preservation Requirements*.

Elements of instantiated (non-parameterized) Requirements

- *Requirement Identifier* (mandatory, non-repeatable): a unique identifier of the *Preservation Requirement.* Could be a URI (data constraint: Requirement ID)
- *Requirement Name* (optional, repeatable): a human readable meaningful descriptor for the *Requirement* (data constraint: string)
- *Requirement Type* (mandatory, repeatable): a type specification of the *Requirement* (data constraint: extensible vocabulary: taken from the specific vocabulary for *Preservation Requirement* Types).
- *Requirement Description* (optional, repeatable): a human readable meaningful description for the *Requirement* (data constraint: Description)
- *Stakeholder* (optional, repeatable): (data constraint: Agent ID)
- *Requirement Source* (optional, repeatable):
- *Requirement Applicability* (optional, non-repeatable): Time range during which the *Requirement* is applicable. If it is not specified, then it defaults to the *Document Applicability*
    - *Start Date* (optional, repeatable): The date the *Requirement* is projected to become valid (data constraint: date)
    - *End Date* (optional, repeatable): The date the *Requirement* is projected to cease, if it is not subsequently extended (data constraint: date)
- *Requirement Specification* (mandatory, non-repeatable):
    - Context (mandatory, repeatable): Specifies the object for which the constraint holds
    - Pre (optional, non-repeatable): Specifies a pre-condition for applying the requirement
    - Post (optional, non-repeatable): Specifies a post-condition for applying the requirement

- *Requirement Importance Factor:* Measure of the importance of the requirement for the Institution (data constraint: none)
- *Has Event* (optional, repeatable): unique identifiers to each of the *Requirement's Event* objects (data constraint: Event ID)

The *Requirement Specification* element complies with OCL for specifying constraints. Each pre- and post-condition is a logical expression which combines constraints and can be evaluated to true or false for a given set of *Characteristics Values* of the institution.

In general a constraint will contain some of the following parts:

- Operator : Operator to be applied to determine whether the requirement is satisfied.
  - *Operator* (mandatory, non-repeatable): Function to be evaluated. e.g. "=", "one of", "MyBooleanFunction". The function should evaluate to true/false. If a tolerance is specified the function might return the degree to which the constraint is satisfied with respect to the tolerance.
  - *Tolerance* (optional, non-repeatable): To what degree deviation from the requirement can be tolerated.
- Property specification: It specifies for which property a value should be retrieved. A *Property* is fully specified by the following elements
  - *Property Identifier* (mandatory, non-repeatable): It specifies for which *Property* a *Value* should be retrieved. The *Property* object implies the *Preservation Object Type* and *Environment Component Type* for which this constraint applies. e.g. FormatType325 which is a *Property* of a "*Bytestream" Preservation Object* and its "*Format" Environment Component.*
  - *Data Constraint Identifier* (mandatory, non-repeatable): It specifies which of the possible data constraints is used to express the constraint e.g. DC3 specifies a MIME data constraint
  - *Value Option Identifier* (optional, repeatable): It specifies which of the possible Value Options is used to extract the value. E.g. CO2 which uses a JHOVE format characterisation to extract the MIME type.
- Constant Specification: It specifies a constant value. A constant is fully specified by the following 2 Elements
  - *Value* (mandatory, non-repeatable):
  - *Unit* (mandatory if applicable, non-repeatable):

Units of values or data constraints must be compatible (be the same or have a conversion in *Has Relationship To Data Constraint* in the Property object).

The *Requirement Importance Factor* and the *Tolerance* elements allow for computing a weighted measure of compliance with the *Requirement.*

Other relationships with Requirement

- The *Preservation Guiding Document* object has a *Has Requirement* aggregation link to the *Preservation Requirement* object.
- The *Preservation Risk* object has a *Has Risk Specifying Requirement* association link to the *Risk Specifying Requirement* object.

Example of Requirement

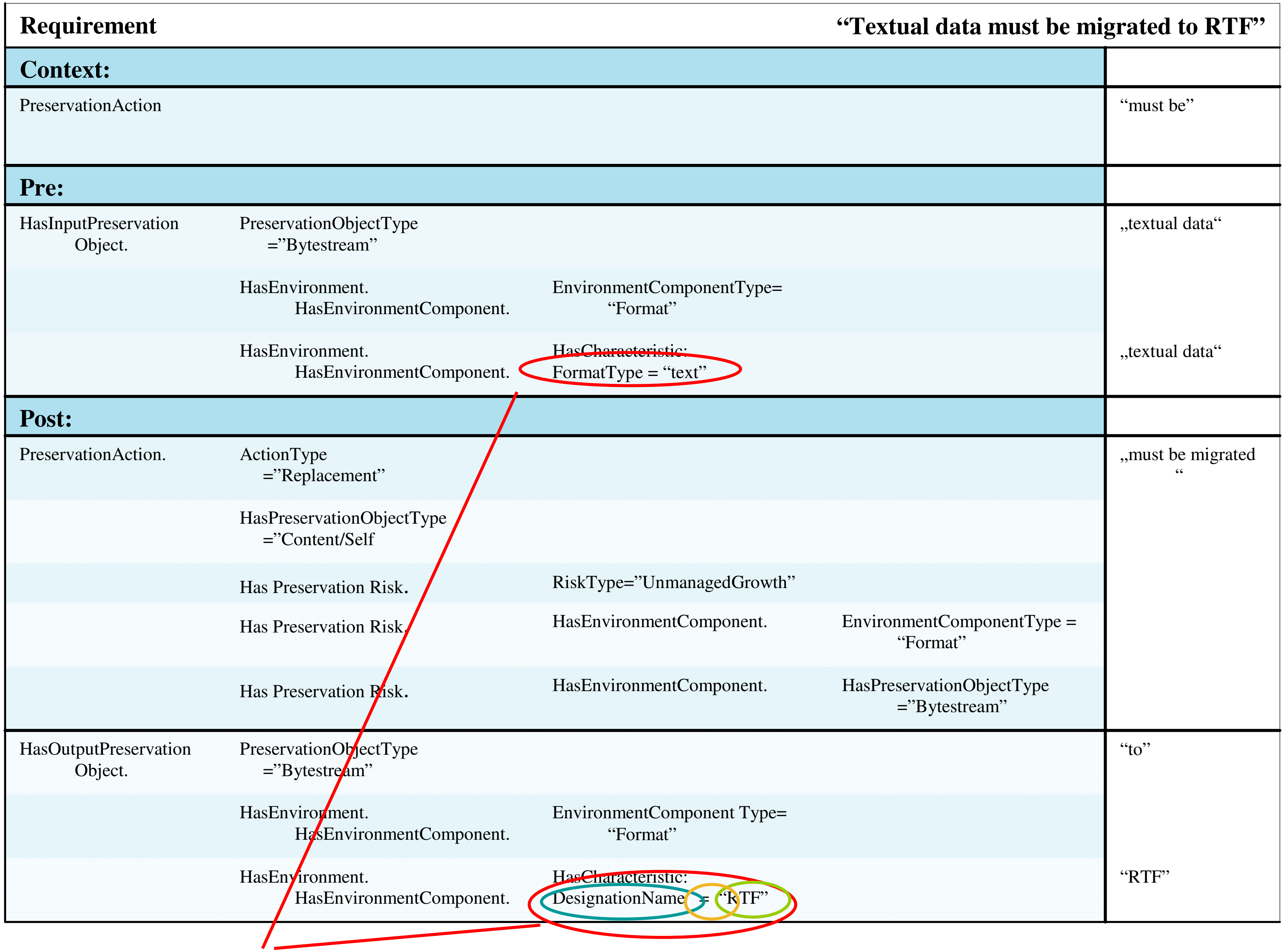
The following example illustrates how a requirement may be expressed solely in terms of model elements and vocabulary.

The requirement "**Textual data must be migrated to RTF**" is being mapped in the following way:

The context of the requirement describes the class to which the precondition, post-condition, or invariant applies. In this example it describes restrictions on eligible *Preservation Actions.*

The precondition describes under which circumstances the *Requirement* applies. This is expressed solely in terms of the *HasInputPreservationObject* relationship between *Preservation Action* and *Preservation Object* and in terms of the
PreservationObjectType
HasEnvironment. *HasEnvironmentComponent*. *EnvironmentComponentType* and
HasEnvironment. *HasEnvironmentComponent.HasCharacteristic*

elements of *Preservation Object*.



Requirement				“Textual data must be migrated to RTF”
Context:				
PreservationAction				“must be”
Pre:				
HasInputPreservation Object.	PreservationObjectType =”Bytestream”			„textual data“
	HasEnvironment. HasEnvironmentComponent.	EnvironmentComponentType= “Format”		
	HasEnvironment. HasEnvironmentComponent.	HasCharacteristic: FormatType = “text”		„textual data“
Post:				
PreservationAction.	ActionType =”Replacement”			„must be migrated “
	HasPreservationObjectType =”Content/Self			
	Has Preservation Risk.	RiskType=”UnmanagedGrowth”		
	Has Preservation Risk.	HasEnvironmentComponent.	EnvironmentComponentType = “Format”	
	Has Preservation Risk.	HasEnvironmentComponent.	HasPreservationObjectType =”Bytestream”	
HasOutputPreservation Object.	PreservationObjectType =”Bytestream”			“to”
	HasEnvironment. HasEnvironmentComponent.	EnvironmentComponent Type= “Format”		
	HasEnvironment. HasEnvironmentComponent.	HasCharacteristic: DesignationName = “RTF”		“RTF”

These are simplified representations of an expression consisting of a Property Specification, Operator and Constant Specification

Figure 19 Example Requirement

The post-condition, finally, describes which condition need to be true after a *Preservation Action* is executed under the given circumstances. Again this is expressed using relationships and elements introduced in the above data model.

Characteristics' Properties and *Values* are specified in a simplified way in the example. In an actual specification they would refer to unique identifiers and might have to include unit specifications and operator tolerances.

This example can now easily be translated into an OCL expression.

5.12 Shared data types

- *Description* (optional, repeatable): a human readable meaningful description which is suitable for object types which describe an abstract concept, such as *Property*, *Requirement*, *stratML:Value*, *stratML:Goal*
    - *Has Definition* (optional, repeatable): A verbal definition of the concept (data constraint: string)
    - *Has Justification* (optional, repeatable): Why this concept is needed for preservation planning (data constraint: string)
    - *Has Example* (optional, repeatable): Examples (data constraint: string)
    - *Has Notes* (optional, repeatable): Notes (data constraint: string)
    - *Has Usage* (optional, repeatable): How this concept is to be used (data constraint: string)
- Version information which is used to manage the history of objects, (such as a history of all *Values* which a certain *Environment Component* takes on over time for a given *Property*) is not included in this model. It is assumed that the system which implements this model will manage versions according to its own needs. Version information that is part of the name of the object (such as a software version or document version) are included.

5.13 Using the data model for preservation planning

Even though process modelling is out of scope for this report, we would like to point out how this model is particularly suitable for uniform processing of all *Preservation Object Types* for all preservation processes (monitoring, planning, characterisation, etc.).

For example, characterisation tools are defined to work on the *Manifestation* and *Bytestream* level. But there are other tools, which characterise on a higher level, e.g. collection profiling tools which analyse *Characteristics* of a *Collection* at a given time and produce profiles describing the *Collection*. They could in principle share a data model and associated processes.

In preservation planning, one needs to consider the *Characteristics* of the *Preservation Object* before and after the execution of a candidate *Preservation Workflow*, and compare them to the institution*'s Requirements*. The result is an evaluation score for how suitable each candidate *Preservation Workflow* is with respect to the Institution's *Requirements*. The utility analysis of the Plato tool is an example of this.

Preservation Requirements express constraints on all levels of *Preservation Objects* in the *Preservation Object* hierarchy (e.g. budgetary and legal constraints on the *Collection* level; preserving interactivity at the *Bytestream*, *Manifestation* or *Deliverable Unit* level) and might even mix *Characteristics* from several levels (e.g. specifying constraints on *Collections* which contain *Bytestreams* with a certain *Characteristic*).

Since each possible *Preservation Workflow* may impact all levels in the *Preservation Object* hierarchy, the evaluation of a *Preservation Workflow* must be determined on all levels. This is to say that for every candidate *Workflow* we can evaluate how well it satisfies the *Requirements* associated with a specific *Bytestream*, but also how well it satisfies the *Requirements* for the whole of the *Manifestation*, or for a *Deliverable Unit*, or even for a *Collection*.

If for example, a concrete *Preservation Workflow* exceeds the I*nstitution's* budget, then it need not be considered for a given *Bytestream*. Equally, if it violates a *Collection* principle, even though it would be very suitable for preserving a specific *Manifestation*, it need not be considered. This sort of higher-level constraint is very useful in immediately ruling out unsuitable candidate *Preservation Workflows* at a lower level.

Conversely, it is necessary to not just evaluate a concrete *Preservation Workflow*'s utility in isolation on a lower level, but rather place it in a higher level context. When combining the evaluations from lower levels, with constraints on the higher level, then the evaluation of a *Workflow* might shift in the more global perspective.

<u>Examples</u>:

- *Preservation Workflow* A is considered more suitable than *Preservation Workflow* B in the evaluation for a digital file. But if we look now onto a higher level then it might not be possible to combine *Preservation Workflow* A with the suggested *Preservation Workflows* for the other files in the *Manifestation*, which is a *Preservation Requirement* on *Manifestation* level. This might, for example, be the case if the *Workflows'* output requires incompatible environments.
- For a .png file we decide that it is best migrated to a .gif file. When we look at the enclosing *Deliverable Unit* "web page" we see that the references to the image are broken and that the best *Workflow* would now add the *Preservation Action* "rename the links". When we look at the next higher *Deliverable Unit* "website" we see that they use java script for their links. The renamed links would not work. The best option is now to use a redirect list for the web server to the image1 on the server side instead of adding the *Preservation Action* "rename the links".

As the example shows, this also means that we have to modify the candidate *Workflows* on higher levels, by either amending the candidate *Workflows* with new *Preservation Actions*, or by replacing parts of the candidate *Workflow*, as needed, or by rejecting *Workflows* which might seem acceptable on a lower level.

It also means that the resulting *Environment* at a higher level will have to be modified compared to lower level *Environments*. If, for example, a *Preservation Workflow* on a file level requires an *Environment* that is insufficient for all the other files in the same *Manifestation*, then the *Environment* needs to be expanded to accommodate all files in the *Manifestation*.

The following figure is supposed to illustrate how lower level evaluations affect the evaluations of candidate *Workflows* on a higher level. The pink "local evaluations" are evaluations you would get if you just looked at the *Characteristics* at that level. They are combined into a contextual (purple) evaluation by combining the local evaluations with the contextual evaluations from all lower levels.

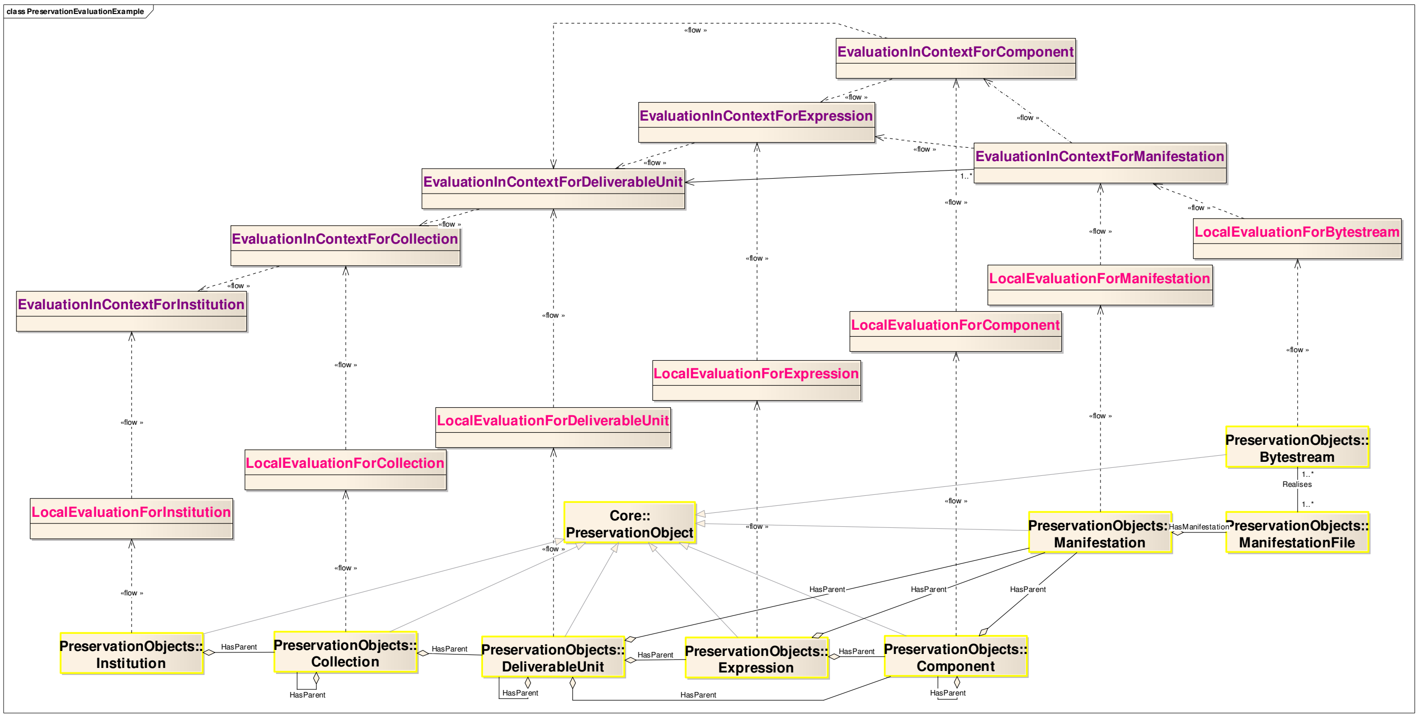

class PreservationEvaluationExample
EvaluationInContextForComponent
EvaluationInContextForExpression
EvaluationInContextForManifestation
EvaluationInContextForDeliverableUnit
EvaluationInContextForCollection
EvaluationInContextForInstitution
LocalEvaluationForBytestream
LocalEvaluationForManifestation
LocalEvaluationForComponent
LocalEvaluationForExpression
LocalEvaluationForDeliverableUnit
LocalEvaluationForCollection
LocalEvaluationForInstitution
Core::
PreservationObject
PreservationObjects::
Bytestream
PreservationObjects::
Manifestation
PreservationObjects::
ManifestationFile
PreservationObjects::
Institution
PreservationObjects::
Collection
PreservationObjects::
DeliverableUnit
PreservationObjects::
Expression
PreservationObjects::
Component
«flow »
1..*
Realises
1..*
HasManifestation
HasParent

Figure 20 Propagating preservation planning results up the Preservation Object hierarchy.

Even though the above diagram shows a separate evaluation process for every level it is actually always the same. It is the process of comparing input and output *Characteristics* with the corresponding *Requirements*, and combining the results from all levels.

This information can now be used to choose the combined *Preservation Workflows* at any level in the *preservation object* hierarchy.

A smaller percentage of requirements is machine-interpretable for higher-level *Preservation Objects*; the non-machine-interpretable ones tend to be described in more abstract preservation guiding documents, such as policy documents. The percentage of machine-interpretable requirements increases as preservation guiding documents become more and more concrete (moving from strategy documents to runtime parameters, and moving from higher-level *Preservation Objects* to lower-level ones). Nevertheless, it is advantageous to incorporate the ones which are machine-interpretable at any level uniformly into the evaluation process.

5.14 Conclusion

The conceptual model presents a simple yet expressive representation of the preservation planning domain.

It builds on the idea of preservation planning as a process to identify and mitigate risks to current and future access to digital objects.

It accommodates all processes which are involved in preservation planning, such as monitoring, characterisation, comparison of characteristics, evaluation of candidate preservation actions, etc..

It allows for uniform execution of these processes on all levels of the *Preservation Object* hierarchy, such as *Collections*, *Deliverable Units*, *Bytestreams*.
Even though the goal of our research was not to deal with characterisation and the related processes on a Bytestream level, it is an invaluable bonus offered by this model that it permits *Characteristics* to be expressed uniformly. This is illustrated in the following excerpts from the preceding text:

- “The term “organisational” does not mean that the model is limited to concepts which model only the organisation as a whole, but rather we include concepts that describe the parts of the organisation at any level, such as dynamic and static collections, deliverable units, expressions, manifestations, components, or files[35], if they affect the preservation planning process and would be expected to be expressed in preservation guiding documents. It is, for example, necessary to refer to characteristics at a lower level to represent requirements at a higher level. For example, in order to specify “collections which contains files that exceed 1 GB”, you need to be able to specify the file property “file size”.
  Even though they are not the focus, the technical aspects of a digital preservation policy or strategy, as well as the state of technology on the basis of which high level constraints can be derived, need to be part of the research scope of this work. Some institutions appear to mandate a particular “technical preservation strategy” (migration, for example) at the preservation policy level, regardless of the lower level technical requirements. This demonstrates the need to integrate institutional and data object considerations in the conceptual model.”
- “Significant Properties are a special class of *Preservation Guiding Requirements* which is in general considered to be limited to *Characteristics* of *Bytestreams*. Often they are limited to *Characteristics* for which it is possible to evaluate the satisfaction of their *Requirements* automatically.”
  “Significant Properties … are not contained in this requirements base since much effort is going into modelling them in other work. If an institution should chose to, it may, however, express them consistent with this model, so that they can be integrated into a holistic planning process for the institution.”

The vocabulary offers a starting point for creating individualised models for individual institutions, even if the institution does not aim for a machine-interpretable document.

[35] Definitions may be found in the Terminology section of this report. See [Core] for a motivation of these concepts.

The goal is to have a deep vocabulary that would be generally acceptable and sharable by different institutions. The current state of the vocabulary is a first phase attempt. More work is needed to expand it, validate it, and harvest community input.

Many of these requirements are by nature not machine-interpretable. In order to translate the rest of them to OCL

- The conceptual model needs to be refined and extended to be able to express all concepts found within the requirements.
- The requirements need to be expressed with more precision. Crisp, measurable definitions are needed that permit evaluation tools to determine whether the constraints are satisfied.

Costing models would make an interesting extension to the requirements base. The model naturally accommodates propagating costing considerations up the *Preservation Object* hierarchy.

The concepts and requirements extracted from the literature and document analysis and the interviews as described in Section 4.5 are fully integrated in the conceptual model.

This model is a first iteration output. It and its vocabulary will be refined, validated and corrected over the coming year.

6. Summary

Digital preservation activities can only succeed if they consider the strategy, policy, goals, and constraints of the institution that undertakes them. Furthermore, because organizations differ in many ways, a one-size-fits-all approach cannot be appropriate.

For digital preservation solutions to succeed, it is essential to go beyond the technical properties of the digital objects to be preserved, and to understand the cultural and institutional framework in which data, documents and records are created, managed, and preserved. Fortunately, organizations involved in digital preservation have created documents describing their policies, strategies, workflows, plans, and goals to provide guidance. They also have skilled staff who are aware of sometimes unwritten considerations.

We have analysed preservation guiding documents and interviewed staff from libraries, archives, and data centres that are actively engaged in digital preservation. This report introduces a conceptual model for expressing the core concepts and requirements that appear in preservation guiding documents. It defines a specific vocabulary that institutions can reuse for expressing their own policies and strategies. The ultimate output of the project will be to machine interpretable models, produced from the basic model and vocabulary, which can be used by preservation planning tools.

To perform the analysis, we used a combination of top-down and bottom-up methods. We examined the scientific literature to create a top-down model from first principles. To complement this, we analyzed actual preservation guiding documents for their content and interviewed decision makers to determine factors that influence their preservation decisions.

The resulting conceptual model presents a very simple and elegant representation of the preservation planning domain. It builds on the idea of preservation planning as a process that identifies and mitigates risks to current and future access to digital objects. It accommodates the full range of processes which are involved in preservation planning, such as monitoring, characterisation, comparison of characteristics, evaluation of candidate preservation actions, and so on. And it allows for uniform execution of these processes on all levels of the preservation object hierarchy, from institutions, collections, down to byte-streams.

The vocabulary can be shared and exchanged by software applications. It also offers a convenient starting point for creating individualised models for an institution; this holds true even if the institution does not require a machine-interpretable document.

We found that many of the requirements we found are by nature not machine-interpretable. In order to express the remaining ones in a machine-interpretable , either the conceptual model needs to be refined and extended to be able to express all concepts found within the requirements, or the requirements need to be expressed with more precision than is currently found in preservation guiding documents. Crisp, measurable definitions are needed that permit evaluation tools to determine whether the constraints are satisfied.

Other key findings can be summarised as follows:

- The features, scope, and level of detail in preservation guiding documents varies tremendously. This reflects a lack of consensus on the use of digital preservation terms, the variety of preservation planning goals, and uncertainty as to how digital preservation should be implemented in practice.
- Preservation policy documents set a general framework for digital preservation, but do not provide specific practical guidance.
- Some existing preservation policies may not accurately reflect the institution’s actual preservation goals. Consequently they may not be particularly useful or fit for purpose.
- Some institutions mandate a specific “technical preservation strategy” (migration, for example) regardless of lower level technical requirements. It is, therefore, important to combine considerations at higher institutional and lower data object levels via a conceptual model for the preservation planning process. Additionally, this use presents a risk of rendering the policy ineffective for some subsets of content.
- Non-technical aspects, such as the regulatory framework, need to be more detailed than they currently are, to support automated preservation planning tools.

- Most current preservation policies and strategies “hard-wire” the choice of preservation action. No on-the-fly preservation planning is needed. These generally fall into 2 categories:
    - Preventive preservation actions, such as format normalisation upon ingest (to focus on a small set of supported formats), diligence during ingest (e.g. rigorously validate and repair errors during ingest) and the use of standards (e.g. in file formats and metadata), avoid difficult preservation situations later on.
    - Data carrier refresh, a re-active preservation action. Now, that the first generations of data carriers are deteriorating at an alarming rate, these replacements are considered urgent and take priority over migration or emulation efforts, which can at this moment safely be postponed.

  Reactive, non-hardware preservation solutions, such as migration and emulation, which require on-the-fly preservation planning are currently avoided, if possible, or not considered necessary or high-priority yet.
- Collections are considered to be well-identified by either their file format types or by the data carrier types of the material. Most institutions have not yet accrued large mixed-format collections which will require automated discovery of existing preservation risks.
- The choice of migration tools is generally considered straight-forward. There are few alternatives to chose from and they are perceived to have clearly identifiable advantages for the given situation.
- Most institutions felt that there are few factors that limit them in their preservation decision making. It was, for example, felt that
    - The legal framework determines **why**, but **not how** things are done.
    - If there is a legal mandate to preserve, then ways will be found to finance preservation and storage.
- Our analysis shows that all institution types studied used very similar concepts. Our confidence in this finding is tempered, however, because (1) we have a very small sample size, so it is not possible to draw statistically significant conclusions; and (2) the documents studied are mostly based on theoretical considerations and may lack the essential details which might differentiate institutional types. Growing practical experience might produce insights into differences which we do not have at the moment; (2) institutions often take on multiple, conflicting roles or take on roles which would be expected to be handles by other institution types.

This is emerging work and this document represents an initial model. We will modify and improve it over the coming year in response to integration efforts with related work, and as the Planets project tries to exploit the ideas in practice. The Methodology section explains our past and intended future approaches. An improved release of this work is planned for May 2009.

7. Appendices

7.1 Modelling approach

In order to specify and document the model, we have drawn on a set of industry standard languages and methods. The model is specified using the Unified Modelling Language (UML) and the Object Constraint Language (OCL). UML models are often depicted as diagrams using a class hierarchy. Several such diagrams appear in this document. OCL is a recent addition to UML that allows more complex constraints to be expressed. As with all Planets models, we are also defining a serialisation in the extensible mark-up language (XML).

7.1.1 UML class diagrams

Object Management Group Inc. (OMG) standardized the initial version of the Unified Modelling Language (UML) in 1997. Since then, UML has been very widely adopted by both industry and academia as the language of choice for describing the architecture of software systems. This is reflected in the fact that it is currently supported by literally hundreds of commercial tools.

UML - the Unified Modeling Language standardizes representation of object oriented analysis and design. A graphical language, its dozen diagram types include Use Case and Activity diagrams for requirements gathering, Class and Object diagrams for design, and Package and Subsystem diagrams for deployment. UML lets architects and analysts visualize, specify, construct, and document applications in a standard way. It is a general-purpose modelling language that can be used with all major object and component methods, and that can be applied to all application domains (e.g., health, finance, telecom, aerospace) and implementation platforms (e.g., J2EE,.NET).

One of the primary goals of UML is to advance the state of the industry by enabling object visual modelling tool interoperability. However, to enable meaningful exchange of model information between tools, agreement on semantics and notation is required. UML meets the following requirements:

- A formal definition of a common MOF (Meta Object Facility)-based meta-model that specifies the abstract syntax of the UML. The abstract syntax defines the set of UML modelling concepts, their attributes and their relationships, as well as the rules for combining these concepts to construct partial or complete UML models.
- A detailed explanation of the semantics of each UML modelling concept. The semantics define, in a technology independent manner, how the UML concepts are to be realized by computers.
- A specification of the human-readable notation elements for representing the individual UML modelling concepts as well as rules for combining them into a variety of different diagram types corresponding to different aspects of modelled systems.
- A detailed definition of ways in which UML tools can be made compliant with the UML-specification. This is supported (in a separate specification) with an XML-based specification of corresponding model interchange formats (XMI) that must be realized by compliant tools.

[OMG Unified Modeling Language (OMG UML), Infrastructure, V2.1.2, November 2007. Online available at http://www.omg.org/docs/formal/07-11-04.pdf (accessed: 26 May 2008)]

The underlying premise of UML is that no one diagram can capture the different elements of a system in its entirety. Hence, UML is made up of a number of diagram types that can be used to model a system at different points of time in the software life cycle of a system.

The policy and strategy model defined in this report makes extensive use of the UML class diagram. The class diagram describes the structure of a system by showing the system's classes, their attributes, and the relationships between the classes. A class is a specification - an object is an instance of a class. Classes may be inherited from other classes (that is they inherit all the behaviour and state of their parent and add new functionality of their own), have other classes as attributes, delegate responsibilities to other classes and implement abstract interfaces.

The Class Model is at the core of object-oriented development and design - it expresses both the persistent state of the system and the behaviour of the system. A class encapsulates state (attributes) and offers services to manipulate that state (behaviour).

7.1.2 OCL – The Object Constraint Language

A UML diagram, such as a class diagram, is typically not refined enough to provide all the relevant aspects of a specification. There is, among other things, a need to describe additional constraints about the objects in the model. Such constraints are often described in natural language. Practice has shown that this will always result in ambiguities. In order to write unambiguous constraints, so-called formal languages have been developed. The disadvantage of traditional formal languages is that they are usable to persons with a strong mathematical background, but difficult for the average business or system modeller to use.

Object Constraint Language (OCL) is a formal language used to describe expressions on UML models. These expressions typically specify invariant conditions that must hold for the system being modelled or queries over objects described in a model. Note, that when the OCL expressions are evaluated, they do not have side effects (i.e., their evaluation cannot alter the state of the corresponding executing system). OCL expressions can be used to specify operations / actions that, when executed, do not alter the state of the system. UML modellers can use OCL to specify application-specific constraints in their models. UML modellers can also use OCL to specify queries on the UML model, which are completely programming language independent.

OCL is a pure specification language; therefore, an OCL expression is guaranteed to be without side effects. When an OCL expression is evaluated, it simply returns a value. It cannot change anything in the model. This means that the state of the system will never change because of the evaluation of an OCL expression, even though an OCL expression can be used to specify a state change (e.g., in a post-condition). OCL is not a programming language; therefore, it is not possible to write program logic or flow control in OCL. You cannot invoke processes or activate non-query operations within OCL. Because OCL is a modelling language in the first place, OCL expressions are not by definition directly executable. OCL is a typed language so that each OCL expression has a type. To be well formed, an OCL expression must conform to the type rules of the OCL language.

In principle, everywhere in the UML specification where the term expression is used, an OCL expression can be used. In UML 1.4 OCL expressions could be used (e.g., for invariants, preconditions, and post-conditions), but other placements are possible too. The meaning of the value, which results from the evaluation of the OCL expression, depends on its placement within the UML model.

For every occurrence of an OCL expression three things need to be separated: the placement, the contextual classifier, and the self instance of an OCL expression.

- The placement is the position where the OCL expression is used in the UML model (e.g., connected to class Person).
- The contextual classifier defines the namespace in which the expression is evaluated. For example, the contextual classifier of a precondition is the classifier that is the owner of the operation for which the precondition is defined. Visible within the precondition are all model elements that are visible in the contextual classifier.
- The self instance is the reference to the object that evaluates the expression. It is always an instance of the contextual classifier. Note that evaluation of an OCL expression may result in a different value for every instance of the contextual classifier.

[Object Constraint Language, OMG Available Specification, Version 2.0., May 2006. Online available at http://www.omg.org/docs/formal/06-05-01.pdf (accessed: 26 May 2008)]

In the next iteration of this work, the requirements of our general model will be represented as OCL expressions. The initial list of requirements, which we extracted during literature analysis, document analysis and interviews, are expressed in natural language. In order to translate them to OCL

- The conceptual model needs to be refined and extended to be able to express all concepts found within the requirements.
- The requirements need to be expressed with more precision.

Examples of the structural breakdown of the requirements below show our approach and will serve as the basis for the formulation of OCL expressions in the next iteration.

7.1.3 XML

Extensible Markup Language (XML) was developed by an XML Working Group (originally known as the SGML Editorial Review Board) formed under the auspices of the World Wide Web Consortium (W3C) in 1996. XML is a simple, very flexible text format derived from SGML (ISO 8879). Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere.

A markup language is a mechanism to identify structures in a document. The XML specification defines a standard way to add markup to documents. Structured information contains both content (words, pictures, etc.) and some indication of what role that content plays (for example, content in a section heading has a different meaning from content in a footnote, which means something different than content in a figure caption or content in a database table, etc.). Almost all documents have some structure.

[W3C (World Wide Web Consortium), Extensible Markup Language (XML). Online available at http://www.w3.org/XML/ (accessed: 26 May 2008)]
The elements in the conceptual model, the specific vocabulary, and the requirements base can be translated into several implementation specific machine interpretable representations. They may for example be represented in an XML schema.

7.2 Reports on interviews

Goal: How do you decide which Preservation Actions to take?

(as recorded in policy documents, strategy documents, business rules, informal decision processes, runtime parameters)

What sort of things are subject to digital preservation?

What guides you in your decision making process?

Who has which functions in digital preservation? (decision making and execution)

Follow up: Why, why not, can you give an example, can you give more granularity

7.2.1 Interview help sheet

This help sheet was intended to guide the interview. The questions were not supposed to be followed rigorously, but were rather meant to inspire a stalled conversation or to provide more depth to the discussion if necessary. The general structure followed the *Environment Component Types* hierarchy. Some questions were borrowed from ERPANET studies.

7.2.1.1 Range of Preservation Actions

Does your digital preservation policy provide guidelines for:

- reformatting data to newer technological platforms
- refreshing data to newer technological platforms
- migrating data to newer technological platforms
- emulating data to newer technological platforms

Which of the following problems are discussed in your digital preservation policy?

- The short-lasting life span and small capacities of media
- The obsolescence of the hardware required to access them
- The obsolescence of software for reading the data and file formats
- The obsolescence of those data and file formats itself
- The technical and structural heterogeneity of the different types of digital documents

7.2.1.2 Preservation Environment – Influencing Factors

Which of the following factors are included in your digital preservation policy?

Which reasons are relevant for your Institution for making digital preservation decisions?

- Legal requirements
- Financial requirements
- Business requirements (e.g., to document important decisions and activities)
- Historical value
- Authority and Responsibility
- Conversion and Reformatting
- Appraisal, Selection and Acquisition
- Storage and Maintenance
- Access and Dissemination
- Standards
- Rules and procedures
- Quality control
- Technical Infrastructure

- Long-term Maintenance
- People
- Organizational structures
- Knowledge

7.2.1.2.1 Internal Influence (fill in with Environment Components from the model)

How do you prioritize which materials need to be preserved?

Which are the required competences for persons responsible for digital preservation?

What type of training or advice is available for them?

How do non-preservation specific guidelines, such as internal policies (e.g., electronic records management handbook) affect your preservation decisions?

Does your policy provide that your Institution takes care of its digital preservation activities itself or are these outsourced?

Do you have direct access to the digital information stored?

7.2.1.2.2 Domain, Business context (Interest groups behind you)

How much is your digital preservation policy influenced by the (business) context in which your organization is working?
What restrictions does this place on your preservation choices?

7.2.1.2.3 Budget

Are there external resources available for digital preservation activities (e.g., government grants, cross-sector funds?

7.2.1.2.4 Legislation, Regulatory Environment

What are the legal requirements and obligations of the institution?

Are there any current national/local/regional rules that affect your digital preservation decisions?

Who is involved in the development of these national/local/regional rules on digital preservation?

How do national rules influence your policy-

- To approve the digital preservation policy
- To implement the digital preservation policy
- To review the digital preservation policy

What restrictions do you have on the use of intermediate repositories, distribution, backing up, format changes, restricted access permissions to material by preservation staff , etc?

Are there specific regulations for the creation of trusted digital repositories which you have to follow?

In the case of outsourcing, are there specific regulations you have to follow?

7.2.1.2.5 Consumer’s Interest

7.2.1.2.6 Producer’s Interest

7.2.1.2.7 Standards (OAI, ISO …)

Do you restrict preservation decisions to ones that comply with certain standards, best practices and guidelines?

If YES, for which area-

- Data formats
- Access

- Organizational policies
- Data Exchange

7.2.1.2.8 State of Technology (Hardware Software)

7.2.1.2.9 File Formats

7.2.1.2.10 Code of Ethics

Are there specific rules to ensure that selected information is complete, accurate and identifiable-(i.e., preservation of defined metadata, of filing plan, registry system data, etc.)

In selecting the preservation method or strategy, has your Institution considered what its effect might be upon the intellectual integrity (e.g., authenticity and reliability) of the digital material-

7.2.1.2.11 Storage (Type of Storage)

How does the quantity of digital materials for which your Institution currently has preservation responsibility affect your preservation decisions?

- approximate number of unique files
- approximate number of volumes (reels of tape, optical disks, etc.)
- total storage volume (in MB, GB, etc.)

7.2.1.3 Decision Making Process

What information would have to be contained in your policy to make it a good basis for making preservation decisions?

Do separate collections have different decision processes or are preservation decisions taken on the institutional level?

Are there any regulations to identify specific persons responsible for the digital preservation?

If your Institution currently has a practice of regular review of items for possible digital preservation treatment, what factors are considered in this process?

Do you use requirements similar to these example requirements?

Comply with all legal requirements (refers to registry of legal requirements)
Don’t exceed the annual budget (by more than x %) (for collection, institution)
Don’t produce output manifestations/files that are larger than x Bytes
Prefer output manifestations whose file formats are supported by existing HW and SW
Prefer output manifestations that are faster, more stable, better supported (compares several output formats and refers to the file format registry)
Preserve colour information
Start preservation action before 1% of sample material is corrupted
The **cost of a preservation action** may not exceed the **value of the object**
Prefer preservation workflows which use software under existing licenses.
Prefer preservation workflows which produce target outputs which satisfy the main user needs.
Preference for implementing preservation actions for which there is expertise
Percentage of objects with a given characteristic at which a certain preservation workflow would be considered amortized
The cost of executing the preservation workflow may not exceed the preservation budget
National legislation may never be violated
The staff cost of supporting new output environments must follow rules in document x
Preserve digital objects for which we do not have printed backup.

7.2.2 National library 1

Interview with the Digital Preservation Manager, at a library, 7 February 2008

7.2.2.1 What sort of things are subject to digital preservation?

The driving factors for what is to be preserved are the following:

- Legislation
- Cultural factors
- Secure Housing of Collection Areas
- Risk analysis

Legislation:

Things that are deposited under legal deposit legislation, currently eJournals and web-archiving, need to be protected through digital preservation.

Cultural Factors:

- Change in publishing from print to e- (eBooks, eJournals, web sites, etc.)
- living authors and composers
- acquisition or deposit of images, sound, etc.

Secure Housing of Collection Areas:

Directorate plan may require *Preservation Action*s which may lead to digitisation and, inevitably, digital preservation. For example one of the newspaper Collections is no longer fully accessible due to its fragile nature, therefore the digital files become more important for long term preservation.

Risk Analysis:

Determines which digital assets are at risk and need to be preserved.

7.2.2.2 What guides you in your decision making process?

Criteria against which risk will be evaluated

- External context:
  - Business/social/regulatory/cultural/competitive/financial/political demands placed on organization
  - External stakeholders
- Internal context:
  - Internal stakeholders
  - Internal capabilities/resources
  - Internal goals & strategies
- Risk management context:
  - The purpose, goals, and objectives of the organization
  - Define what kind of recommendations and decisions can/should be made in response to the analysis
  - Define the depth and breadth of the risk analysis
- risk criteria are established:
  - what kinds of consequence are considered
    - i.e. do we need to consider legal ramifications? Social? *Environment*al? Etc.
  - how will likelihood be defined

- how will risk levels correspond to treatment activities

How do you prioritize the importance of these factors?

Of all factors **media format obsolescence** (= format obsolescence and/or decay of data carriers) is the strongest driver for *Preservation Action*.

As a secondary driver one can look at **content types** (What is on the CDs?) and **software** and **hardware dependencies**.

Legislative constraints should not form a *decisive* factor since all digital *Preservation Action*s must assume a library privilege exemption. If legislation would limit digital preservation then it is mandatory to lobby for change of legislation rather than to compromise the preservation.

Copyright and **IPR** legislation may impact preservation planning decisions; for example, when archiving web-material, permission of the content-owners needs to be obtained, which limits the choice of automatic actions.

Data protection and **freedom of information** are not relevant factors.

Under **legal deposit** there is an obligation to preserve cultural assets.

Publishers are a limiting factor on information access, but not on *Preservation Action*s.

Internal factors exist.

There are finite resources on **personnel**, **funding**, **equipment**. Currently the majority of *Preservation Action*s have to be handled manually. Necessary automatic preservation **tools** and **techniques** don't exist.

There is no limit on **access to material**. What needs to be preserved can be accessed without obstacles. The library has no interest in preserving confidential or corporate records beyond the legally bound timeframes.

There may be **conflicting corporate factors**, for example the archiving prioritisation may be different from the preservation prioritisation of actions.

Availability of hardware and software tools is not a driving factor. The library needs to accommodate a large set of materials, and has an obligation to handle exceptions that other institutions can not support. It therefore needs to have as large an arsenal of tools as necessary to manage the deposited material. To enable this the library uses broad registries, such as PRONOM, GDFR, and Planets techniques.
Normalisation (restricting the supported SW or HW) may be used for large programmes, such as eJournals, but is not applied universally.

The content's **business domain** is not a decision driver. The library has no plans to exclude material of a certain subject from preservation; rather, if at-risk material of a certain media format is being preserved, then all types of content are being batched together, independent of the business domain. This policy may change as more experience is gained.

Interest groups are not an important factor at this time.
However, **specialist curators** raise concerns and influence the prioritisation of preservation decisions.
Funding bodies may influence the choice of *Preservation Action* or preservation content. (varying economic, policy, organisational, technical focus).
Shifting consensus in the **digital preservation community** informally influences the decision process.
Users are currently not a factor in the decision making process; but it is library practice to consult readers on preferences, and it is conceivable that reader priorities may be considered in future preservation decisions.

Co-ordination with other internal library systems is a decision factor since it is necessary to chose actions that are fully **supported**, **avoid duplication** of efforts with other groups, and ensure that outputs are **inter-operable** with other systems.

Use of **standards** is desired in the library, but that has not been a factor in the choice of certain *Preservation Action*s.

Desired access formats are a factor in preservation choices since we need to limit what is published, for example, to the web.

Authenticity of the output is a main driver in the choice of tools.

7.2.2.3 Remarks

Diligence during ingest avoids difficult preservation situations (e.g. rigorously validate and repair errors during ingest).

There is little **need for automatic preservation planning** support **at the organisational level** at the moment, because

- at the moment there are very few *Preservation Action*s to chose from
- there is no economic driver to develop a large set of competing tools.
- There are very powerful tools, such as
    - CD Inspector,
    - characterisation tools, such as Jhove and the NLNZ characterisation tool,
    - DeBabelizer, which recognises hundreds of unknown file formats and converts for PC, Mac and Unix platforms
    - Photoshop which converts fast and deals with exceptions competently for image formats

In the setting of a national library with a preservation obligation, budgets and legislation need to be adjusted to **accommodate the preservation of the cultural objects** rather than vice versa, if the risk is real.

Practical experience so far is limited and the choice of factors will change over time.

7.2.3 Data centre 1

Interview with the head of Digital Preservation and Systems of Data Centre 1, 3 March 2008

What sort of things are subject to digital preservation?

- Data centre 1 is the primary repository for digitised social science research data of it's country.
- 6000+ social science studies are deposited with data centre 1. They comprise data and documentation (e.g. codebook).
- The core studies take up approx. 750GB. Within the historical Collection one study has 3.5TB. The total capacity of all Collections is approx 4.5TB for a single copy. The expected increase is 7 to 10 GB/year.
- Terminology: dataset ≈ study ≈ Collection

Preservation Strategy

- The preservation strategy is based upon open and standardised file formats, data migration and media refreshment.
- *Preservation Action*s are performed for one data set at a time. There is no bulk preservation of data. Bulk preservation is conceivable for a set of documentation documents of a shared text format.
- Most *Preservation Action*s are performed as part of the ingest process.
  - Data and documentation are validated.
  - Data and documentation are, with consent of the depositor, cleaned up if coding errors are found (e.g. semantic inconsistencies, such as male persons giving birth). This is considered to be the original.
  - A preservation copy and dissemination copies are created.
  - Exact documentation about all *Preservation Action*s taken is (but has not always been in the past) associated with the derived copies to guarantee authenticity.
- The original is always kept.
- Later ad-hoc dissemination copies are created on-demand from previous dissemination copies. If for example the latest supported SPSS version is 15, and a requested data-set is currently in version 12, then it can be upgraded to version 15 on demand.

File formats

- The preservation strategy is based upon open and standardised file formats.

Deposited file formats

- About 90% of the studies deposited are based on "rectangular" (column and row formatted) data. The majority of them use the 3 main statistics packages SAS, Stata or SPSS. These are preferred original formats, as they are easily converted to ASCII and are generally platform independent.
- The goal is to reduce all data to ASCII text to facilitate the reading of the data by any program. Approx 80% of data is numeric and is easy to convert.
- More challenging input formats:
  - A decision has to be made on whether deposited PDFs should be converted to PDF/A.
  - Should SQL output be flattened (This is the choice) or be retained as is?
  - Difficult Scenario: TIFF images from microfilm. The cost of preservation might possibly exceed the cost of recreation. Though well-formed, associated metadata could be more valuable.
- Forensic effort is needed for older data-sets. Example: 1967 punch cards (sets of three cards per record) with incomplete information on the order of the cards, the interpretation of which set of digits establishes which data column. Some documentation for the interpretation of the value codes is available.
- Deposited material typically does not contain executables. The data centre may produce executables which transform deposited data into a new output format. These executables may be necessary for viewing the data. Since they are written in the statistics packages which are already in use they do not prompt additional preservation needs.

Target file formats

- The minimum number of preservation formats that are necessary to manage the full range of data types in the data centre's Collections has been identified as a list of acceptable formats.
- The choice of *Preservation Action* is pre-determined. For any file format type a pre-defined migration path into a target format is defined.
- New incoming formats might trigger the purchase of software to migrate to. Attempts are made to accommodate any incoming format and studies are not rejected for the Collection because of file format. However, studies that do not fit within the generally approved list of preserved formats may better be referred to other agencies. The data centre has a sound network of institutions with which it exchanges expertise or to which it can refer deposits.
- Hardware and software developments are monitored continuously to update target formats as needed.
- Goal: The derived formats should last indefinitely (though we tend to say that long-term is 25 years.)

What guides you in your decision making process?

Hardware

- For every Collection, for security reasons, 5 copies are kept, in multiple locations and on different carriers. A strategy has been adopted to store data on at two different forms of storage media (i.e., optical and magnetic). These are reviewed regularly and data are copied onto new media when appropriate.
- Technologically the structure is easily expandable.

The availability of hardware does not restrict the choice of *Preservation Action*s.

The acceptable formats are not constraint by software or hardware.

Physical Data Carriers

All deposits are made under a license agreement which permits the data centre to copy from the original media. The data centre is not obliged to preserve or care for original data carriers. Physical media are eliminated.

All deposits are stored in multiple locations on magnetic as well as optical media for security.

Operating Systems

- Probably approx. 30% of end-users use Linux, almost all the rest use Windows. Main software packages used by end-users and depositors are pretty much identical on all platforms.

Operating systems therefore do not limit the choice of *Preservation Action*s.

Triangle of depositors, consumers, legislation / regulation / ethical constraints:

These are the most important factors on preservation decisions.

Legislation / regulation / ethical constraints:

The 2005 altered status as Place of Deposit, a new archival status of the data centre, means a shift from the focus on **usability** to **authenticity**. In addition to a user-centred workflow an archival workflow had to be introduced, with new cost-implications. This impacts all workflows, but mostly ingest and access. A direct impact on preservation itself is that enough documentation has to be included with the preserved data sets, so that the user can, if not re-create original data sets, understand exactly what preservation processes the data centre has carried out. More than a question of legal admissibility, this is archival good practice.

Statistics and Registration Service Act 2007

Access to data which identifies individuals in a compromising way is no longer a summary offence, but incriminating.

One may have to make decisions on what can be disseminated. Unscrupulous users might amalgamate data from different sources to create a composite picture of an individual. To avoid that, data has to be redacted (through sampling, anonymisation, removal of variables or reclassification of data at a greater level of granularity (e.g. use age instead of date of birth)).

The data centre cannot control use of data and therefore might have to control access by making different versions available to different user groups.

This impacts ingest and access, but not *Preservation Action*s.

Data Protection Act

In practice the depositor has to make a statement that the deposited material complies with the Data Protection Act (i.e. individuals can not be identified from the data set or must have given consent) and while the data centre attempts verification of the depositors' assertion, it can not consistently verify this. If it appears that offending data may be included, then identifying information will not be disseminated.
In theory the data centre might be required to remove non-compliant data.[36]
This impacts ingest and access, but not *Preservation Action*s.

Freedom of Information Act

The data centre draws out a contract with the depositor which specifies the access conditions for the deposited material. Freedom of Information Act requests are redirected to the depositor. Since, in practice, depositors might not exist any longer or may not be able to be identified, the data centre has, at times, to modify access conditions without the depositor's explicit consent.
This impacts ingest and access, but not *Preservation Action*s.

Copyright Legislation

Depositors make a copyright assertion. Practically it is hard to assert factual circumstances in an academic world, since it is not always clear whether a study was authored on a scientist's personal clock or on the employer's clock.
Increasingly, copyright assertions apply to parts of the documentation (e.g. an outside company using a survey document which is protected and has copyright of its own) rather than to the whole of the survey.
This impacts ingest and access, but not *Preservation Action*s.

Internal / External factors

- There is no strong division of internal and external factors. They are mostly dictated by funding agencies.
- There is no right to deposit. Depositors may offer data for deposit. The ARC (Archive Review Committee) can decide on which studies are brought into the Collection, based on rights concerns, ethics concerns, cost, scholarly and historical value, and user accessibility.
- Also, the acquisition process is well-defined ("red folder"), which makes the need for acquisition resources for a data set transparent.

This means that accepted deposits can be planned such that there is no back-log developing, that there is no compromise on the data quality, and that service level throughput can be met.

Funding Agencies

- Most internal and external factors are dictated by funding agencies. The expectations of the funding agencies are strongly aligned with the data centre's strengths and pose no restriction on the choice of *Preservation Action*s.
- The 'institutional repository' model has changed the attitudes of some funding agencies (especially that AHRC). There may be a tendency to assume that institutions can ingest and manage their own repositories. This leads to a difference in funding, but it has not impacted the choice of *Preservation Action*s for the data centre. Having said that, the data centre is constructing a self-archiving repository for data Collections, which will a) function as part of the ingest process and b) act as a dissemination platform for researchers' work. The material archived by end-users that is not selected for permanent preservation will not be subject to any active preservation techniques.

Service Description

The service description is clearly defined and based on targets set by the funding agencies; for example deposited data should be made available within 1 month of deposit. Longer delay may be needed under special circumstances and is regulated by a prioritisation process.
The service description does not restrict preservation planning.

Personnel/ Skills

There are 60 staff in the data centre; beyond that tasks may be out-sourced to sub-contractors. Personnel numbers and skills do not pose a restriction on the choice of *Preservation Action*s.

[36] Example: The Data centre was offered (and accepted) an AHRC funded study on Members of the British Communist Party. The study has been preserved but is not disseminated because no consent has been given. It was originally thought that a "hundred year" rule could be applied but this proved impractical.

Budget
Since the data centre has control over which studies are accepted for ingest it can adjust the preservation work so that it can be managed within the existing resources.

Access restrictions
At present all material deposited is open to the data centre's staff (no access restrictions) for preservation purposes, though this may change in the future.

Standards

- The data centre strives to conform with OAIS and to use OAIS terminology; but OAIS does not in all cases fit the business process of the data centre (e.g. OAIS expects on-the-fly DIP production).
- The data centre accepts data of any of the versions of the standard statistics packages. For data exchange they are transformed to the DDI metadata standard and wrapped up into an XML preservation version. [This is a simplification, because we will accept GIS, relational databases, structured and coded texts etc.]
- DDI is generally used for descriptive metadata on the data centre's resource discovery and delivery side. It is archived but not preserved, as it is dynamic. Some of the material within this metadata may form part of the original deposit.
- METS has been used for some descriptive metadata by the data centre for resource discovery and delivery. Experiments have been made to use METS for preservation metadata, but is not yet consistently used. Preservation metadata is kept separately on the system side. The goal is to join them up to remove redundancy.

Tool support

- Data exchange tools for raw data are used (outsourced); but there is no need for decision support on which data exchange tool to use.
- A migration tool for upgrading to the newest supported version of a statistical package is being developed.

Competitors
There is no influence from competitors on the choices of *Preservation Action*s.

Peer organisations
Peer organisations substantially influence policies and strategies, but have no direct impact on individual *Preservation Action* choices.

Business context
The value of a Collection is taken into consideration when the ARC decides on which data-sets to ingest. Once the decision is made to ingest a Collection, all preservation processes will be the same as for other Collection items.

Data quality
Some data needs pre-processing. Insufficient data quality might restrict access, but will not influence the choice of *Preservation Action*s.

Remarks
The new 2008 Preservation policy document describes policy and terminology.
The strategy document (forthcoming) will describe procedure. The 2005 policy document had combined these aspects.

The hierarchy of interpretation of legal instruments is tempered by the relationship with the depositors. This is different to other institutional types who have no choice of depositor, deposited material, or agreements made. [Worth noting that the data centre explicitly requires depositor to sign a form allowing us to preserve these digital materials.]

The legal framework does not limit the choice of hardware, software, data carriers, *Preservation Action*s, or formats used. It determines why, but not how things are done.

Diligence during ingest avoids difficult preservation situations (e.g. rigorously validate and repair errors during ingest).

There is little need for an **automatic preservation planning** support tool. The choice of *Preservation Action* to take is clearly laid down in a strategy document. There is no great choice amongst tools.

Desired tools

- A ***Preservation Action*** **tool** which transforms old versions of statistics packages to newer versions on demand is being developed.
- A **preservation policy checklist** which assists in writing preservation policy documents would be useful, especially for institutional repository implementations. Parts of this might be informed by internal research at the data centre in due course.
- A **checklist** which assists in collecting all necessary facts and documents when ingesting new material would be useful. Underlying **representation information** / **registries** which let you link to information e.g. to registered license agreements with all universities would be useful.

7.2.4 National archives 1

Interview with the technical manager Digital Preservation of National Archives 1, 28 February 2008

Introductory – general question

What is the National Archives doing in the realm of digital preservation?

The National Archives accepts digital data on physical media from government agencies. There is process involving other parts of the National Archives before the data are being transferred. After this, the data are transferred to the Records Search System, where intellectual control information is added. Basically, the National Archives gets intellectual control information (for Records Search), a manifest indicating which digital objects Digital Preservation will receive, and the digital objects themselves. Links between the parts are made through unique identifiers.

At the Digital Preservation Section, there is a process that is followed to deal with the digital objects. It starts with the quarantine process that is aimed to verify whether the correct data were received, and includes a virus malware check. After this, the digital objects are copied to a device, locked away in a safe and left for 28 days. After 28 days and constant updating of virus systems, the data are virus checked again. If the check is positive, the objects are transferred to the preservation system.

In the preservation system, the file format of the objects is identified and on the basis of that, a conversion to a standard based open format will be executed. For this, our system is used.

Preferred formats

The National Archives has a range of open formats that has been selected for each genre of file type.

The four main criteria are standard-based, community developed, multiple software implementations and no patents or no property rights restrictions. Sometimes, there is more than one format that is accepted. For example, for images, the National Archives uses two formats at the moment, jpeg and png. But png is better as it includes lossless compression and it has three built-in methods of integrity checking. That's very important for the integrity of the data. All Microsoft Office formats are transformed in odf. This choice for odf is based on the four criteria.

Procedures for quality assurance

The National Archives takes a sample of objects (with at least one of each type) and this sample is analysed after conversion and compared with the original or a proxy of the original. On the basis of the manual evaluation, for which there are no clear procedures or criteria, it will be decided whether the conversion was successful or not.

Quality assurance has been automated as much as possible, but examining the objects is not automated.

Development of tools

All tools are developed in house. The National Archives developers write code in Java for plug ins.

Personnel and expertise

The National Archives adheres to some best practices for Java development. The Java code is standard compliant and all development practices are documented. Also, procedures for enhancing and bug tracking etc. are public and easy to access.

This approach reduces the risk that knowledge is concentrated too much.

Collection and ingest

All records are acquired and accepted, and there is no difference in value for these records. All have to be preserved.

After records are ingested, they are immediately analysed and converted. This approach has the advantage that problems are identified as close to the moment of ingest, so that immediate action can be taken.

Costs of digital preservation

There have not been any cost or budget issues at the moment. Operational costs are no issue. There is prioritising of developments of code writing, but money has not been a limiting factor in digital preservation at the National Archives.

Legislation

The Archives Act indicates that the National Archives have the same responsibility for digital information as for paper records. This responsibility is to preserve it and make it accessible. How we do this is an institutional decision.

User input

User needs are secondary to preservation needs. User needs are related to accessibility. In the National Archives, the user experience is separated from preservation.

Authenticity/integrity

The institution has established a work flow that is able to show for each step in the process what has been done. Various things are recorded.

Conclusions

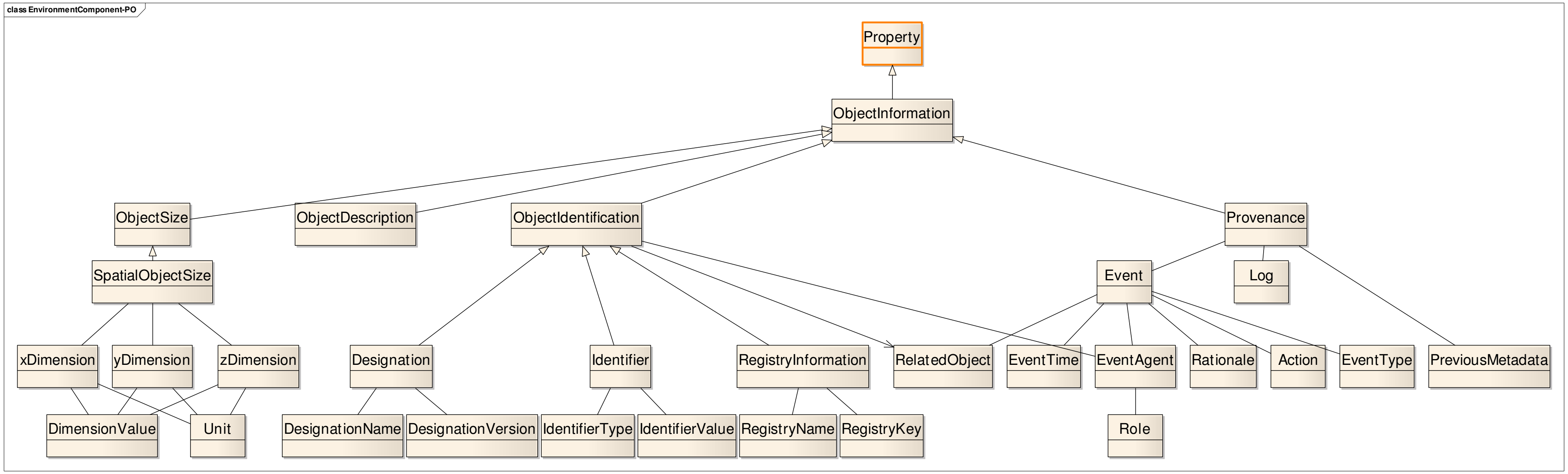
Main issues are automation of processes that are still done manually at the moment, and development of new plug-ins for conversion of file formats that are not yet included in the preservation system.

Budget, users, legislation, etc. are no real issues or are not inhibiting digital preservation.

7.3 Vocabulary for specifying Properties

In the following we list an initial collection of *Property* vocabulary for a subset of the *Environment Component Type* - *Preservation Object Type* combinations. The goal is to have a deep vocabulary that would be generally acceptable and sharable by different institutions. The current state of the vocabulary is a first phase attempt. More work is needed to expand it, validate it, and harvest community input. For certain subsets one can refer to related work. For example, the PREMIS preservation metadata defines Properties on a Manifestation (≈ PREMIS Representation), and Bytestream (≈ PREMIS File and Bitstream).

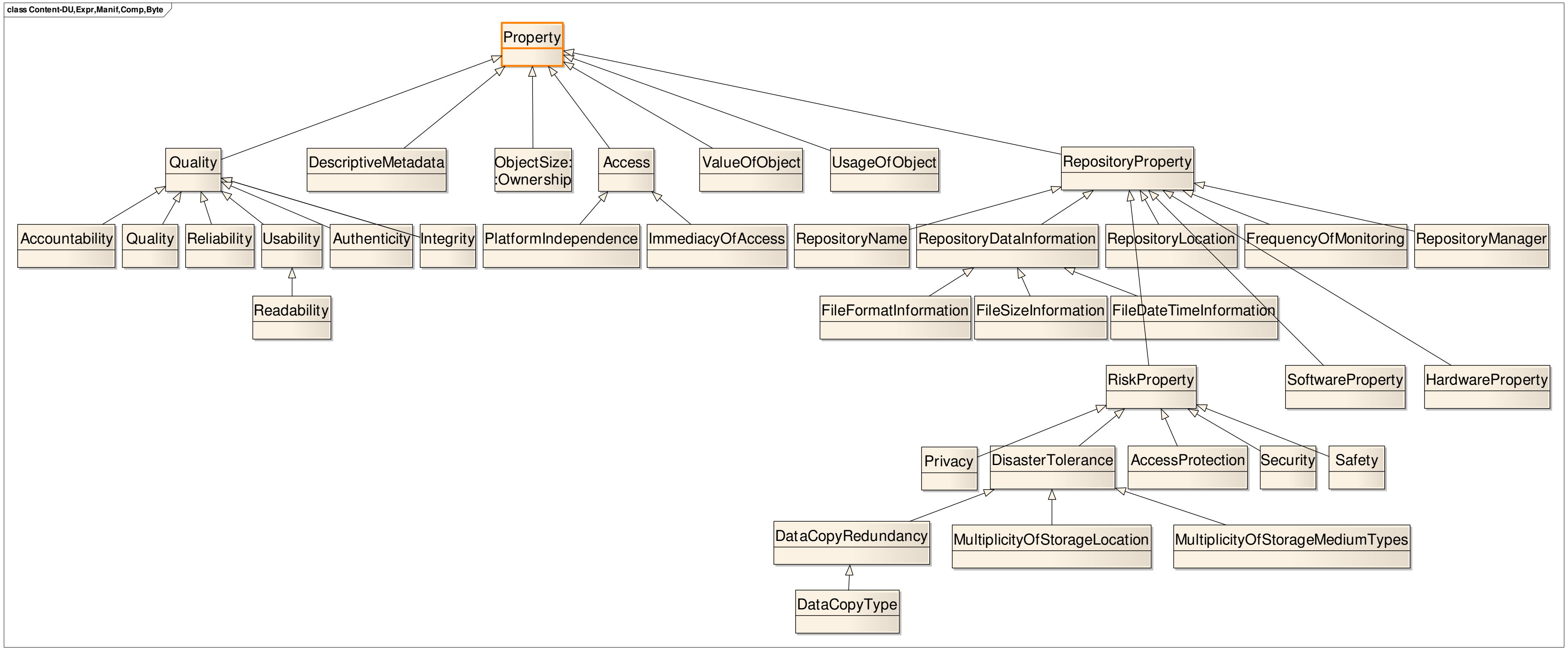
Vocabulary for any *Environment Component* of any *Preservation Object*



Every *Preservation Object Type* inherits these *Properties* from the general *Preservation Object*. A *Collection*, for example, might have collection-specific *Properties* such as TimeRange and Subject, and inherit *Properties* such as ObjectSize (i.e. CollectionSize in this context).

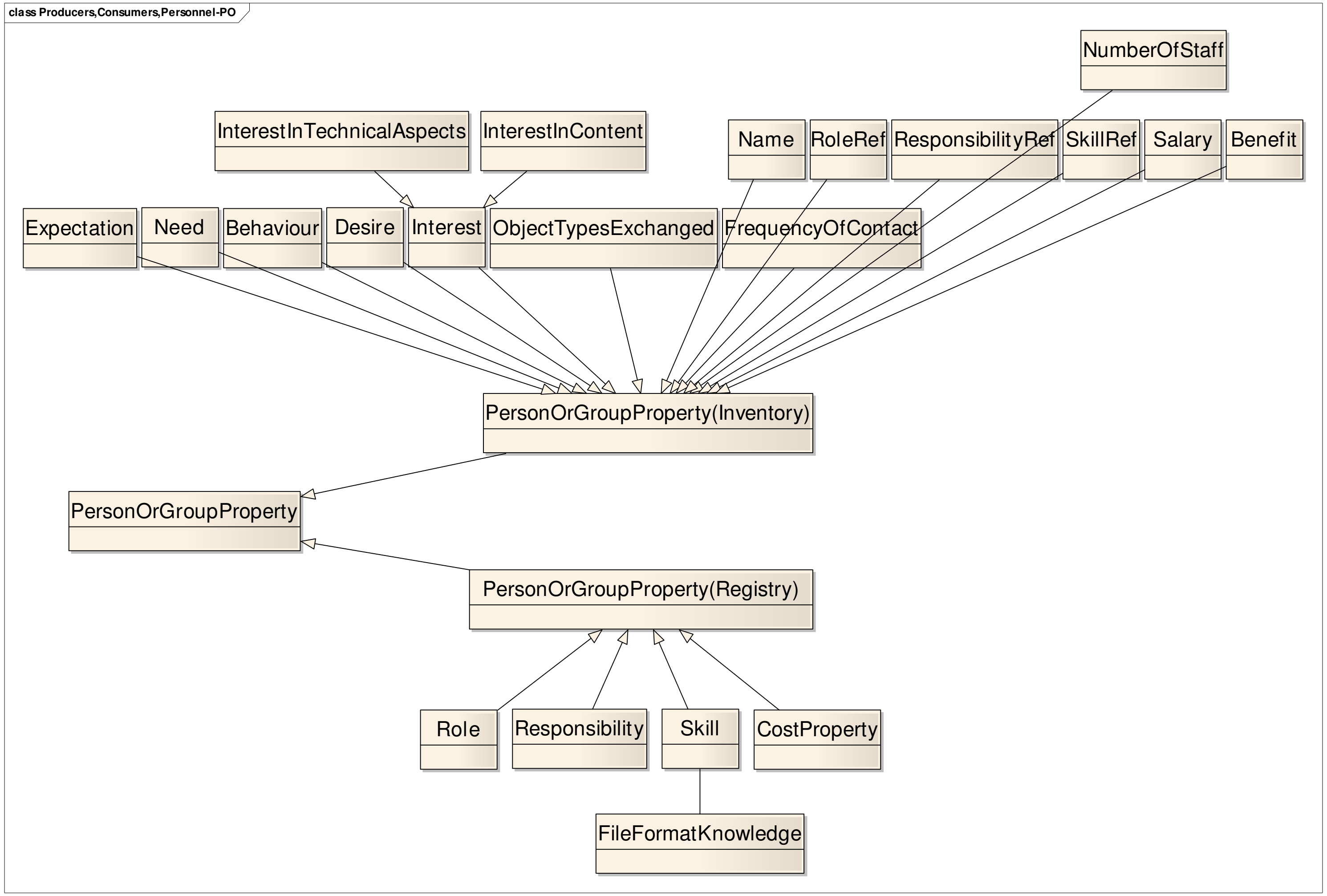
Vocabulary for the Content/Self *Environment Component* of any Collection, Deliverable Unit, Expression, Component, Manifestation or Bytestream

Our model does not prescribe which descriptive metadata an institution should use and the vocabulary does **not** explicitly list traditional descriptive metadata for digital objects, since they are described in great depth elsewhere[1,2,3, etc.] Rather, it provides for an extension point. The model might refer to descriptive metadata, however, in order to express a condition in a requirement. “Publisher”, for example, is a typical descriptive metadata element, which might be taken from the MODS metadata framework[3]. A requirement might use this element. Equally, institutions can write their own requirements referring to their own metadata schema of choice. An example requirement might be:“If the publisher is Elsevier then normalise the Bytestream using the“Elsevier_Normaliser2.0” tool.The machine-interpretable representation of this requirement might use the MODS concept “publisher”: If MODS:publisher = “Elsevier” Then PreservationActionTool= “Elsevier_Normaliser2.0.”

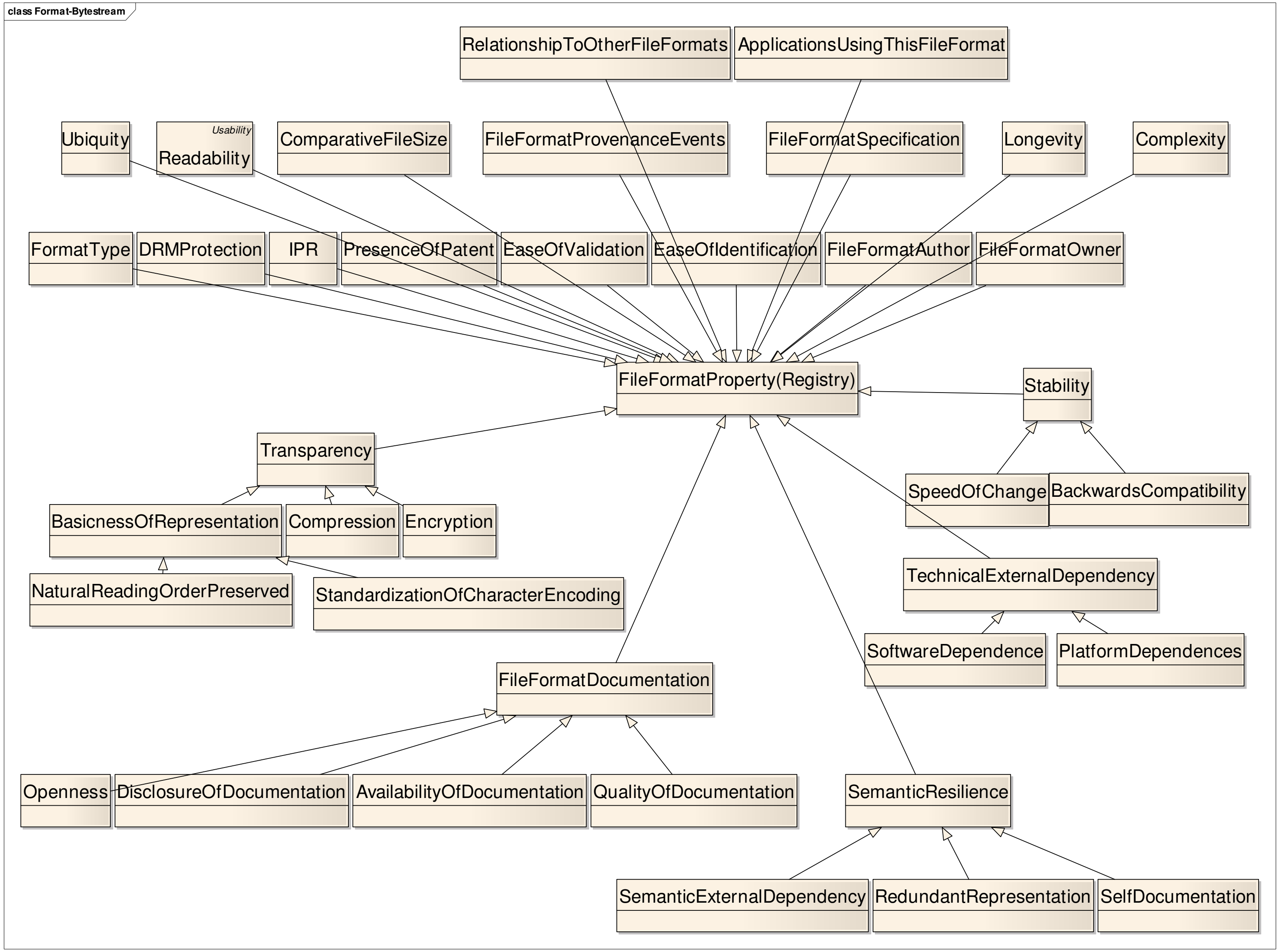

class Content-DU,Expr,Manif,Comp,Byte
Property
Quality
DescriptiveMetadata
ObjectSize:
:Ownership
Access
ValueOfObject
UsageOfObject
RepositoryProperty
Accountability
Quality
Reliability
Usability
Authenticity
Integrity
PlatformIndependence
ImmediacyOfAccess
RepositoryName
RepositoryDataInformation
RepositoryLocation
FrequencyOfMonitoring
RepositoryManager
Readability
FileFormatInformation
FileSizeInformation
FileDateTimeInformation
RiskProperty
SoftwareProperty
HardwareProperty
Privacy
DisasterTolerance
AccessProtection
Security
Safety
DataCopyRedundancy
MultiplicityOfStorageLocation
MultiplicityOfStorageMediumTypes
DataCopyType

Vocabulary for the Producer, Consumer and Personnel *Environment Components* of any *Preservation Object*

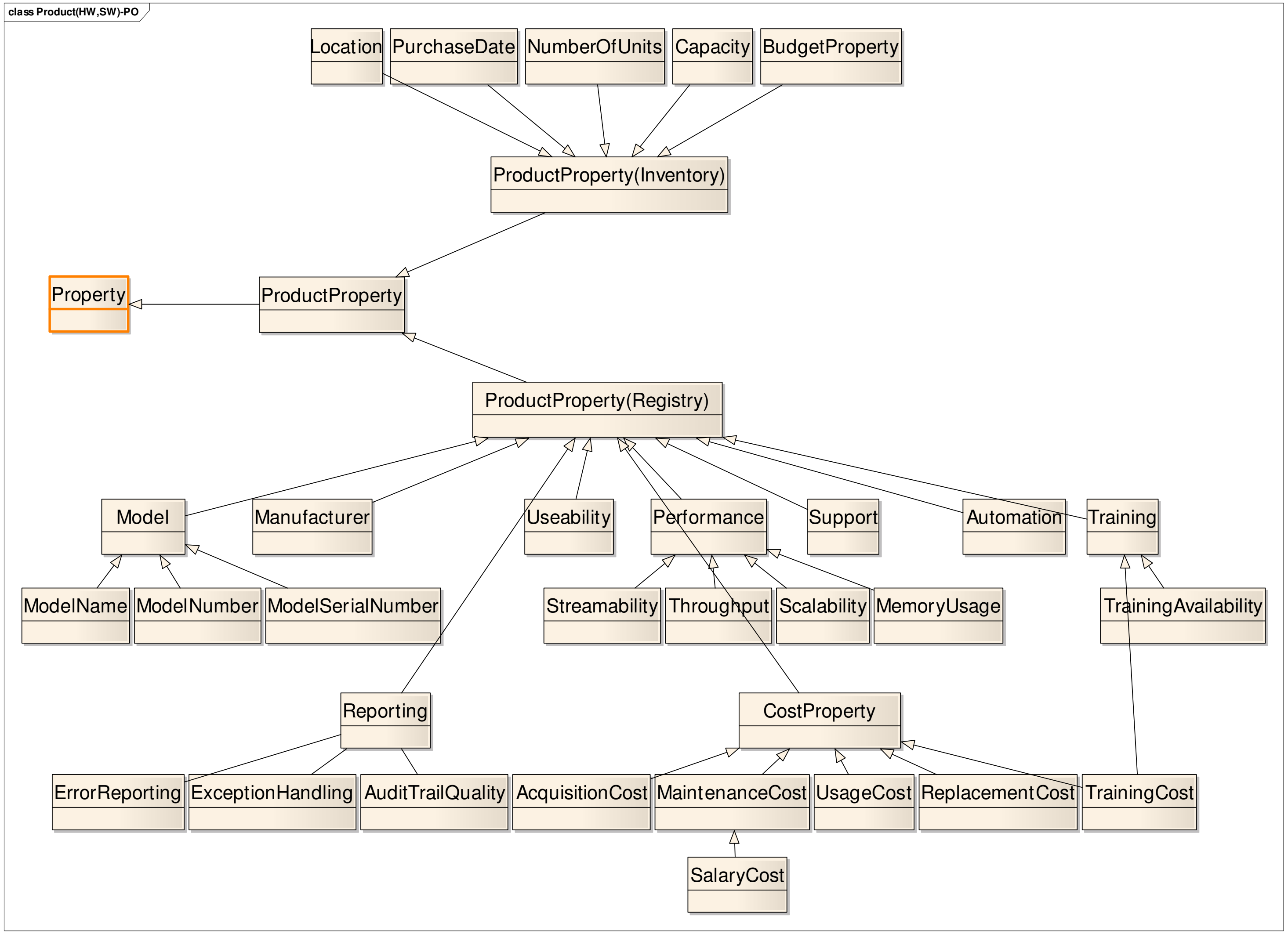
Properties, such as the *Identification*, *Description* and *Size* of the group can be represented in the *ObjectInformation Property* that every *Preservation Object* has.



Vocabulary for the Format *Environment Component* of any Bytestream *Preservation Object*

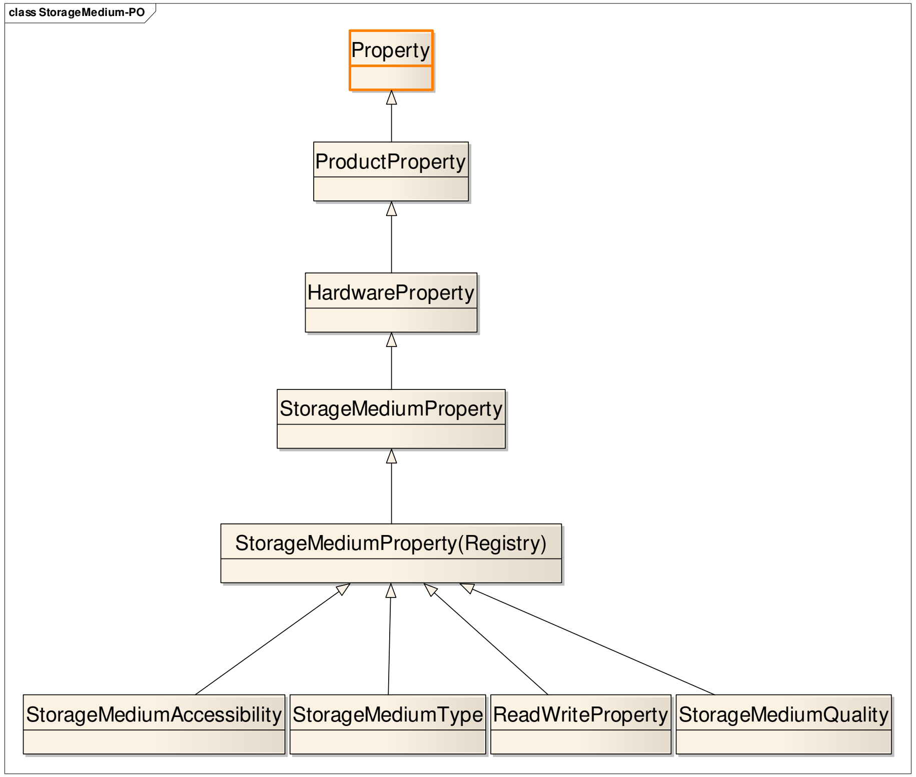


Product vocabulary for the Software or Hardware *Environment Component* of any *Preservation Object*

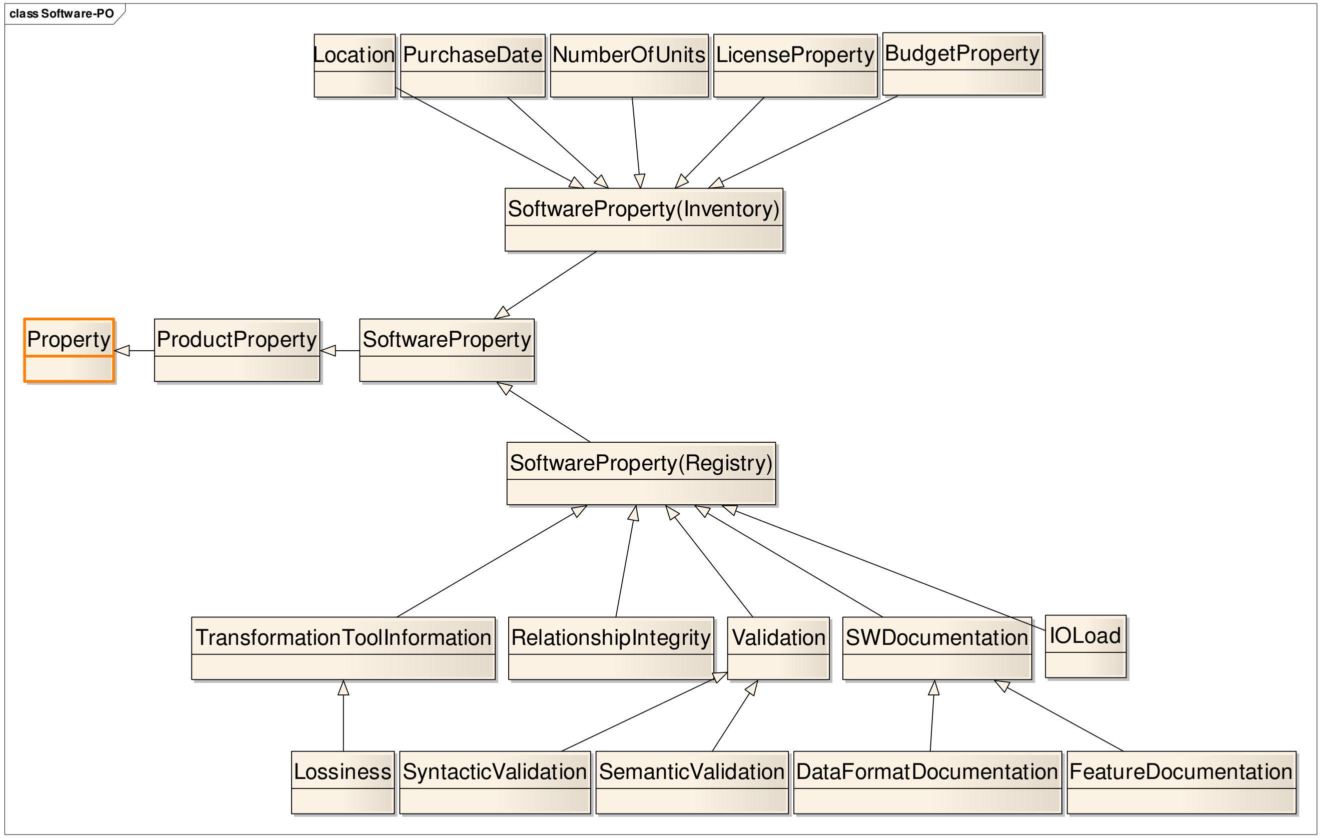


Vocabulary for the Storage Medium *Environment Component* of any *Preservation Object*

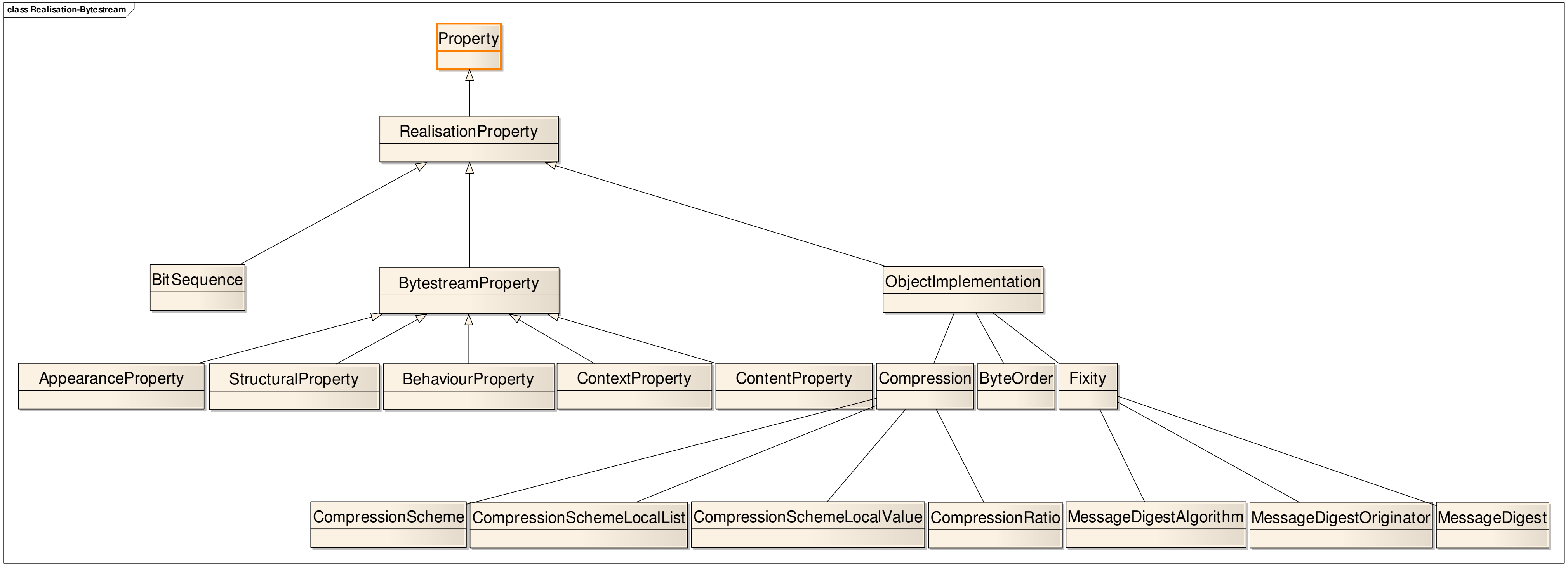
Properties are inherited from Hardware *Environment Component*s and Products.

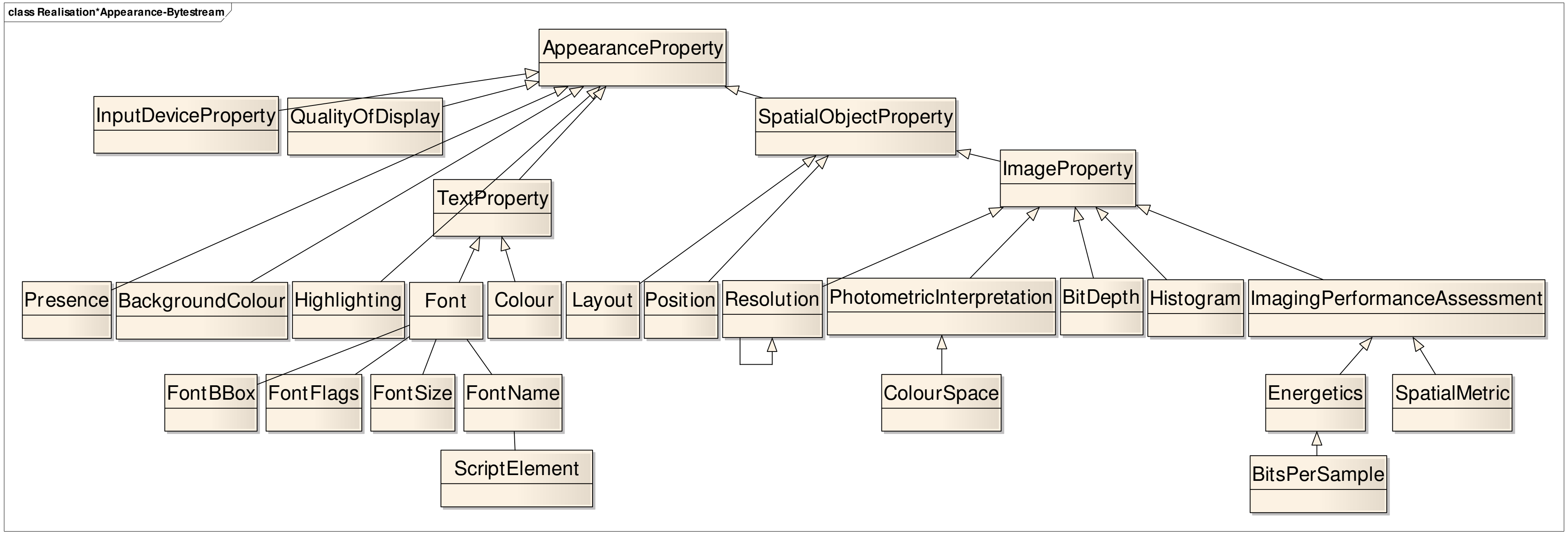


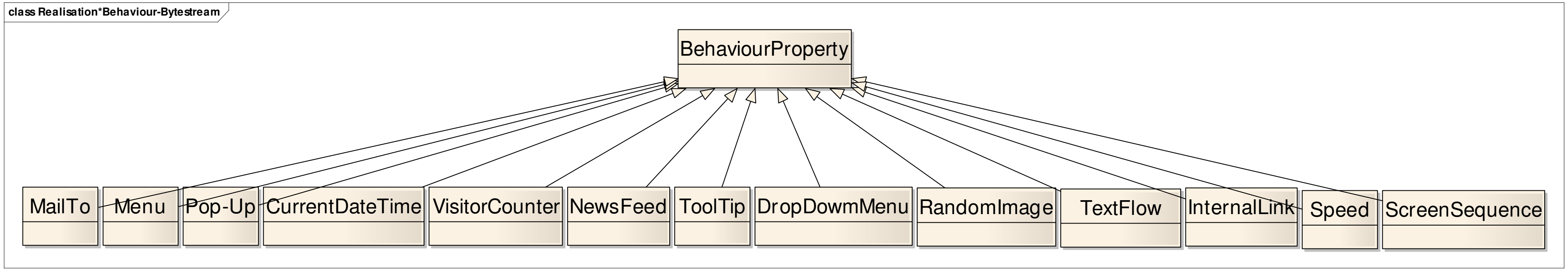
Vocabulary for the Software *Environment Component* of any *Preservation Object*

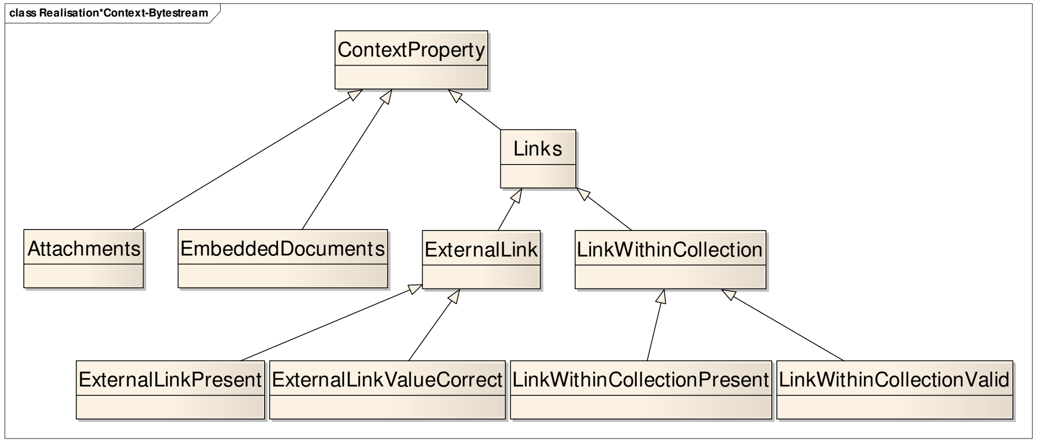


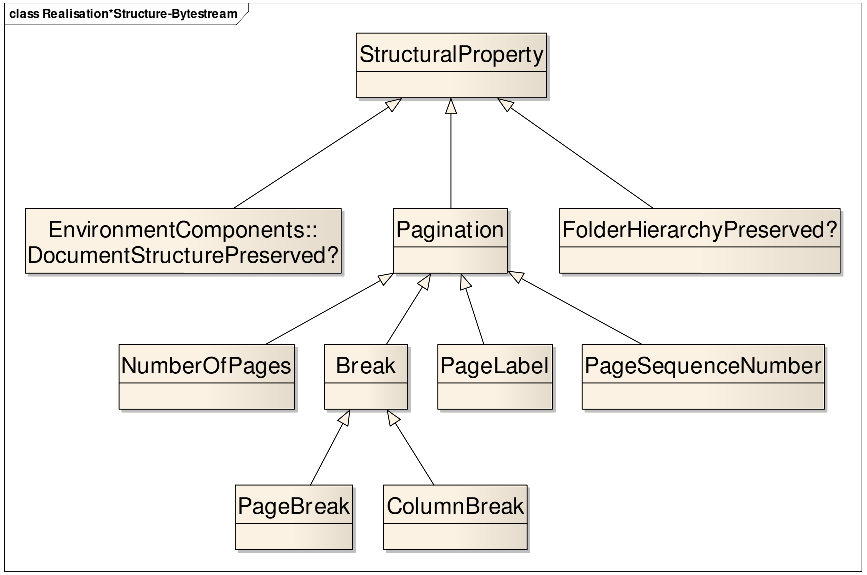
Vocabulary for the Realisation *Environment Component* of a Bytestream *Preservation Object*

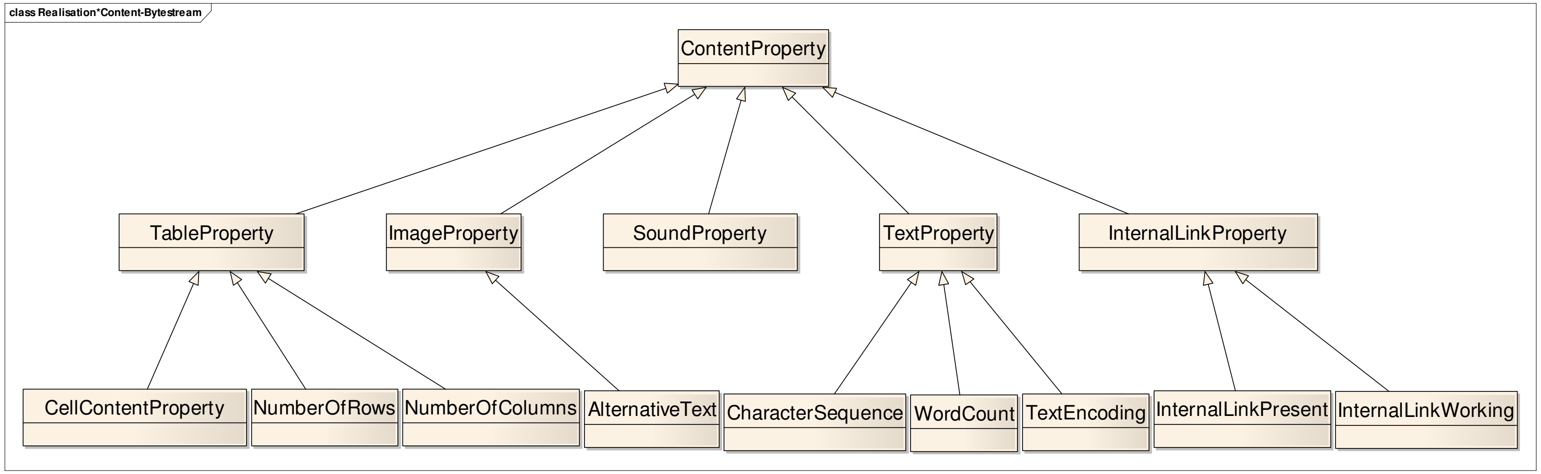



class Realisation*Appearance-Bytestream
AppearanceProperty
InputDeviceProperty
QualityOfDisplay
SpatialObjectProperty
ImageProperty
TextProperty
Presence
BackgroundColour
Highlighting
Font
Colour
Layout
Position
Resolution
PhotometricInterpretation
BitDepth
Histogram
ImagingPerformanceAssessment
FontBBox
FontFlags
FontSize
FontName
ColourSpace
Energetics
SpatialMetric
ScriptElement
BitsPerSample


class Realisation*Behaviour-Bytestream
BehaviourProperty
MailTo
Menu
Pop-Up
CurrentDateTime
VisitorCounter
NewsFeed
ToolTip
DropDowmMenu
RandomImage
TextFlow
InternalLink
Speed
ScreenSequence


class Realisation*Context-Bytestream
ContextProperty
Links
Attachments
EmbeddedDocuments
ExternalLink
LinkWithinCollection
ExternalLinkPresent
ExternalLinkValueCorrect
LinkWithinCollectionPresent
LinkWithinCollectionValid


class Realisation*Structure-Bytestream
StructuralProperty
EnvironmentComponents::
DocumentStructurePreserved?
Pagination
FolderHierarchyPreserved?
NumberOfPages
Break
PageLabel
PageSequenceNumber
PageBreak
ColumnBreak


class Realisation*Content-Bytestream
ContentProperty
TableProperty
ImageProperty
SoundProperty
TextProperty
InternalLinkProperty
CellContentProperty
NumberOfRows
NumberOfColumns
AlternativeText
CharacterSequence
WordCount
TextEncoding
InternalLinkPresent
InternalLinkWorking

7.4 Citations

Digital Preservation Policies can be accessed for the following institutions (non-exhaustive list). It should be noted that these documents cover a variety of issues, and vary widely in depth and detail.

National Library of Australia (Australia) Digital preservation policy
http://www.nla.gov.au/policy/digpres.html

National Library of Australia (Australia) Digital preservation strategies
http://www.nla.gov.au/padi/topics/18.html

State Library of Victoria (Australia)
http://www.slv.vic.gov.au/about/information/policies/digitalpreservation.html

National Archives of Australia (Australia)
http://www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper.pdf

British Library (UK) http://www.bl.uk/about/Collectioncare/pdf/bldppolicy1102.pdf

British Library (UK) Digital Preservation Strategy
http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digpresstrat.pdf

British Library (UK) Digital Preservation Plan for Microsoft Digitisation
http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digpresmicro.pdf

British Library (UK) Digital Preservation Plan for eJournals
http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digpresejournal.pdf

British Library (UK) Risk Assessment 2007
http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/riskassessment.pdf

National Library of Wales (UK)
http://www.llgc.org.uk/fileadmin/documents/pdf/digital_preservation_policy_and_strategy_S.pdf

Hampshire Record Office (UK) http://www.hants.gov.uk/record-office/policies/digital.html

UK Data Archive (UK) http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0905.pdf

UK Data Archive (UK) Preservation Policy 03 2008 http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0308.pdf

Arts and Humanities Data Service (UK) http://ahds.ac.uk/documents/colls-policy-preservation-v1.pdf

University of Sterling (UK) Index to policy documents http://www.is.stir.ac.uk/aboutis/policy.php

The Digital Archives of Georgia (USA)
http://www.sos.state.ga.us/archives/who_are_we/rims/digital_History/policies/policy%20-%20Digital%20Preservation%20Policy.pdf

North Carolina, Department of Cultural Resources (USA)
http://statelibrary.dcr.state.nc.us/digidocs/policy_framework.pdf

Cornell University Library (USA) http://commondepository.library.cornell.edu/cul-dp-framework.pdf

Columbia University Library (USA)
http://www.columbia.edu/cu/lweb/services/preservation/dlpolicy.html

Florida Digital Archive (USA) http://www.fcla.edu/digitalArchive/pdfs/DigitalArchivePolicyGuide.pdf

In addition, some examples of practical guidelines:

eDavid-project http://www.expertisecentrumdavid.be/eng/index.php

ICTU series ‘Van digitale vluchtigheid naar digitale houvast’
http://www.digitaleduurzaamheid.nl/bibliotheek/docs/bewaren_van_email.pdf
http://www.digitaleduurzaamheid.nl/bibliotheek/docs/bewaren_van_tekstdocumenten.pdf

http://www.digitaleduurzaamheid.nl/bibliotheek/docs/Bewaren_van_spreadsheets.pdf
http://www.digitaleduurzaamheid.nl/bibliotheek/Bewaren_van_databases.pdf

Some additional resources can be found here:

stratML http://www.xml.gov/stratml/index.htm

Reference Model for an Open Archival Information System (OAIS)
http://public.ccsds.org/publications/archive/650x0b1.pdf

ISO 15489: 2001 Records Management standard
http://www.iso.org/iso/catalogue_detail?csnumber=31908

Domea-Konzept
http://www.kbst.bund.de/cln_006/nn_836960/Content/Standards/Domea__Konzept/domea__node.html__nnn=true

DIRKS http://www.records.nsw.gov.au/recordkeeping/dirks-manual-print_1923.asp

RLG-OCLC audit tool (2002) http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf

Trustworthy Repositories Audit & Certification (TRAC) Criteria and Checklist (2007)
http://www.crl.edu/PDF/trac.pdf

nestor Catalogue of Criteria for Trusted Digital Repositories (2006) http://www.nbn-resolving.de?urn:nbn:de:0008-2006060703

Ten basic *Characteristic*s of digital preservation repositories
http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=92

DCC/DPE Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)
http://www.repositoryaudit.eu/

Noark-4, functional requirements http://www.riksarkivet.no/noark-4/Noark-eng.pdf

InterPARES models www.interpares.org

InterPARES2 project glossary
http://www.interpares.org/ip2/display_file.cfm?doc=ip2_glossary.pdf&CFID=243105&CFTOKEN=70677126

InterPARES2 project Business-Driven Recordkeeping (BDR) model
http://www.interpares.org/display_file.cfm?doc=ip2_BDR_model(consultation_draft_20070730).pdf

JISC records management http://www.jiscinfonet.ac.uk/InfoKits/records-management

JISC: Digital Preservation briefing paper
http://www.jisc.ac.uk/publications/publications/pub_digipreservationbp.aspx

JISC/NPO studies on the preservation of electronic materials
http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/rept011.pdf

PP2-D1: Report on policy and strategy models – result of the first iteration, 9/11/2007, Internal Planets report at http://www.planets-project.eu/private/pages/wiki/index.php/Planets_Internal_Deliverables_1st_June_2007_-_31st_November_2008

Cornell University Library: Digital Preservation Management: implementing short-term strategies for long-term problems http://www.library.cornell.edu/iris/tutorial/dpm/terminology/strategies.html

Digital Preservation an Overview, 2007 Seamus Ross
http://www.dpc.delos.info/ss07/attendees/delos_SS2007_SR_DPanIntroduction.pdf

Report on Comparison of Planets with OAIS, PP/7-D1 Internal Planets Deliverable

Sustainability of Digital Formats, Planning for Library of Congress Collections
http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

Kenneth Thibodeau: "If you build it will it fly?", *Journal of Digital Information*, Vol 8, No 2, (2007)
http://journals.tdl.org/jodi/article/view/197/174

ISO/TR 18492:2005 http://manage.committees.standards.org.au/COMMITTEES/IT-021/N0001/ISO-TR_18492-2005.pdf

DPC: Mind the Gap: Digital Preservation Needs in the UK
http://www.ariadne.ac.uk/issue48/semple-jones/ and
http://www.dpconline.org/graphics/reports/mindthegap.html (full report)

BL LIFE cost models http://www.life.ac.uk/

ERPANET: Digital Preservation Policy Tool
http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf

DPE http://www.digitalpreservationeurope.eu/

TASI http://www.tasi.ac.uk/advice/delivering/digpres2.html

eDavid-project http://www.expertisecentrumdavid.be/eng/index.php

S. Strodl, C. Becker, R. Neumayer, A. Rauber (2007). How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure, Proceedings of the 2007 conference on Digital libraries, ACM, p. 29 - 38

Ayre and A. Muir: "Right to Preserve? Copyright and licensing for digital preservation project". Final Report, Loughborough, 2004
http://www.lboro.ac.uk/departments/dis/disresearch/CLDP/DOCUMENTS/Final%20report.doc

F. Boudrez: "E-mails: hoe klasseren en goed archiveren?" Antwerp, 2006
http://www.expertisecentrumdavid.be/docs/emailrapport_lr.pdf

Solinet: Contents of a Digital Preservation Policy
http://www.solinet.net/preservation/preservation_templ.cfm?doc_id=3678

M. Guercio, L. Lograno, A. Battistelli and F. Marini: "Legislation, Rules and Policies for the Preservation of Digital Resources, a survey". (Draft)
http://eprints.erpanet.org/65/01/Dossier1_English_version_Full.pdf

R. Pearce-Moses: "A glossary of archival & records terminology". Chicago, 2005.

Policies for Digital Preservation” Seminar Report ERPANET Training Seminar, Paris January 29-30, 2003 http://www.erpanet.org/events/2003/paris/ERPAtraining-Paris_Report.pdf

Journal Archiving and Interchange Tag Suite (NLM DTD), National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM) http://dtd.nlm.nih.gov/

AONS - An obsolescence detection and notification service for Web archives and digital repositories http://www.informaworld.com/smpp/content~content=a780448483~db=all~order=page
http://www.informaworld.com/smpp/content~content=a780448483~db=all~order=page

RAND report on digital preservation Addressing the uncertain future of preserving the past
http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/rand_report_e-depot_TR510_3c_Cover.pdf