Preservation Policy Levels in SCAPE

Barbara Sierman
KB National Library of the Netherlands
PO Box 90407
2509 LK The Hague
+31 70 314 01 09
Barbara.Sierman@KB.nl

Catherine Jones
Science and Technology Facilities Council
Harwell Oxford, Didcot OX11 0QX
+44 1235 445402
Catherine.jones@stfc.ac.uk

Sean Bechhofer
University of Manchester
Kilburn Building, Oxford Road
Manchester M13 9PL
+44 161 274 6282
sean.bechhofer@manchester.ac.uk

Gry Elstrøm
State and University Library Denmark
Victor Albecks Vej 1
8000 Aarhus C
+45 8946 2314
gve@statsbiblioteket.dk

ABSTRACT

This paper describes the Preservation Policy model as designed in the European project SCAPE and an experiment to test the viability of the model against two real life preservation policies.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Information Storage and Retrieval – *Digital Libraries; I.2.4* [**Computing Methodologies**]*:*Artificial Intelligence – *Knowledge Representation Formalisms and Methods*

Keywords

Digital preservation, policies, watch, planning.

1. INTRODUCTION

There is a shared recognition that the existence of preservation policies for long term digital preservation is important. Not only because it is for example stated in the ISO standard 16363 Audit and Certification of Trustworthy Digital Repositories, but also because digital preservation needs a well defined underlying basis. The creation of these policies seems to be rather difficult and we see that organizations are struggling to write them. Many organizations who are preserving collections for the long term have not yet published their policy on their website. While these organizations often have a legal mandate and are funded by public money, the general public does not know how these digital collections are treated. Nor can they see how these organizations plan to handle various challenges.

A preservation policy is a "Written statement authorized by the repository management that describes the approach to be taken by the repository for the preservation of objects accessioned into the repository".[1]

Preservation Policies are not a goal in itself, they are there to support the activities of the organisation with respect to the maintenance and preservation of the digital collection. "Without a policy framework a digital library is little more than a container for content" [5] . In an ideal situation, the preservation policies will guide the preservation activities in an organisation. As the field in which the organizations act is rapidly changing, and the insights in digital preservation change, the preservation policy documents should be a regularly revised and updated.

The European project SCAPE has designed a Preservation Policy Model that will support organizations to build their preservation policy documents. Before this, several European projects investigated preservation policies. These results are input for the current work in the SCAPE project.

The DL.org project investigated "interoperability" as an important means to enable digital libraries to get the most value out of their collections and to enable "sharing" and "building by re-use". By being "interoperable" on various aspects, it would be possible to share collections and to collaborate between organisations. Digital libraries is here more broadly defined, not restricted to digital libraries in a traditional sense, but to "a potentially virtual organisation, that comprehensively collects, manages and preserves for the long depth of time rich digital content, and offers to its target user communities specialised functionality on that content, of defined quality and according to comprehensive codified policies [4]. One of the areas for interoperability identified in this report is "preservation policies", for which the DL.org project designed a conceptual approach.

The PLANETS project introduced the "preservation guiding document" [6] including a conceptual model and a vocabulary for preservation guiding documents. The key focus was the digital collection and the risks that might threaten that collection. The

[1] http://www.alliancepermanentaccess.org/index.php/knowledge-base/member-resources/digital-preservation-glossary/

preservation object, within a digital collection, has characteristics and lives in an environment. The identification of a preservation risk will lead to a preservation action, that takes into account the characteristics of the object and the environment in order to formulate requirements.

The Shaman project defined a number of catalogues and processes needed in digital preservation from the business governance viewpoint, such as a Policy Catalogue that provides a list of all the preservation policies, a Driver/policy/goal/objective Catalogue that provides a breakdown of preservation drivers, policies, goals and objectives within the organisation. Further a Contract/measure Catalogue: providing the list of all policies and associated strategies and finally the Preservation Management Processes representing the processes which manage the preservation in the organization [1].

The SCAPE project is dedicated to the challenges of large scale, heterogeneous collections of complex digital objects. The digital objects are held in the collections of various participating content holders, like libraries, web archives and data centres. The scale of these digital collections implies that preservation activities that need to be performed will limit the possibility of manual involvement, and require more automation through the use of workflows and high-performance systems. Preservation activities need to be guided by a preservation policy.

The SCAPE project will run until 2014. The experiment described in this article is an intermediate result that gave us input to shape further work. The scope in this experiment has been limited to preservation policies that are relevant for preservation watch and preservation planning.

2. PRESERVATION AREAS

Preservation Policies will guide Preservation Actions. In digital preservation however, a preservation action will often be preceded by an identified risk, based on monitoring several areas of interest, and a combination of the outcomes leading to a decision to act. The identification of the most appropriate action is done in the Preservation Planning process, which produces a preservation plan. Enacting the preservation plan will result in the Preservation Action. In SCAPE the Preservation Watch area will be enriched by the SCOUT system [9]. SCOUT is an automatic preservation watch system that will detect preservation risks and opportunities. The Preservation Planning will be extended by new versions of Preservation Planning tool PLATO[2]. In both cases, a detailed level of preservation policies will be needed to enable the planning and watch services to act according to a specific set of institutional preservation policies.

2.1 Preservation Watch

In the Planets project an extension of the OAIS model was designed, the Planets Functional View [14], in which special attention was paid to a Preservation Watch function that brings together several monitoring functions. One could imagine that in case of large collections, not all the areas to be monitored can be covered by activities, done manually by humans. Instead an organisation should identify which elements should be monitored and this information could then be fed into an automatic monitoring system. The focus will be determined by the content of the preservation policies. Take for example a preservation policy that would limit the diversity of file formats that an organisation is willing to accept. Monitoring the developments related to file formats can then be restricted to the file formats that are allowed and subsequently be automatically monitored.

2.2 Preservation Planning

Preservation Planning is another area where preservation policies provide important input. If one wants to plan preservation actions that can support the long term preservation of a digital collection, input for this process should come from the preservation policies that are related to the digital material as defined by the organisation and its goals [3].

3. SCAPE PRESERVATION POLICY MODEL

3.1 Policy levels

The SCAPE Preservation Policy Model consists of three preservation policy levels that will support an organisation to create their preservation policies set. By connecting these three levels and identifying clearly which level is fit for which purpose, we intend to make the creation of a preservation policy for organizations more straightforward.

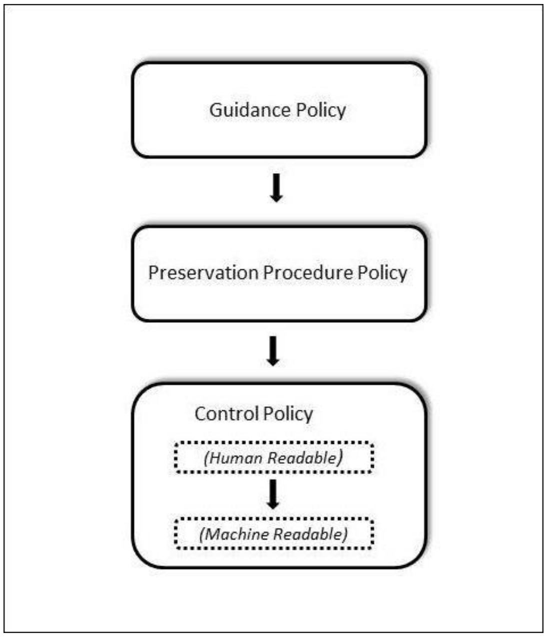

Figure 1 SCAPE Preservation Policy Model

The three levels of policies identified in SCAPE are:

1. **High level or guidance polices.** On this level the organisation describes the general long term preservation goals of the organisation for its digital collection(s). One example is that an organization decides to act according the OAIS model.

[2] http://www.ifs.tuwien.ac.at/dp/plato/intro.html

2. **Preservation Procedure policies.** These policies describe the approach the organisation will take in order to achieve the goals as stated on the higher level. They will be detailed enough to be input for processes and workflow design but can or will be at the same time concerned with the collection in general. These are likely to be made publically available.

3. **Control policies**. On this level the policies formulate the requirements for a specific collection, a specific preservation action, for a specific designated community This level can be human readable, but should also be machine readable and thus can be used in automated planning and watch tools to ensure that preservation actions and workflows chosen meet the specific requirements identified for that digital collection. These are likely to be kept internally within the organisation.

It is the interaction between the Preservation Procedure level and the Control Policy level that is the focal point of study. How much information is enough to transform the decisions and statements in the Guidance Policies and the Preservation Procedure Policies into actionable Control Policies.

3.2 Control Policy Model

The control policies created through the translation of natural language policy are intended to capture the whole policy intent, enabling automatic checking of the state of the world in watch or potential preservation plan in planning. They provide the local organisational environment within generic tools and ensure that these automated tools are not concerning themselves with areas which the organisation is not interested in; honing the tools to the specific circumstance. By using a standard model to represent this information, then two separate tools can use the same policy basis to achieve different aims enabling policy interoperation. This is the SCAPE Control Policy Model (figure 2.)

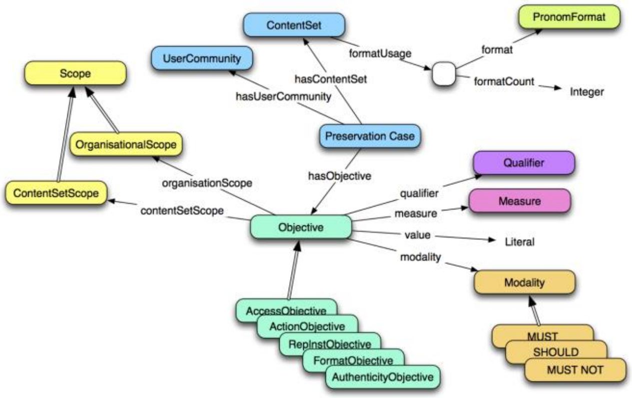

Figure 2 Overview of Control Policy Model

The SCAPE Control Policy Model provides a controlled vocabulary or set of terms and relationships that allow for the description of policies. A key aspect here is that the control policies are expressed in a, unambiguous, machine readable way, rather than as natural language. A policy that states (in English) that "Most formats used must be ISO standardised" is potentially open to interpretation -- what do we mean by "most formats" or even "ISO standardisation"? The controlled policy vocabulary provides a common set of terms that can be used, and on whose interpretation there is a shared agreement. The states of affairs that the objectives define and describe can then be tested or evaluated through some automated processes (without an agreement on the interpretation of terms it is very difficult, if not impossible, to automate this). For example, the policy above states that most formats used for a particular content set must be ISO standardised. A content profiler, such as the c3po tool[3], can analyse document collections and provide information about the formats used in that collection. Format registries (e.g. PRONOM[4]) provide detailed information about the characteristics of formats. By integrating all this information along with an unambiguous interpretation of the policy, the conditions expressed in the policy can be automatically checked, and suitable actions planned. Further advantages of a machine readable policy expression include the ability to validate or check for conflicting or subsuming policies.

The Control Policy Model provides vocabulary that is used to describe particular domain entities: situations, formats, content sets etc. Key entities described in the model are Content Sets, Objectives and Preservation Cases. A Content Set represents a collection of objects that are the focus of the policy. Objectives are the atomic building blocks of the policies. In general, an Objective will refer to a property (see below) along with a value for the property and a Modality that indicates whether or not the expected value is an absolute requirement or prohibition, expressed as MUST/MUST NOT/SHOULD etc.[5] Objectives are generic in that they describe states of affairs without referring to specific content sets or organisations. This facilitates the sharing of Objectives across policies. A Preservation Case ties objectives to a Content Set and intended User Community. Objectives may refer to properties that representations of content have; properties of the formats themselves; tools used and so on.

The properties in Objectives are taken from a collection of measures[6] -- properties that describe particular characteristics of items, formats or actions. For example, "Number of free tools that are open source"[7] is a measure that gives some indicator for the adoption of a format. Measures are further organised into "attributes"[8] -- collection of measures relating to particular characteristics and "categories"[9] -- high level groupings of attributes. A number of measures have been defined by the SCAPE project. In the future we expect measures to be shared across communities -- improving opportunities for sharing and exchange of practice. It may also be the case that particular domains or organisations will want to define their own particular measures -- extending the vocabulary in this way is possible.

Note that the model is simply there to enable the objectives to be stated in an unambiguous way. The model itself does not attempt to check whether or not the statements are true. Such checking will be done by other tools (for example the PLATO planning tool). Further details of the policy models and their use in the SCAPE preservation ecosystem are discussed in [11].

The Control Policy Model of SCAPE uses the W3C's family of representation languages. The models are defined as OWL [12]

[3] http://ifs.tuwien.ac.at/imp/c3po

[4] http://www.nationalarchives.gov.uk/PRONOM/

ontologies, with particular objectives being represented as an RDF [13] knowledge base. This use of standardised representations allows the possibility of existing tools to support the creation, management and manipulation of the policy instances.

Tools that support the user in defining policies using the control policy model are essential -- we cannot expect users to work directly with representations such as RDF. The model itself assists in this process as it can provide constraints as to what users can express, controlling and focusing the expression of the policies. A prototype web application that supports the user in defining objectives has been developed. As we discuss below, however, the process of moving from a high level expression to the specific control policy elements is non-trivial.

4. Verification of the Model using two real life Policies

Having defined the SCAPE policy model, we have verified this approach by using existing policy documents from two of the SCAPE partners to create control policies, both in human and machine readable forms

We used the policies of the State and University Library Denmark and the ISIS Data Management Policy of the Science and Technologies Facilities Council.

Although these policies could not strictly be categorized as either a Guidance Policy or a Preservation Procedure Policy, they were the currently available information with respect to the preservation intentions of both organizations and would reflect the situation in many organizations.

4.1 Policies at the State and University Library

A few years ago the State and University Library created a Digital Preservation Policy (DP Policy[7]) and a Digital Preservation Strategy [8]. The DP policy is at a very high level declaring the purpose and scope of the State and University Library's digital preservation. The DP Policy works at a management level and consists of very general statements. It is revised once a year.

In addition to this policy the State and University Library developed a DP Strategy. This details the high level policies formulated in the DP Policy and is concerned with the overall collection management. It does not specify anything about specific collections but defines how to make the right decisions according to the State and University Library policies. For instance the DP Strategy does not specify precisely what format to use for a specific collection, instead it states that the choice of format for a specific collection must be in line with the policies in the DP Strategy, in the case of formats it must be an open format, it must be well-documented etc.

The DP Strategy is the link between the high level policy, and the preservation plans that have been developed at the State and University Library for specific collections. The collection specific preservation plans transform the policies on the Preservation Procedure Level, in case of the State and University Library the DP Strategy, into human readable control policies that, combined with the general statements from the DP Strategy, form the basis for developing machine readable Control Policies.

In SCAPE The State and University Library has performed an experiment with transforming DP Strategy on the Preservation Procedure Level and the collection specific preservation plans into machine readable Control Policies.

4.2 Policies at the Science and Technology Facilities Council (STFC)

STFC's high level, organizational wide Data Policy [15] states that underlying data should be kept for at least ten years after the end of a project or in perpetuity if it is unrepeatable observational data and that all data should have a Data Management Plan. This data management plan should address preservation as part of the data lifecycle, the focus within STFC is on data management rather than preservation due to the nature of STFC's business which is supporting the processes of creating new scientific data and ensuring this remains useable.

The ISIS Neutron Spallation Source, one of the large scale scientific facilities provided by STFC has a Data policy for users of the facility [10]. Although this is not exclusively concerned with preservation, it addresses some of the topics covered in preservation procedure policy and has been used as the starting point for the creation of control policies to support the Research Data Testbed scenarios provided by STFC elsewhere in the SCAPE project.

4.3 Applying the model to a real life situation

To enable to generation of control policy statements which can be used elsewhere in the SCAPE project a process of elaborating these statements needed to be identified. There are two key differences between policy aimed at a human audience and policy to be used automatically:

There is a difference in intent and viewpoint between written, human readable policies, especially at the higher levels and the control level policy. High level policy is trying to set the boundaries of acceptable states whereas control level policy is aiming to be precise in defining conditions for those states

The second difference is the implicit/explicit dilemma. A person will need less documented facts as they can use other implicit information, whereas a computer system only knows what it is told. Being able to ensure all implicit information is made explicit is a hard task to undertake.

4.3.1 Process for creation of control policies

There are two possible starting positions: (1) that the natural language control level policy is already documented and (2) that natural language preservation procedure level policy exists but natural language control level policy is implicit and is not contained in a single document describing detailed preservation decisions for the collection. For our experiments both of these states applied.

During the experiment we identified the following stage and steps. The three stages are (1) steps which apply to the whole policy

[5] cf RFC 2119 <http://www.ietf.org/rfc/rfc2119.txt

[6] http://purl.org/DP/quality/measures

[7] http://purl.org/DP/quality/measures#139

[8] http://purl.org/DP/quality/attributes

[9] http://purl.org/DP/quality/categories

document, (2) steps which need to be applied to each policy statement and (3) final review of the results.

Whole Policy Steps

1. Define the content set that the policy addresses

The content set is an intellectual cohesive collection of digital objects to which all the objectives within a preservation case apply.

The differences between the two organisations showed clearly a different approach in identifying the collections, for STFC the policy created a single content set related to the way the data were created and collected, and at SB the collection was a heterogeneous set of Radio Television Collection, as the policies were written on this level and reflect the organisation's view of their information. It should be noted that the STFC ISIS formats are specialised and consist of a local format for early data and a domain specific format for later data, and so for data management purposes there is no need to further divide the data; however for preservation purposes where we are interested in the semantics within the files, then there may be a need to describe collections in a different manner.

2. Identify the user community/ roles required by the policy

It is important to be able to identify who will be enacting the policy statement. Although the SB and STFC user communities identified had different names, they both were aligned to the DL.org [1, p.23] End Users which identifies three types: creators, consumers and administrators.

3. Map policy statements to high level concepts

To assist in identifying the risk or preservation case that the particular policy statement addresses, it is mapped to one (or more) of the high level concepts we already identified in SCAPE.

So the ISIS Data Management policy fragment **"3.1.1 *All raw data will be curated in well-defined formats for which the means of reading the data will be made available by the Facility*",** maps to the high level concepts of format and access and so the final preservation case will be concerned with these aspects.

Steps for each line of policy

1. Clarification to implicit meaning

This stage is designed to ensure that the natural language version being worked on does not have any "hidden" meaning within the words.

2. Identification of Control Policy Model Preservation Case

A Preservation Case ties Objectives to a Content Set (defined in step 1) and intended User Community (defined in step 2) This step should assist in identifying a particular Preservation Case for this particular policy statement.

3. Identification of Objectives for this content set

The Objectives are the measurable machine readable statements to be generated from the policy fragment being considered. These for example can be access objectives *rendering tools should exist for specific environments in use by the user community* or file format objectives *only ISO standard file formats should be in the collection*. The Objectives need to be phased in clear statements (MUST, SHOULD, >, < etc.)

4. Generate control statements

Tooling with a GUI will support the end user to create the machine readable control statements; in this case we use a set of already created attributes and measures.

Review the Preservation Cases

1. Review the preservation cases identified

Having completed the whole policy, then a check should be made as to whether any control policies and/or preservation cases overlap and whether it might be advisable to merge the outcomes or identify those which apply to the whole organisation.

5. CONCLUSIONS

Several conclusions can be drawn from this experiment. Firstly, it is possible to create machine readable control policies based on existing policy documents and using the Control Policy Model described in this document. The ease of doing so depended on the level of policy documents and the familiarity of the creator with the preservation intent and specific collection knowledge and in both cases the policy documents were too generic and detailed information needed to be gathered from other sources. This process also assumes that all relevant topics will be covered in the Preservation policies; there may be occasions where the control policy may come from another source – such as a specific requirement of the software used.

There are two main challenges still be to be worked on. The first is that the process moving from the often implicit to the explicit; is in practice a difficult task and the requirement to make control policies unambiguous may not be achievable for all policy elements. Secondly the granularity of the preservation case is still under discussion. The preservation case groups the objectives, content set and users together around the mitigation of a risk and will be used in the Watch and Planning tools. What is the appropriate level of granularity working from the policy, may not be the same as that required for Watch or for Planning for a Preservation Action. Both of these use triggers to action and the linkage between these and preservation cases are still under discussion. Currently we suggest that as it is not easy to identify the right level of granularity when defining control policies, we recommend to creating fine distinctions first and merging categories during the final stage.

This process leads from the natural language to machine readable policy, there is no process available to check that this machine readable policy is actually the same intent as the natural language policy, although ensuring specific linkages/relationships to be made between statements in the two levels would assist in this. Further development of a catalogue of policy elements related to the controlled vocabulary will contribute to solving these problems.

6. ACKNOWLEDGMENTS

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

7. REFERENCES

[1] Antunes, G, Barateiro, J, Becker, C et all: *Shaman Reference Architecture. Final version – update year 4* [2012. Retrieved 22-04-2013 from http://shaman-ip.eu/sites/default/files/SHAMAN-REFERENCE%20ARCHITECTURE-Final%20Version_0.pdf

[2] Bradner, S: *Key words for use in RFCs to Indicate Requirement Levels*. RFC 2119. 1997. Retrieved 22-04-2013 from : http://www.ietf.org/rfc/rfc2119.txt

[3] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., & Hofman, H. (2009). Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International journal on digital libraries*, *10*(4), 133-157.

[4] Candela, L and Nardi, A (ed.) *The Digital Library Reference Model* , 2011 p. 17. Retrieved 22/04/2013 from: from http://www.dlorg.eu/index.php/publications ver, Crafty Content

[5] Candela, L. and A.Nardi (ed.) *Digital Library Technology and Methodology Cookbook* p. 68 Retrieved 24-04-2013 from http://www.dlorg.eu/index.php/publications

[6] Dappert, A: *Report on the Conceptual Aspects of Preservation, Based on Policy and Strategy Models for Libraries, Archives and Data Centers*. Planets Project, 2009. Retrieved 22-04-2013 from http://www.planets-project.eu/docs/reports/Planets_PP2_D3_ReportOnPolicyAndStrategyModelsM36_Ext.pdf

[7] *Digital Preservation Policy for the State and University Library Denmark*. 2012 version 2.0 Retrieved 22-04-2013 from: http://en.statsbiblioteket.dk/about-the-library/ddpolicy

[8] *Digital Preservation Strategy for the State and University Library*, Denmark Version 2.0 June 2012 Retrieved 22-04-2013 from:http://en.statsbiblioteket.dk/about-the-library/dpstrategi

[9] Faria, L., P. Petrov, K. Duretec, C. Becker,M. Ferreira, and J. C. Ramalho. Design an architecture of a novel preservation watch system. In: *International Conference on Asia-Pacific Digital Libraries (ICADL)*. Springer, 2012

[10] *ISIS Data Management Policy*. Retrieved 22-04-2013 from http://www.isis.stfc.ac.uk/user-office/data-policy11204.html

[11] Kulovits, Hannes, Kraxner,Michael, Plangg, Markus, Becker, Christoph, Bechhofer, Sean. *Open Preservation Data: Controlled vocabularies and ontologies for preservation ecosystems* 10th International Conference on Preservation of Digital Objects (iPRES 2013), 2013

[12] *OWL 2 Web Ontology Language Document Overview* (Second Edition) 2012. Retrieved 22-04-2013 from: http://www.w3.org/TR/owl2-overview/

[13] *Resource Description Framework (RDF*) Retrieved 22-04-2013 from: http://www.w3.org/RDF/

[14] Sierman, B. and Wheatly, P: *Evaluation of Preservation Planning within OAIS, based on the Planets Functional Model. Planets Project* 2010 Retrieved 22-4-2013 from http://www.planets-project.eu/docs/reports/Planets_PP7-D6_EvaluationOfPPWithinOAIS.pdf

[15] *STFC scientific data policy* Retrieved 22-04-2013 from http://www.stfc.ac.uk/Resources/pdf/STFC_Scientific_Data_Policy.pdf