
# Final version of Policy specification model

Authors
Sean Bechhofer (University of Manchester), Barbara Sierman (National Library of the Netherlands), Catherine Jones (Science and Technologies Facilities Council), Gry Elstrøm (State and University Library Denmark), Hannes Kulovits, Christoph Becker (Vienna University of Technology)

July 2013

*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).*


# Executive Summary

The SCAPE Policy specification model is a controlled vocabulary for describing control policies. It provides a common language that can be used for the specification of policies which are then consumed and enacted upon by elements of the Preservation and Watch infrastructure. The vocabulary is specified using standard Semantic Web languages (RDF, OWL, SKOS) as defined by the W3C. This report provides an overview of model.


# Table of Contents

# 1 Introduction

There is a shared recognition of the importance of preservation policies for long term digital preservation. This is apparent in standards such as ISO standard 16363 Audit and Certification of Trustworthy Digital Repositories. 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] Such policies then support activities of an organisation with respect to the maintenance and preservation of collections. Ideally, these policies guide the preservation activities in the organisation.

This deliverable describes a machine-readable policy model that has been developed for the representation of control policies – low level statements about the states of affairs of a preservation system that can be tested. These control policies can then be used by components in a preservation ecosystem (for example a planning or watch tool) with suitable actions being taken dependent on the outcome of such tests. The provision of a machine readable representation for the policies helps to reduce issues of ambiguity and can increase opportunities for interoperation between those components.

The remainder of this document is structured as follows. Section 2 describes the context of this work within the SCAPE project. Section 3 discusses the background technologies used for representation. Section 4 presents the details of the model. Section 5 discusses policy creation, Section 6 describes concrete details of delivery and access, Section 7 presents some examples of control policy elements expressed in the model and Section 8 discusses prototype tools to assist users in writing control policies using the model. Note that this document contains content sourced from recent publications [opd, ppl].

Note that the deliverable does not give all the details of the models, but rather provides an overview of the content and structure. Those wishing to see the details are directed to the locations discussed in Section 6.1.

# 2 Preservation Policies

Most usages of *policies* correspond to what the Object Management Group (OMG) standards call b*usiness policies*. According to these standards, policies are "element[s] of governance" that are "not directly enforceable" and they "exist to govern; that is, control, guide, and shape the strategies and tactics" [sbvr, bmm]. Preservation policies hence should provide the mechanisms to document and communicate key aspects of relevance, in particular drivers and constraints and the goals and objectives motivated by them. At present, there are no established standards for preservation policies relevant to planning or for aspects such as monitoring specifications, Service Level Agreements for preservation operations, or system interfaces.

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

Preservation Policies are not a goal in themselves, but are to support the activities of an organisation with respect to the maintenance and preservation of a digital collection. “Without a policy framework a digital library is little more than a container for content”[2] . In an ideal situation, the preservation policies will guide the preservation activities in an organisation. The SCAPE project has designed a preservation policy model that will support organisations to build their preservation policy documents.
Preservation and Watch activities within SCAPE have identified three levels of policy – guidance policies, procedure policies and control policies.

1. **High level or guidance polices.** At this level an organisation describes the general long term preservation goals of the organisation for its digital collection(s). For example, an organisation may desire that “all content collected will be actively preserved” or “the organisation aims to preserve the digital collections in data formats suitable for digital preservation”.
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 holdings in general. For example a policy may require the organisation to consider the *openness* of a format used: Is the format well described and is documentation available? Is the format subject to any patents? Is a licence or permission required to use the format?
3. **Control policies**. At 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 available for use in automated planning and watch tools to ensure that preservation actions and workflows chosen meet the specific requirements identified for that digital collection. These may be kept internally within the organisation, although sharing control policies could help other organisations in authoring or developing policies.

The first two of these – guidance and procedure – are high level policies that are expressed in natural language forms. An investigation of relevant literature such as OAIS, TRAC etc. led to a set of areas that were seen as representing the guidance policies. Examples are Preservation Strategies, Authenticity, Standards, Access and Rights Management.

Control policies are expressions of states of affairs or desired states of affairs. Ideally these policies can be tested and appropriate action taken depending on the result. For example a control policy may express characteristics of the formats used to represent content in a content set held by an organisation. By expressing these low level policies in a structured, machine readable way, such as through the SCAPE Content Policy Model, we allow the possibility of actioning, testing or validating those policies through some computational process.

Organisations will have other policies which are not directly labelled "preservation" but may have an impact on how preservation is undertaken - such as IT back-up and recovery plans. In addition, these preservation procedure policies are likely to be supported by internal procedure documents, job role descriptions and on occasion specific project plans. When compiling the control level statements, it is not yet clear how much dependency there will be on these other supporting documents.

[2] DL.org 3.4 **Digital Library Technology and Methodology Cookbook p. 68**
http://www.dlorg.eu/index.php/outcomes/dl-org-cookbook

The expectation here is that information regarding different aspects of SCAPE – collections of content, content profiling, formats, the communities for whom preservation is being undertaken, etc. – is represented in a common way using a common information structure. The advantages of such an approach include the ability to query across organisational information, policies, watch requests, plans, actions etc. using a single framework.

By expressing policies using standardised, common vocabularies we facilitate the sharing of information between components of a preservation ecosystem and the potential reuse of policy elements between organisations. The use of open APIs and a common language then support loose coupling between the components in that ecosystem.

The requirements on our control policy model are:

- to ensure the coverage of key domain entities;
- to develop a representation using open standards
- to follow design principles of modularity, openness and extensibility

While policies are in general "non-enforceable", the control policies discussed here are concerned with concrete aspects of collections, formats or environments that can be checked/verified by machine (for example through an examination of the values for particular properties).

The control policies described here may be concerned with the representations of objects; the formats that are used by those representations; the tools that are used to access or view content, the platforms that tools run on; content profiles of content sets; or technologies available to users.

The model described here defines a vocabulary – a collection of terms and relations that can be used to define and describe control policies. A set of control *policy instances* will be described using a *policy vocabulary*. These descriptions will also make use of terms from a *domain vocabulary* to describe particular *domain entities*: situations, formats, content sets etc.

Figure 1 below illustrates these dependencies.

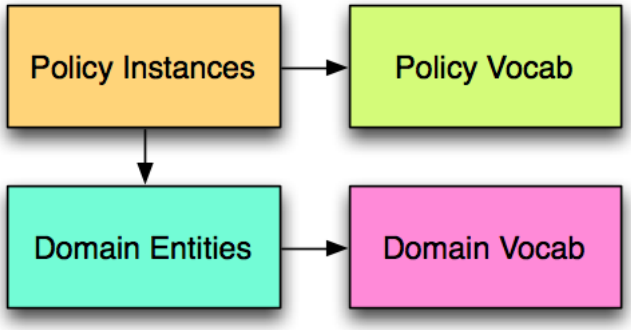



Figure 1 Dependencies between Policy Model Components

## 2.1 Control Policies and SCAPE components

The controlled vocabulary for policies is used to support interoperability between components in the SCAPE ecosystem, including the planning tool Plato, watch system Scout and c3po, a tool for processing content profile information.

An organisation specifies control policies for a specific preservation case (see below for discussion of how this is represented in the model). Scout can then detect violations in those policies (potentially using content profiling information created by c3po), triggering the creation of a preservation plan by the planning tool Plato. During this planning process, Plato can make use of the objectives and constraints specified in the control policies. See Section 5 for further discussion of this.

# 3 Background Technologies

According to the W3C, "The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries." This vision has given rise to the development of a family of representation languages that are specifically targeted at representing data and metadata in a web context. More specifically, the languages RDF, RDFS, OWL and SKOS provide frameworks for representation. These languages are open standards, fitting the requirement introduced above. In addition, RDF provides flexibility and extensibility, as vocabularies and schema can be extended with additional domain vocabulary. As discussed in [odp], this allows the provision of a core ontology that can be extended when necessary to support an evolving ecosystem.

## 3.1 RDF

The Resource Description Framework (RDF) [rdf] is a generic language for the representation of data and metadata that essentially allows for the expression of subject-predicate-object triples, describing the characteristics of resources (a resource being more or less anything that can be identified via an IRI[3]). Triples can be collected together to form RDF graphs, allowing for rich statements about resources.

## 3.2 Ontologies and Vocabularies

Such assertions or graphs have little value without common agreement on their interpretation, and this is provided, in part, by vocabularies and ontologies. An ontology provides a collection of shared terms, along with explicit characterisations of the assumptions that should be made when interpreting those terms. This achieves two things:

1. The use of common terms, so that applications are “speaking the same language”
2. A guarantee that interpretation of those terms is made consistently, so that applications “mean the same thing”.

### 3.2.1 RDFS and OWL

The schema and ontology languages RDFS [rdfs] and OWL [owl] provide mechanisms for the specification of vocabularies and ontologies. They allow for description of classes (collections of things) and properties (relationships between things) along with characterisations of those classes and properties. RDFS is a relatively simple language, providing the opportunity to define class hierarchies. OWL provides more expressive operators, allowing for the definition of necessary and sufficient conditions for class membership, restrictions on properties, Boolean operators and so on. OWL has a well-defined semantics – describing precisely how one should interpret a complex

[3] Internationalised Resource Identifier, a generalization of the Uniform Resource Identifier (URI), most commonly encountered as the Uniform Resource Locator (URL).

expression – that facilitates the use of reasoning, supporting ontology management and querying of data.

### 3.2.2 SKOS

SKOS [skos] is further language that allows for the representation of knowledge organisation systems (KOS). It serves a different purpose to OWL in that it does not seek to provide rich intentional definitions of entities, but is rather targeted at those KOS such as thesauri and controlled vocabularies that are used primarily for indexing and retrieval.

### 3.2.3 SPARQL

RDFS, OWL and SKOS all build on RDF as a representation. Thus data, metadata and schema definitions can all be represented (and queried) in one common format. SPARQL provides a query language and protocol that allows for query of RDF graphs.

# 4 SCAPE Policy Model

As discussed above, the SCAPE preservation policy model consists of three preservation policy levels that will support an organisation to create their preservation policies set, those levels being high level/guidance, preservation procedure and control. We focus here primarily on the *control level* policies.

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 an 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. An unambiguous reworking of this statement could then assert that 80% of formats used must be developed by ISO according to information from a format registry such as PRONOM. A content profiling tool (such as c3po) can analyse document collections and provide information about the formats used in that collection. Format registries (e.g. PRONOM) 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.

## 4.1 Model Content

Requirements for the Control Policy Model were largely taken from an internal document produced by SCAPE WP12 and WP14, in particular [framework].

The policy model provides vocabulary that is used to describe particular domain entities: situations, formats, content sets etc. Key entities described in the model are as follows:

- **Content Set.** A Content Set represents a collection of objects that are the focus of the policy

- **User Community.** The community for whom digital content is preserved for.
- **Preservation Case.** A Preservation Case ties objectives to a Content Set and intended User Community
- **Objective.** Objectives are the atomic building blocks of the policies. Objectives may refer to properties that representations of content have; properties of the formats themselves; tools used and so on. Objectives are defined in terms of measures (see below) which are taken from a catalogue.

To enable successful communication between decision makers and automated operations, we have developed a core model of specific policy elements that can be represented in a machine-understandable way.

We define control policies as practicable elements of governance that relate to clearly identified entities in a specified domain model. An element of governance is practicable if it is "sufficiently detailed and precise that a person who knows the element of guidance can apply it effectively and consistently in relevant circumstances to know what behaviour is acceptable or not, or how something is understood"[rules].

A control policy contains quantified, precise statements of facts, constraints, objectives and directives about domain entities such as representations of content or format and their properties. Decision making processes such as preservation planning translates these policies into a specified set of rules in a plan. This rule set is then testable and enforceable, and it controls operations. For example, constraints about data formats to be produced by conversion processes can be automatically enforced in a straightforward way.

The policies thus provide an input to the planning and watch processes. Note that the model is simply there to allow us to state the objectives 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).

For expressing control policies, we introduce a policy vocabulary that is used to describe concrete control policy instances. These policies use vocabulary from a domain vocabulary to describe particular domain entities such as formats, and content. Figure 1 (see p.3) illustrates these interactions, while Figure 2 illustrates the overall policy model including classes and properties discussed. Note that this figure is not intended to be a complete and exhaustive depiction of the content of the vocabularies – for this we refer the reader to the resources described in Section 6.1).

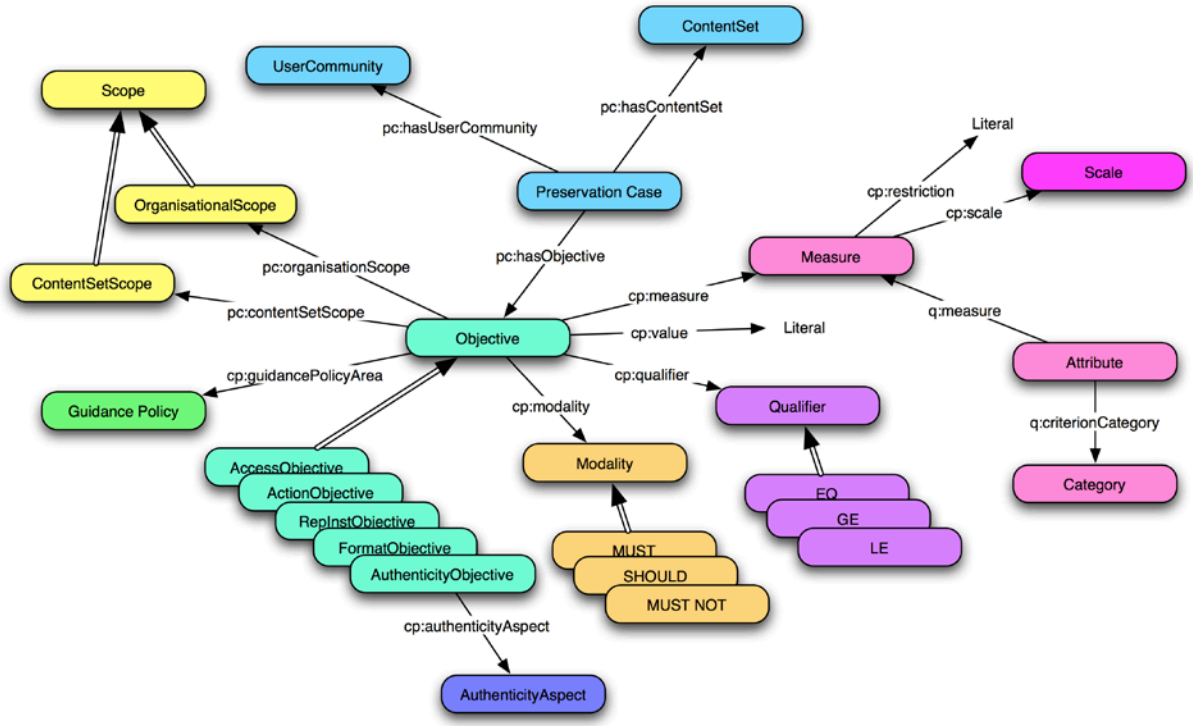



Figure 2 Overview of Policy Model

Central to a control policy statement is the notion of a *preservation case*, which links a *content set* to a *user community* with particular *objectives*. Before decision makers embark on a preservation endeavour, the context of "what" has to be achieved for "whom" needs to be established. As Webb et al. describe in [preserve], an identified set of objects is being preserved for a certain user community, such as images preserved in a library for the general public, or business processes in a company for internal usage to ensure legal compliance.

Ultimately, ensuring that the objectives associated with a case are met is the target of preservation planning. To achieve this, objectives need to be associated with measurable outcomes. To this end, we define a *measure* as the result of measurement of an *attribute*. Objectives are thus based on attributes that are represented by measures. Following the definition in ISO/IEC 15939:2002, an attribute is an "inherent property or characteristic of an entity that can be distinguished quantitatively or qualitatively by human or automated means" [measurement, quality].

An example is the attribute *compression* which indicates the compression used. Measures for this attribute include the *compression type* (none, lossless, or lossy), *compression algorithm*, and *compression algorithm covered by patent* which indicates whether licencing fees might occur when using a certain compression algorithm.

### 4.1.1 Measures

In the vocabulary we define a measure as `m(s,r)` with

`s` Scale used for conducting measurement. This includes Boolean, number, and ordinal.

`r` Restriction limiting the possible range of measurement values. Specification of a restriction is optional.

A core set of measures have been defined (see Section 5). Organisations may define their own catalogues of measures, although the use of the bespoke vocabularies may restrict opportunities for sharing. The association of measures with attributes (and categories) allows for query across all objectives relating to a specific attribute (see Section 7).

### 4.1.2 Objectives

We further define a control policy objective as `cp(m,v,q,mo)` with

`m` A measure pertaining to an authenticity, access, action, or representation instance objective. Measures are taken from a measures catalogue (See Section 6).

`v` A value associated with the measure.

`q` A qualifier (equals, less than, greater than, less or equal, greater or equal).

`mo` A modality that describes whether the particular property- value pair is present or not (MUST, SHOULD, MUST NOT[4]).

An Objective refers to a particular property (measure) 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. 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 fragment from a preservation case taken from the Danish Statsbiblioteket is shown below (in RDF Turtle format).

```
:global_institutional_policies
    pw:hasContentSet :radio_television_collection ;
    pw:hasObjective :AutomaticFormatValidationMustBePossible,
       :FormatDocumentationShouldBeFree,
       :FormatIdentificaitonMustBePossible,
       :FormatMustHaveNoLicenseCosts,
       :FormatShouldBeValid,
       :NumberOfFreeOpenSourceToolsMustBeGT3,
       :UseOfFormatShouldBeWidespread ;
    pw:hasUserCommunity :library ;
    a pw:PreservationCase;
    skos:prefLabel "SB Global Institutional" .


:AutomaticFormatValidationMustBePossible
    control-policy:modality modalities:MUST ;
    control-policy:measure <http://purl.org/DP/quality/measures#164> ;
    control-policy:value "automatic"^^xsd:string ;
   a control-policy:FormatObjective;

:radio_television_access
    pw:hasContentSet :radio_television_access_copies ;
    pw:hasObjective :FormatMustBeSupportedByBrowsers ;
    pw:hasUserCommunity :researchers ;
    a pw:PreservationCase;
    skos:prefLabel "SB Radio and Television Access" .

:FormatMustBeSupportedByBrowsers
    control-policy:measure <http://purl.org/DP/quality/measures#156> ;
    control-policy:value true ;
```

[4] As used in RFC 2119, http://www.ietf.org/rfc/rfc2119.txt

```
pw:contentSetScope :radio_television_access_copies ;
a control-policy:RepresentationInstanceObjective.
```

This example includes two objectives. The first is a format objective concerning automatic validation of formats. The second is a representation instance objective requiring that a format be supported by browsers. See below for discussion of objective types.

### 4.1.3 Preservation Case

The preservation case collects the particular objectives resulting from the combination of user community and content set intended for preservation. This includes the time horizon and the goals, objectives and constraints associated with a case.

### 4.1.4 Objective and Constraint

To be able to preserve the content for a specific user community, clear objectives and constraints on several aspects have to be defined. Objectives have been broken down as follows:

- **Format Objective.** This describes an objective referencing a particular property that formats in general should or must have. Most importantly, this corresponds to a risk profile of formats.
- **Authenticity Objective.** This denotes an objective describing the requirements for the preservation of a certain significant property in a preservation case. The set of significant properties can then be used to determine whether a particular preservation action will preserve the authenticity of the performance of each digital object.
- **Representation Instance Objective**. These describe objectives referencing a property that representations of content, such as files and byte streams, should or must have. This includes aspects such as compression, encryption, size, or validity.
- **Access Objective.** This is an objective that describes the requirement for the preservation of a certain characteristic in a particular scenario with respect to accessing the digital object.
- **Action Objective**. This describes constraints on the preservation action process, such as the maximum time or memory resources available or a restriction on allowed licensing.

We observe that the above collection of objectives may not be exhaustive – different kinds of data, for example science data, may require the definition of further objective types.

## 4.2 Validation

It is important to note that this model provides a framework for the expression of control policies – it provides (consistent and shared) vocabulary for those policies. Issues concerning, for example modalities (whether characteristics must, should, may take particular values) or comparisons between values are not explicitly handled in the semantics of the model. It is simply a container for information that is then passed on to other components (for example planning, monitoring or watch). Those applications will then take appropriate actions.

Note also that here we are considering atomic objectives. Kolovski et. al. [service] describe an approach to representing *combinations* of policy elements using OWL-DL, allowing inference of relationships such as policy inclusion. Such an approach may potentially also be used here, but has not been explored as yet.

# 5 Process for Control Policy Creation

There are three stages involved in translating natural language policy into a machine understandable representation. The first stage has steps which apply to the whole policy document. The content set that the policy applies to must be identified and consideration made as to whether this content set needs any further changes for all the machine readable policy to apply to it. The user communities/roles that the policy applies to must be identified - the minimum set is likely to be creators, users and staff responsible for managing the content. Finally to assist in identifying the appropriate preservation case, the sections of the policy should be mapped to high level topics, this may help in deciding which type of objective is to be identified for a specific piece of policy. High level topics that have been identified within SCAPE are

- Authenticity, measures to establish authenticity
- Preservation Goals an organizations want to achieve
- Preservation Strategies methods or strategies for preservation of achieving the goals
- Digital Object, policies related the object being preserved
- Metadata, policies related to metadata
- Organisation, policies related to the behaviour and tasks of the archival organization
- Standards, the applicability of standards
- Designated Community, policies related to the users of the digital archive
- Storage, policies related to the storage of digital objects
- Formats, policies related to file formats
- Rights, policies related to access, preservation, IP etc. rights
- Trustworthy Digital Repositories, policies related to the aim to become a TDR

The second stage is repeated for each policy statement. Each statement needs to any implicit information made explicit, the preservation case linking the content set, user community and specific objectives should be decided upon; the precise objectives then identified and finally the control statements generated using the attributes and measures described elsewhere.

The final stage is to review the outputs of the second stage looking for overlaps and duplication; there may be objectives that apply across all preservation cases.

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. 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.

# 6 Implementation

## 6.1 Delivery

The policy model described above has been implemented as a collection of OWL ontologies and SKOS vocabularies.

These are delivered via URLs defined within a general namespace of http://purl.org/DP (for Digital Preservation), with the PURL currently redirecting to content served from http://www.ifs.tuwien.ac.at/dp/. A number of individual vocabularies/namespaces are provided:

- http://purl.org/DP/control-policy
  Vocabulary describing Objectives and the general properties of those objectives.
- http://purl.org/DP/control-policy/modalities
  SKOS Vocabulary providing modalities (cf RFC-2119);
- http://purl.org/DP/control-policy/qualifiers
  SKOS Vocabulary providing qualifiers;
- http://purl.org/DP/preservation-case
  Vocabulary describing preservation cases, organisations and user communities;
- http://purl.org/DP/quality
  Basic classes describing categories, attributes and measures;
- http://purl.org/DP/quality/categories
  Categories. Generated from a spreadsheet (see below);
- http://purl.org/DP/quality/attributes
  Attributes. Generated from a spreadsheet (see below);
- http://purl.org/DP/quality/measures
  Measures. Generated from a spreadsheet (see below);
- http://purl.org/DP/quality/scales
  SKOS Vocabulary describing scales;
- http://purl.org/DP/quality/scopes
  SKOS Vocabulary describing scope;
- http://purl.org/DP/quality/authenticity
  SKOS Vocabulary describing aspects of authenticity.
- http://purl.org/DP/guidance
  SKOS Vocabulary describing high level guidance policy areas.

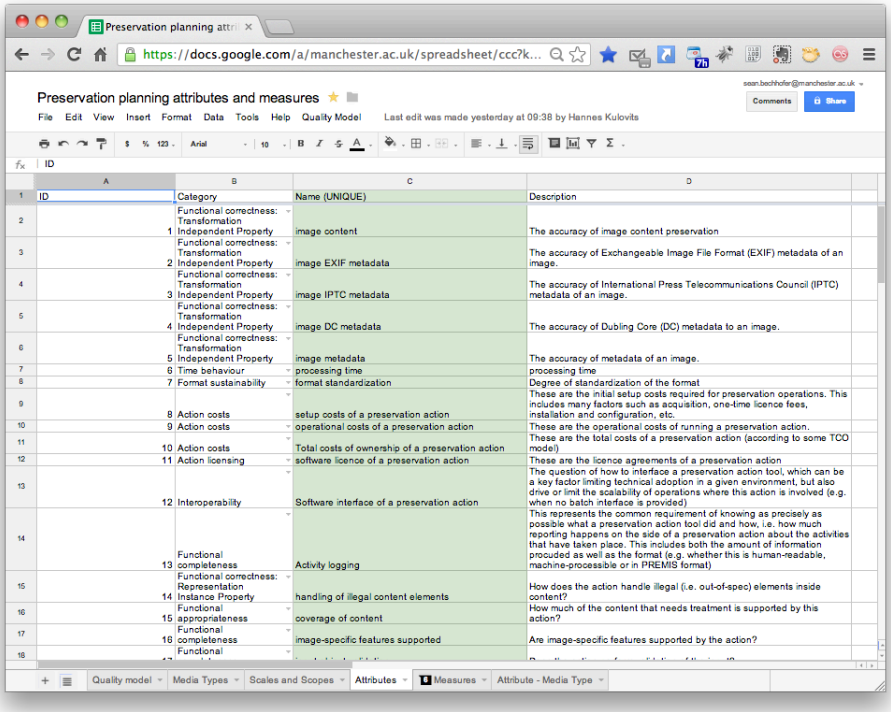
All vocabularies are delivered using content-negotiation, providing RDF representations or human-readable HTML documentation depending on `Accept:` headers in the HTTP request. Human readable documentation is generated from the source ontologies.

The separation of the model into a number of connected, modular pieces provides flexibility and allows for extensibility. For example, the basic control-policy vocabulary allows for the description of Objectives as discussed above, but does not refer to specific measures. These are specified in the measures vocabulary. This gives the possibility of using the top level "schema" with an alternative collection of measures.

Where possible, external existing vocabularies and URIs have been incorporated, rather than redefining concepts. For example, URIs for media types (http://purl.org/NET/mediatypes). These are then used to annotate attributes. For example, measure http://purl.org/DP/quality/attributes#1 (image content) relates to media type http://purl.org/NET/mediatypes/image. The Library of Congress provides SKOS vocabularies for the representation of countries (http://id.loc.gov/vocabulary/countries) and languages (http://id.loc.gov/vocabulary/iso639-2). Terms from these vocabularies are used to annotate user communities and organizations.

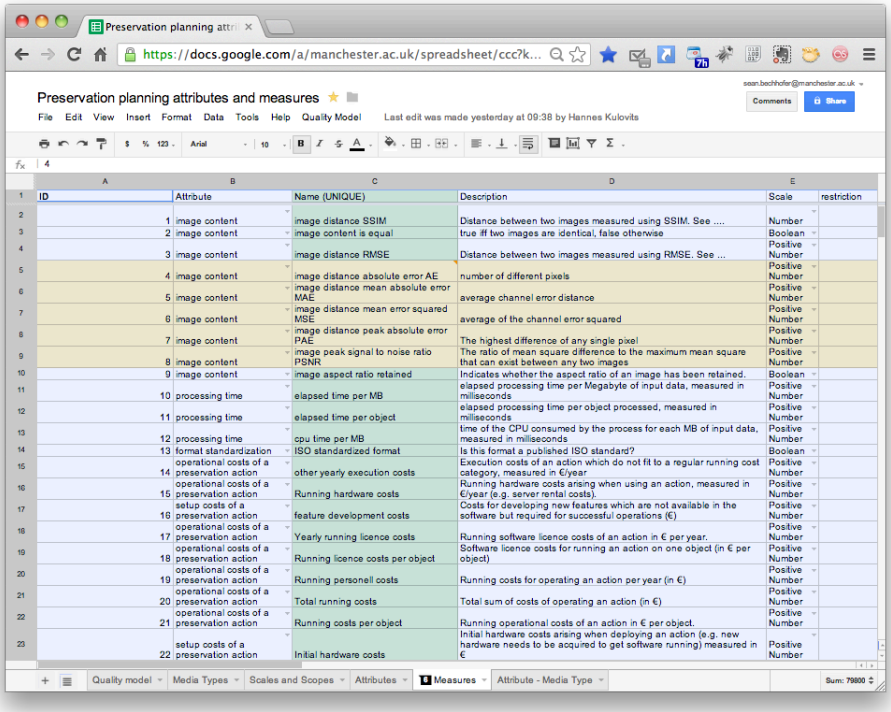
## 6.2 Authoring

The vocabularies are represented using RDF, OWL and SKOS. The core vocabularies that describe the schema level classes and properties have been authored using standard OWL tools (Protégé). However, the authoring of the vocabularies describing categories, attributes and measures is done via a spreadsheet (see Figure 3 and Figure 4). These spreadsheets allow the description of the categories, attributes and measures (names, description) along with additional information such as scales or expected values.




| | ID | Category | Name (UNIQUE) | Description |
|---|---|---|---|---|
| 2 | 1 | Functional correctness: Transformation Independent Property | image content | The accuracy of image content preservation |
| 3 | 2 | Functional correctness: Transformation Independent Property | image EXIF metadata | The accuracy of Exchangeable Image File Format (EXIF) metadata of an image. |
| 4 | 3 | Functional correctness: Transformation Independent Property | image IPTC metadata | The accuracy of International Press Telecommunications Council (IPTC) metadata of an image. |
| 5 | 4 | Functional correctness: Transformation Independent Property | image DC metadata | The accuracy of Dubling Core (DC) metadata to an image. |
| 6 | 5 | Functional correctness: Transformation Independent Property | image metadata | The accuracy of metadata of an image. |
| 7 | 6 | Time behaviour | processing time | processing time |
| 8 | 7 | Format sustainability | format standardization | Degree of standardization of the format |
| 9 | 8 | Action costs | setup costs of a preservation action | These are the initial setup costs required for preservation operations. This includes many factors such as acquisition, one-time licence fees, installation and configuration, etc. |
| 10 | 9 | Action costs | operational costs of a preservation action | These are the operational costs of running a preservation action. |
| 11 | 10 | Action costs | Total costs of ownership of a preservation action | These are the total costs of a preservation action (according to some TCO model) |
| 12 | 11 | Action licensing | software licence of a preservation action | These are the licence agreements of a preservation action |
| 13 | 12 | Interoperability | Software interface of a preservation action | The question of how to interface a preservation action tool, which can be a key factor limiting technical adoption in a given environment, but also drive or limit the scalability of operations where this action is involved (e.g. when no batch interface is provided) |
| 14 | 13 | Functional completeness | Activity logging | This represents the common requirement of knowing as precisely as possible what a preservation action tool did and how, i.e. how much reporting happens on the side of a preservation action about the activities that have taken place. This includes both the amount of information procuded as well as the format (e.g. whether this is human-readable, machine-processible or in PREMIS format) |
| 15 | 14 | Functional correctness: Representation Instance Property | handling of illegal content elements | How does the action handle illegal (i.e. out-of-spec) elements inside content? |
| 16 | 15 | Functional appropriateness | coverage of content | How much of the content that needs treatment is supported by this action? |
| 17 | 16 | Functional completeness | image-specific features supported | Are image-specific features supported by the action? |
| 18 | | Functional | | |

Figure 3 Attributes Spreadsheet



Preservation planning attributes and measures

| ID | Attribute | Name (UNIQUE) | Description | Scale | restriction |
|---|---|---|---|---|---|
| 1 | image content | image distance SSIM | Distance between two images measured using SSIM. See .... | Number | |
| 2 | image content | image content is equal | true iff two images are identical, false otherwise | Boolean | |
| 3 | image content | image distance RMSE | Distance between two images measured using RMSE. See ... | Positive Number | |
| 4 | image content | image distance absolute error AE | number of different pixels | Positive Number | |
| 5 | image content | image distance mean absolute error MAE | average channel error distance | Positive Number | |
| 6 | image content | image distance mean error squared MSE | average of the channel error squared | Positive Number | |
| 7 | image content | image distance peak absolute error PAE | The highest difference of any single pixel | Positive Number | |
| 8 | image content | image peak signal to noise ratio PSNR | The ratio of mean square difference to the maximum mean square that can exist between any two images | Positive Number | |
| 9 | image content | image aspect ratio retained | Indicates whether the aspect ratio of an image has been retained. | Boolean | |
| 10 | processing time | elapsed time per MB | elapsed processing time per Megabyte of input data, measured in milliseconds | Positive Number | |
| 11 | processing time | elapsed time per object | elapsed processing time per object processed, measured in milliseconds | Positive Number | |
| 12 | processing time | cpu time per MB | time of the CPU consumed by the process for each MB of input data, measured in milliseconds | Positive Number | |
| 13 | format standardization | ISO standardized format | Is this format a published ISO standard? | Boolean | |
| 14 | operational costs of a preservation action | other yearly execution costs | Execution costs of an action which do not fit to a regular running cost category, measured in €/year | Positive Number | |
| 15 | operational costs of a preservation action | Running hardware costs | Running hardware costs arising when using an action, measured in €/year (e.g. server rental costs). | Positive Number | |
| 16 | setup costs of a preservation action | feature development costs | Costs for developing new features which are not available in the software but required for successful operations (€) | Positive Number | |
| 17 | operational costs of a preservation action | Yearly running licence costs | Running software licence costs of an action in € per year. | Positive Number | |
| 18 | operational costs of a preservation action | Running licence costs per object | Software licence costs for running an action on one object (in € per object) | Positive Number | |
| 19 | operational costs of a preservation action | Running personell costs | Running costs for operating an action per year (in €) | Positive Number | |
| 20 | operational costs of a preservation action | Total running costs | Total sum of costs of operating an action (in €) | Positive Number | |
| 21 | operational costs of a preservation action | Running costs per object | Running operational costs of an action in € per object. | Positive Number | |
| 22 | setup costs of a preservation action | Initial hardware costs | Initial hardware costs arising when deploying an action (e.g. new hardware needs to be acquired to get software running) measured in € | Positive Number | |

Figure 4 Measures spreadsheet

A tool (available from the OpenPlanets code repository[5]) takes this spreadsheet and produces the RDF vocabularies describing categories, attributes and measures.

We expect that the ontologies and vocabularies used within planning and watch processes will evolve over time – in particular it is likely that organisations will wish to extend the measures catalogue with additional measures to cover additional scenarios. The use of the spreadsheet for authoring of measures supports this, allowing non-Semantic Web Technology experts to author properties to be used in control policies. A certain level of control is needed though, in order to ensure that, for example, the central measures catalogue is not extended unnecessarily with duplicate elements. The vocabularies are also currently being managed via the OPF github repository. Processes are being put in place that will support requests for extensions and changes via issue tracking – it is important that the vocabularies are seen as open, with encouragement for community contributions and extensions

[5] https://github.com/openplanets/policies

## 6.3 Control Policy Authoring and Usage

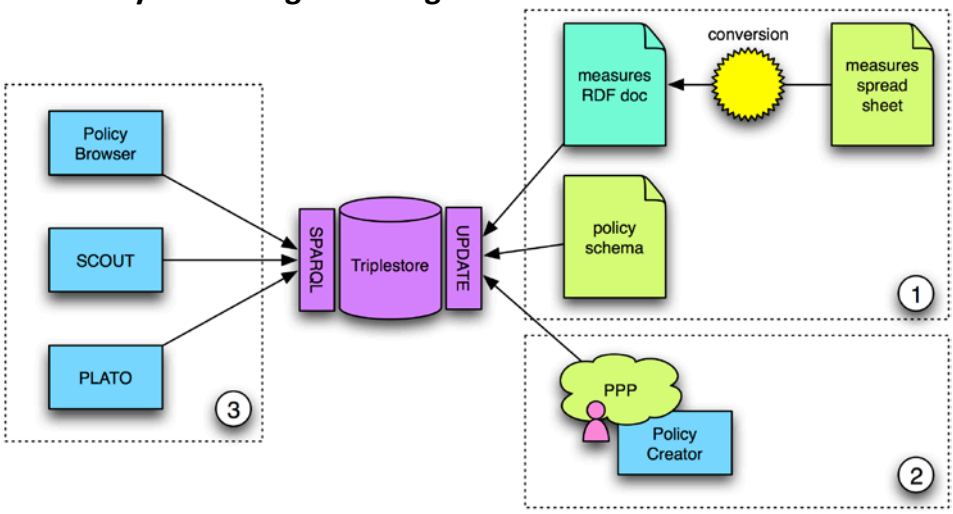



Figure 5 Authoring and Usage

Figure 5 illustrates the authoring process along with other uses of the model. In (1), the measures sheet and policy schema are edited. These are then updated in a central triple store. In (2), the Policy creation tool (see Section 7) can then access the vocabularies via SPARQL, and can add information via UPDATE. Finally, in (3), the created policy elements can be queried by tools such as SCOUT and PLATO via SPARQL.

In the current implementations, control policies represented in RDF form are directly uploaded to Plato. In the first stage of planning, Plato queries the preservation cases (see the queries discussed below) and allows selection of a case. The preservation plan will then be based on this preservation case (i.e. a preservation plan is created for a specific content set which shall be preserved for a specific user community). Linked to the preservation case are the control policies on which Plato bases the creation of the objective tree. This process is also discussed in [odp].

# 7 Example

In order to validate the model, example policies from SCAPE partners SB, TUW, BL and STFC have been produced.

## 7.1 Austrian State Archives

We include here an example relating to policies for the Austrian State Archives that have been produced by TUW. The example is presented using Turtle, a text-based RDF format.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix : <http://www.oesta.gv.at/policies#> .
@prefix quality: <http://purl.org/DP/quality#> .
@prefix measures: <http://purl.org/DP/quality/measures#> .
@prefix preservation-case: <http://purl.org/DP/preservation-case#> .
@prefix control-policy: <http://purl.org/DP/control-policy#> .
@prefix modalities: <http://purl.org/DP/control-policy/modalities#> .
@prefix qualifiers: <http://purl.org/DP/control-policy/qualifiers#> .
@prefix authenticity: <http://purl.org/DP/authenticity#> .
@prefix org: <http://www.w3.org/ns/org#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
```

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix pronom: <http://reference.data.gov.uk/technical-registry/> .
@prefix premis:
<http://multimedialab.elis.ugent.be/users/samcoppe/ontologies/Premis/premis
.owl> .

<http://www.oesta.gv.at/policies/austrian_state_archives>
    a org:Organization ;
    org:identifier "ASA" .

<http://www.example.com/ImageSpecialInterestGroup>
    a foaf:Group ;
    foaf:name "Image Special Interest Group" .

<http://www.oesta.gv.at/policies/public>
    a foaf:Group .

<http://www.oesta.gv.at/policies/researchers>
    a foaf:Group .

<http://www.oesta.gv.at/policies/ministries>
    a foaf:Group .

<http://www.oesta.gv.at/policies/photographs>
    a preservation-case:ContentSet .

<http://www.oesta.gv.at/policies/scanned_papers>
    a preservation-case:ContentSet .

<http://www.oesta.gv.at/policies/documents_executions>
    a preservation-case:ContentSet .

<http://www.oesta.gv.at/policies/documents_notes>
    a preservation-case:ContentSet .

<http://www.oesta.gv.at/policies/emails>
    a preservation-case:ContentSet .

<http://www.oesta.gv.at/policies/ministries_administrative_records>
    a preservation-case:ContentSet .

<http://www.oesta.gv.at/policies/ministries_executions_scenario>
    preservation-case:hasObjective
<http://www.oesta.gv.at/policies/FormatDocumentationMustBeFreelyAvailable>,
<http://www.oesta.gv.at/policies/FormatIdentificaitonMustBePossible>,
<http://www.oesta.gv.at/policies/FormatShallBeISOStandardized>,
<http://www.oesta.gv.at/policies/NumberOfFreeOpenSourceToolsMustBeGT3>,
<http://www.oesta.gv.at/policies/UseOfFormatShouldBeWidespread> ;
    preservation-case:hasUserCommunity
<http://www.oesta.gv.at/policies/ministries>,
<http://www.oesta.gv.at/policies/public>,
<http://www.oesta.gv.at/policies/researchers> ;
    a preservation-case:PreservationCase .
```

```
<http://www.oesta.gv.at/policies/ministries_scanned_papers_scenario>
    preservation-case:hasObjective
<http://www.oesta.gv.at/policies/ImageHeightMustBeUnchanged>,
<http://www.oesta.gv.at/policies/ImageMustBeIdentical>,
<http://www.oesta.gv.at/policies/ImageWidthMustBeUnchanged> ;
    preservation-case:hasUserCommunity
<http://www.oesta.gv.at/policies/ministries>,
<http://www.oesta.gv.at/policies/public>,
<http://www.oesta.gv.at/policies/researchers> ;
    a preservation-case:PreservationCase .

<http://www.oesta.gv.at/policies/FormatDocumentationMustBeFreelyAvailable>
    control-policy:measure <http://purl.org/DP/quality/measures#147> ;
    control-policy:modality modalities:MUST ;
    control-policy:value "yes-free"^^xsd:string ;
    a control-policy:FormatObjective .

<http://www.oesta.gv.at/policies/FormatIdentificaitonMustBePossible>
    control-policy:measure <http://purl.org/DP/quality/measures#153> ;
    control-policy:modality modalities:MUST ;
    control-policy:value "automatic_specific"^^xsd:string ;
    a control-policy:FormatObjective .

<http://www.oesta.gv.at/policies/FormatMustHaveLosslessCompression>
    control-policy:measure <http://purl.org/DP/quality/measures#117> ;
    control-policy:modality modalities:MUST ;
    control-policy:value "lossless" ;
    a control-policy:FormatObjective .

<http://www.oesta.gv.at/policies/FormatShallBeISOStandardized>
    control-policy:measure <http://purl.org/DP/quality/measures#13> ;
    control-policy:modality modalities:SHOULD ;
    control-policy:value "true"^^xsd:string ;
    a control-policy:FormatObjective .

<http://www.oesta.gv.at/policies/ImageHeightMustBeUnchanged>
    control-policy:authenticity authenticity:Content ;
    control-policy:measure <http://purl.org/DP/quality/measures#53> ;
    control-policy:value true ;
    a control-policy:AuthenticityObjective .

<http://www.oesta.gv.at/policies/ImageMustBeIdentical>
    control-policy:authenticity authenticity:Content ;
    control-policy:measure <http://purl.org/DP/quality/measures#2> ;
    control-policy:value true ;
    a control-policy:AuthenticityObjective .

<http://www.oesta.gv.at/policies/ImageWidthMustBeUnchanged>
    control-policy:authenticity authenticity:Content ;
    control-policy:measure <http://purl.org/DP/quality/measures#51> ;
    control-policy:validInScope
<http://www.oesta.gv.at/policies/austrian_state_archives> ;
    control-policy:value true ;
    a control-policy:AuthenticityObjective .

<http://www.oesta.gv.at/policies/NumberOfFreeOpenSourceToolsMustBeGT3>
    control-policy:measure <http://purl.org/DP/quality/measures#139> ;
```

```
    control-policy:modality modalities:MUST ;
    control-policy:qualifier qualifiers:GT ;
    control-policy:value 3 ;
    a control-policy:FormatObjective .

<http://www.oesta.gv.at/policies/ObjectsMustBeValid>
    control-policy:measure <http://purl.org/DP/quality/measures#120> ;
    control-policy:value true ;
    a control-policy:RepresentationInstanceObjective .

<http://www.oesta.gv.at/policies/ObjectsMustBeWellFormed>
    control-policy:measure <http://purl.org/DP/quality/measures#121> ;
    control-policy:value true ;
    a control-policy:RepresentationInstanceObjective .

<http://www.oesta.gv.at/policies/SSIMMustBeGT095>
    control-policy:authenticity authenticity:Appearance ;
    control-policy:measure <http://purl.org/DP/quality/measures#1> ;
    control-policy:qualifier qualifiers:GT ;
    control-policy:validInScope
<http://www.example.com/ImageSpecialInterestGroup> ;
    control-policy:value "0.95" ;
    a control-policy:AuthenticityObjective .

<http://www.oesta.gv.at/policies/UseOfFormatShouldBeWidespread>
    control-policy:measure <http://purl.org/DP/quality/measures#162> ;
    control-policy:modality modalities:SHOULD ;
    control-policy:value "widespread"^^xsd:string ;
    a control-policy:FormatObjective .
```

The example illustrates the definition of simple preservation cases that relate to a number of content sets (identified via URIs such as http://www.oesta.gv.at/policies/scanned_papers and user communities, again identified for example as http://www.oesta.gv.at/policies/researchers. There are then a number of objectives defined, some of which are concerned with properties of the formats used (e.g. http://www.oesta.gv.at/policies/FormatShallBeISOStandardized), while others relate to the characteristics or properties of the objects themselves (such as http://www.oesta.gv.at/policies/ObjectsMustBeValid).

This representation of the policies can be queried by the PLATO tool to discover particular objectives that will impact on planning. For example, the following SPARQL query

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX quality: <http://purl.org/DP/quality#>
PREFIX cp: <http://purl.org/DP/control-policy#>
PREFIX pc: <http://purl.org/DP/preservation-case#>
PREFIX oesta-policies: <http://www.oesta.gv.at/policies#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

select ?cp
WHERE {
       ?cp rdf:type ?o .
       ?o rdfs:subClassOf cp:Objective .
```

```
        ?m rdf:type quality:Measure .
        ?m quality:attribute ?a .
        ?a quality:relatesToMediaType
          <http://purl.org/NET/mediatypes/image>.
        ?cp cp:measure ?m .
}
```

retrieves all those objectives that use measures associated with an attribute that relates to images. This uses the associations to media types described in Section 6.1. This particular query will then return the following objectives:

```
<http://www.oesta.gv.at/policies/SSIMMustBeGT095>
<http://www.oesta.gv.at/policies/ImageMustBeIdentical>
<http://www.oesta.gv.at/policies/ImageWidthMustBeUnchanged>
<http://www.oesta.gv.at/policies/ImageHeightMustBeUnchanged>
```

The query

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX quality: <http://purl.org/DP/quality#>
PREFIX cp: <http://purl.org/DP/control-policy#>
PREFIX pc: <http://purl.org/DP/preservation-case#>
PREFIX oesta-policies: <http://www.oesta.gv.at/policies#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

select ?cp ?m ?mod ?qual ?v
WHERE {
        <http://www.oesta.gv.at/policies/ministries_executions_scenario>
          pc:hasObjective ?cp.
        ?cp cp:measure ?m .
        ?cp cp:value ?v
        OPTIONAL {
          ?cp cp:modality ?mod.
          ?cp cp:qualifier ?qual.
        }
}
```

retrieves all the objectives that have been defined for the preservation case http://www.oesta.gv.at/policies/ministries_executions_scenario along with the measures and values associated with those objectives.

# 8 Tooling

The planned deliverable from this task was a suite of vocabularies that provide representation of low level control polices. It is clear, however, that support for end users in using those vocabularies is needed – in particular for authoring objectives. Early experiments in validating the initial models involved authoring control policies at the RDF level, which is unsuitable for users. In the long term, richer policy authoring tools will be required. Tools that use familiar environments such as spreadsheets would be ideal, for example, extending the RightField [rightfield] tool to support progressive drop down lists and selections. Such activity is outside the scope of this task.

In the short term, to address this, a simple prototype web application has been built which allows the construction of control policy objectives through progressive drop down lists driven by the content of the model and subsequent browsing of those objectives. This is sufficient to support initial policy authoring.

The tool is implemented using SPARQL queries (c.f. the examples above in Section 7) against a triple store containing the vocabularies and data.

The screen shots below show the tool in use.

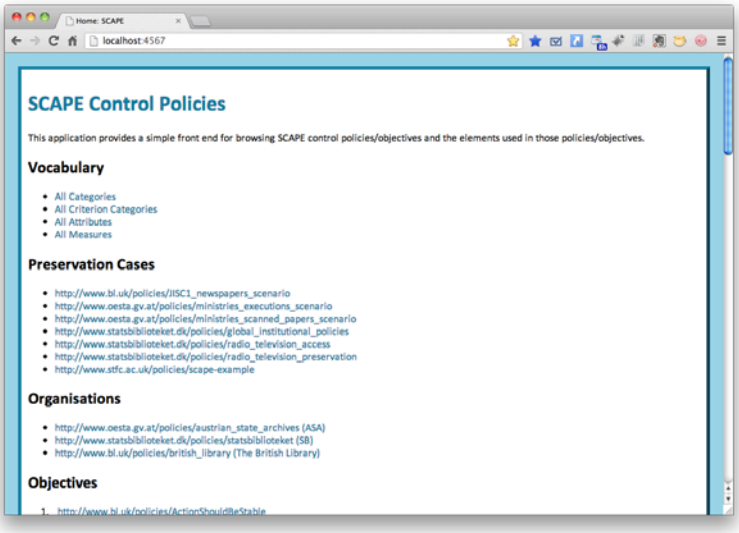



Figure 6 Control Policy Vocabulary Browser

Figure 6 shows the front page of the policy browser. This provides links for browsing the vocabularies themselves (see Figure 7 and Figure 8) and links to the policy instances.

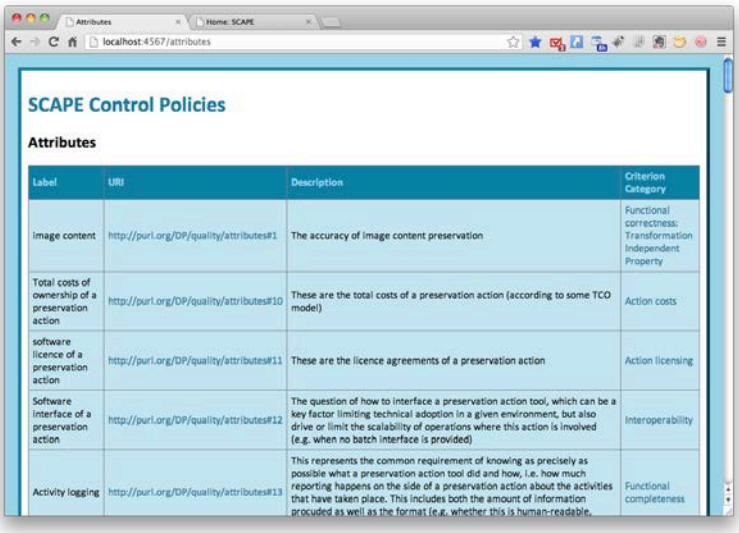



Figure 7 Browsing Attributes

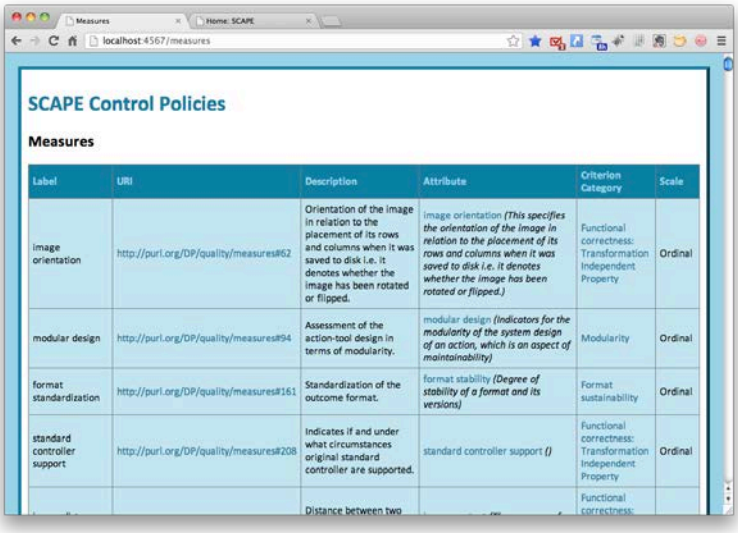



Figure 8 Browsing Measures

Figure 7 and Figure 8 show the browsing of attributes and measures. The particular characteristics of each attribute or measure are shown. Measures are cross-linked to the attribute that the measure is related to. Other aspects of the vocabulary can also be browsed, for example the categories that group attributes.

Figure 9 shows objectives that are defined for a particular Preservation Case. This lists the values associated with the various attributes of the objective.

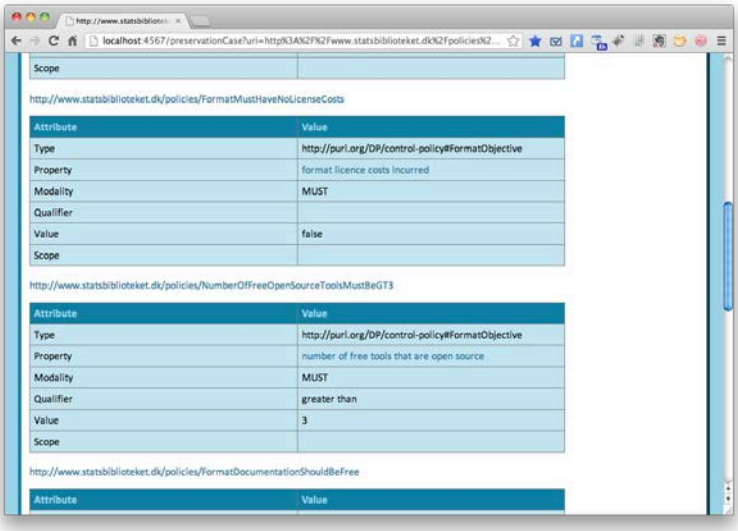



Figure 9 Browsing Objectives

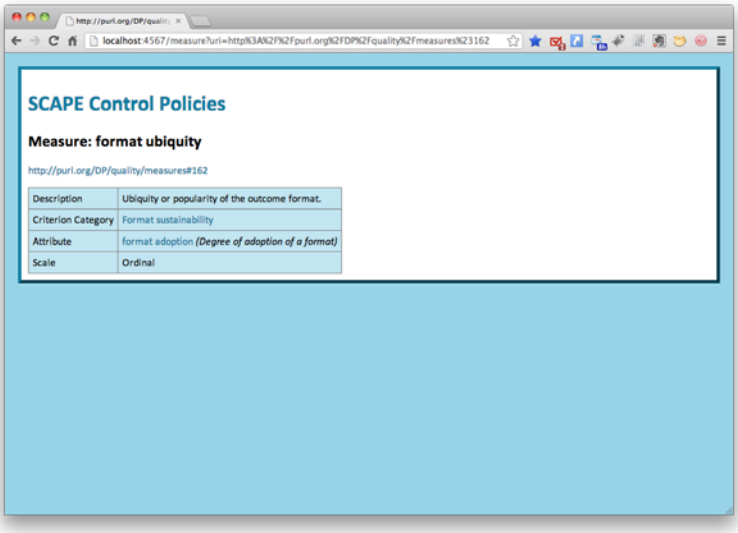



Figure 10 Details for a Measure

Figure 10 shows the details for a particular measure.

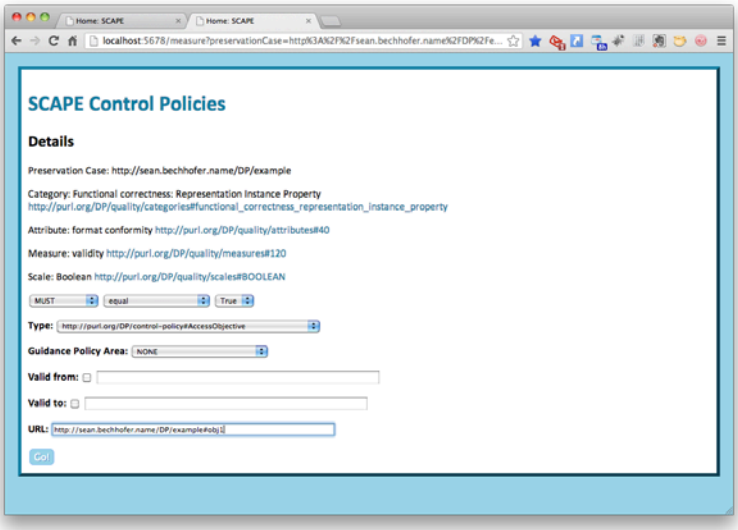



Figure 11 Policy Objective Creation

Figure 11 illustrates the creation of an objective. Here the measure has already been selected. The user is then offered a number of drop down lists that allow for selection of appropriate values for attributes of the objective including the modality, qualifier, value and validity (if appropriate). Once values are selected, an RDF description of the objective is produced and can be inserted into the triple store.

# 9 References

[whitepaper] *A framework for the integration of policies and planning.* Cristoph Becker and Hannes Kulovits. SCAPE Project Internal Whitepaper.

[sbvr] Object Management Group. Semantics of Business Vocabulary and Business Rules (SBVR), Version 1.0. OMG, 2008.

[bmm] Object Management Group. Business Motivation Model 1.1. OMG, May 2010.
[opd] *Open Preservation Data: Controlled vocabularies and ontologies for preservation ecosystems.* Hannes Kulovits, Michael Kraxner, Markus Plangg, Christoph Becker, Sean Bechhofer. 10th International Conference on Preservation of Digital Objects (iPres 2013). 2013
[ppl] *Preservation Policy Levels in SCAPE* - Barbara Sierman, Catherine Jones, Sean Bechhofer and Gry Elstrøm. 10th International Conference on Preservation of Digital Objects (iPres 2013). 2013
[rules] Semantics of Business Vocabulary and Business Rules (SBVR), Object Management Group. Version 1.0. OMG, 2008.
[preserve] *"oh, you wanted us to preserve that?!" statements of preservation intent for the national library of australia's digital collections.* C. Webb, D. Pearson, and P. Koerbin. D-Lib Magazine, 19(1/2), January/February 2013. http: //www.dlib.org/dlib/january13/webb/01webb.html.
[rdf] *Resource Description Framework (RDF): Concepts and Abstract Syntax*. Graham Klyne, Jeremy J. Carroll http://www.w3.org/TR/rdf-concepts/
[rdfs] *RDF Vocabulary Description Language 1.0: RDF Schema.* Dan Brickley, R.V. Guha, http://www.w3.org/TR/rdf-schema/
[owl] *OWL 2 Web Ontology Language Document Overview (Second Edition).* W3C OWL Working Group. http://www.w3.org/TR/owl2-overview/
[skos] *SKOS Simple Knowledge Organization System Reference.* Alistair Miles, Sean Bechhofer, http://www.w3.org/TR/skos-reference/
[measurement] *Software engineering – Software measurement process (ISO/IEC 15939:2002).* ISO/IEC. International Standards Organisation, 2002.
[quality] *Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models (ISO/IEC 25010).* ISO/IEC. International Standards Organisation, 2011.
[service] *Representing Web Service Policies in OWL-DL.* Vladimir Kolovski, Bijan Parsia, Yarden Katz, James A. Hendler: International Semantic Web Conference 2005
[rightfield] *RightField: Embedding ontology annotation in spreadsheets.* Wolstencroft K, Owen S, Horridge M, Krebs O, Mueller W, Snoep JL, du Preez F, Goble CA (2011), Bioinformatics (2011) 15;27(14):pp2021-2