Open Preservation Data: Controlled vocabularies and ontologies for preservation ecosystems

Hannes Kulovits, Michael Kraxner,
Markus Plangg, Christoph Becker
Vienna University of Technology
Vienna, Austria
{kulovits,kraxner,plangg,becker}
@ifs.tuwien.ac.at

Sean Bechhofer
University of Manchester
Manchester, UK
sean.bechhofer@manchester.ac.uk

ABSTRACT

The preservation community is busily building systems for repositories, identification and characterisation, analysis and monitoring, planning and other key activities, and increasingly, these systems are linked to collaborate more effectively. While some standard metadata schemes exist that facilitate interoperability, the controlled vocabularies that are actually used are rare and not powerful enough for the requirements of emerging scalable preservation ecosystems. This article outlines key requirements and elements of such an open ecosystem and discusses the starting points for building such a common language. We then present a core set of controlled vocabulary elements for preservation quality, objectives, policies, and components, and demonstrate how these elements are instantiated to connect preservation planning, preservation watch, and experimentation with preservation policies. We show how these vocabularies are used to enable automation and enable the preservation community to collaborate effectively, and point out extension points and future work.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Information Storage and Retrieval; H.3.7 [**Information Systems**]: Digital Libraries

Keywords

Digital Preservation, Preservation Planning, Preservation Watch, Linked Data, Ontologies, Semantic Interoperability, Workflows

1. INTRODUCTION

Digital preservation aims at keeping digital information authentic, understandable, and usable over a long period of time and across changing technical environments [20]. In recent years the preservation community has come up with a number of independent systems and tools to solve distinct problems in this domain. These systems include repositories, tools for identification and characterisation, analysis and monitoring, and planning. With the digital preservation domain being strongly community-driven, many of today's available systems have been developed directly by individual or collaborating problem owners.

The capabilities a preservation system needs to possess include planning, operations, and monitoring. Preservation planning focuses on the creation of operational preservation plans that contain a decision for a specific preservation action that fulfills clear objectives. Operations focuses on executing the preservation action on content in the production system along with adequate quality assurance measures. Monitoring focuses on gathering and analysing information from different sources internal and external to the organisation, and checking compliance to the organisation's preservation objectives. This needs to be based on a good understanding of organisational policies, which provide the context for preservation. In general terms, policies guide decisions taken within the organisation to achieve long-term goals.

Preservation decision making is guided by information on specific characteristics of actions and aspects such as file formats. Sources providing this kind of information have been implemented and range from online registries and catalogues for file formats and software, to technology watch reports of recognised organisations. Each information source uses its own way to structure data internally and provide it to users. This variety makes it difficult for preservation systems to truly scale up. Furthermore, the information these registries cover is far from complete and often covers only a specific area.

In recent years, several operational software systems have been presented supporting the discussed capabilities. These systems will often be deployed in conjunction with a repository environment. This requires open interfaces and demonstrated integration patterns in order to be useful in practice. We envisage a preservation ecosystem with the following goals:

1. Connect existing systems in a loosely coupled manner.
2. Enable knowledge discovery and exploitation of potential synergies.
3. Facilitate open growth and community participation.
4. Enable automation and scalability.

Apart from open interfaces and reference implementations, this also requires a common language that provides the necessary semantics for the connecting points of these systems, where they communicate about the same concepts. These include objects, file formats, preservation actions and tools, decision criteria and measures, events and conditions.

In this work, we present a loosely-coupled preservation ecosystem where community members in different roles can use an evolving set of tools to collaborate effectively. These tools are linked on the syntactic and semantic level which enables open growth and eases community participation. This leads us to the following questions, which will be discused here:

1. Which elements in a preservation ecosystem play which role towards achieving information longevity, and what are their information requirements?
2. What are the requirements on a language enabling these elements to be connected in a loosely-coupled manner?
3. How can such a language be leveraged in an evolving ecosystem?

The article is structured as follows. The next section outlines key aspects of preservation systems that require integration and discusses some of the major starting points that provided the backdrop and motivation of this work. Section 3 discusses the SCAPE ecosystem of policy-aware operations, planning, and monitoring components, while Section 4 presents the key elements of the common language that enables these systems to exchange information. Section 5 discusses existing applications and outlines benefits and current gaps. Section 6 summarizes the current state of art and points to future work ahead.

2. BACKGROUND

2.1 Systems and tools

Several different systems with specific aims collaborate in a preservation environment and together support the capability of preserving digital information over time. A plethora of software tools exists that perform identification, characterisation, and migration of digital objects. Characterisation tools such as the Digital Repository Object Identification tool (DROID)[1] and JSTOR/Harvard Object Validation Environment (JHove)[2] perform identification, characterisation and validation of digital objects. The File Information Tool Set (FITS)[3] uses a number of tools including DROID, JHove, and Exiftool[4] and provides a unified output. Examples for migration tools include ImageMagick[5] for converting image files, ffmpeg[6] for audio files, and Ghostscript[7] for converting to PDF. The number of available tools however decreases very fast with increasing complexity of the objects.

The service registry CRiB was one of the earliest attempts to wrap migration tools into web services and making them discoverable and usable [10]. The Planets Testbed strived to provide an experimentation environment to evaluate preservation strategies and sharing results [2]. The SCAPE preservation toolset[8] comprises dozens of migration tools ready to install as Debian packages.

The planning tool *Plato*[9] provides systematic decision making support for preservation planning and implements the method introduced in [5]. It includes a model of relevant aspects, entities, and properties that guide preservation planning and offers a standardised view on decision criteria [12]. An integral part of the preservation plan is the decision for a specific preservation action along with concrete quality assurance. Preservation actions may be entire workflows performing complex operations involving identification, migration, and characterisation tools. The workflow management system *Taverna*[10] allows for the definition, and execution of such workflows on different platforms [13]. The platform *myExperiment*[11] integrates with Taverna and makes it possible to share, discover, and reuse workflows [21]. *Preservation operations* is the activity responsible for the execution of this action and reporting on its success. Preservation plans in Plato are specified following a published XML schema.

The preservation monitoring system *Scout*[12] gathers data from various information sources, analyses it and notifies upon the occurrence of configurable events [9]. Scout is an exensible, evolving knowledge base. The information sources it aims at drawing together include content profiles, format registries, software catalogues, experiments carried out in preservation planning, repository systems, organisational objectives, simulation, and human knowledge [4].

The scalable content profiling tool *c3po*[13] (*Clever, Crafty, Content Profiling of Objects*) analyses the technical properties of large sets of objects based on metadata generated by characterisation tools such as FITS and Apache Tika[14]. The generated profile offers a comprehensive and deep insight into the characteristics of the content set in question. Hence it helps to find outliers, objects with particular properties, and combinations of such. Experimentation in preservation planning aims at using samples from the content set that feature a highest possible coverage of occurring properties. Hence the decision making process directly benefits from a thorough analysis of the content set subject to planning.

Technical registries provide information on relevant aspects such as file formats and risks, software products, potential migration paths, and platforms. Such registries have been available for many years and include: the well-established registry *PRONOM*[15] which is curated by the The National Archives UK, the *Global Digital Format Registry (GDFR)*[16] [1], and the *Unified Digital Format Registry (UDFR)*[17] developed by the University of California Curation Center at the California Digital Library. UDFR is a semantic registry and endeavours to unify the content held by PRONOM and

[1] http://digital-preservation.github.io/droid/
[2] http://jhove.sourceforge.net/
[3] http://code.google.com/p/fits
[4] http://www.sno.phy.queensu.ca/~phil/exiftool
[5] http://imagemagick.org
[6] http://www.ffmpeg.org
[7] http://ghostscript.com
[8] http://github.com/openplanets/scape/tree/master/pc-as
[9] http://www.ifs.tuwien.ac.at/dp/plato
[10] http://www.taverna.org.uk/
[11] http://www.myexperiment.org/
[12] http://github.com/openplanets/scout
[13] http://github.com/openplanets/c3po
[14] http://tika.apache.org/
[15] http://www.nationalarchives.gov.uk/PRONOM/
[16] http://www.gdfr.info
[17] http://www.udfr.org

GDFR. The semantically enhanced *P2* registry [25] pulls together content from PRONOM and enriches it with data from *dbpedia*[18]. The *Conversion Software Registry (CSR)*[19] focuses on software packages that support migration of files. CSR finds migration paths of configurable length based on input and output formats provided by the user.

All these registries have been designed with a specific concern in mind. For example, CSR provides migration pathways with some information on the tools but lacks information about file formats. PRONOM on the other hand gives detailed information about some file formats and selected software tools for migration, but does not provide evidence about their quality. P2 is yet sparsely filled and used in a limited number of scenarios. Its successor, LDS3 (Linked Data Simple Storage Specification)[20], provides an open data publication platform based on Linked Data principles. It does not itself provide a common language for describing published preservation data [24], but of course supports the usage of ontologies.

In reality, the information content of moderated registries tends to be modest in coverage, with many important information needs left unfulfilled. We observe that the design assumption of moderated registries, expecting that a controlled point of reference will be able to cope with evolving facts, leads to knowledge gaps. For instance, a migration tool may be considered as stable in one of the registries, but large-scale experiments conducted by an organisation using the tool on content with specific properties might reveal that the tool crashes in particular cases or does not run on a particular platform. On the other hand, open information models are better positioned to capture the evolving facts and knowledge, and technologies such as RDF provide the opportunities to design an ecosystem made for an open world and evolving technologies.

2.2 Policies

Policies provide the context for successful preservation planning, operation, and monitoring. They govern and control decisions within the organisation. Policies often provide guidance on a high-level, for instance by expressing value propositions to customers. However, there is no clear specification of the exact meaning of a "preservation policy". Sometimes it is used as describing the overall strategy of a cultural heritage institution and its commitment to keep digital material accessible over time. Common examples for policy statements also specify strategies and commitments of an organisation, based on regulatory compliance such as statements in the ISO 16363 Repository Audit and Certification catalogue [14] or on industry practice such as statements collected in a recent preservation policy study [3]. These are well known, but do not separate concerns clearly and often mix objectives with functional means to implement capabilities. Hence, their impact is not always well-understood, and operations based on these are complex to implement.

Most usages of "policies" correspond to what the Object Management Group (OMG) standards call "business 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 [s]trategies and [t]actics" [18, 19]. Preservation policies hence should provide the mechanisms to document and communicate about key aspects of relevance, in particular drivers and constraints and the goals and objectives motivated by them. At present, there are no establishd standards for preservation policies relevant to planning or for aspects such as monitoring specifications, Service Level Agreements for preservation operations, or system interfaces. Smith et al. [23] point out that preservation systems operate on a rule level and presents policies that have been translated into rules to be enforced in a repository.

2.3 Standardisation

The digital preservation community has embarked on numerous endeavours towards standardisation of certain aspects required to achieve information longevity. The Planets project[21] presented a conceptual model and vocabulary for representing an organisation's values and constraints[7]. The SHAMAN project[22] has approached digital preservation from an Information Systems point of view and provides a contextualized capability-based view on digital preservation. The SHAMAN Reference Architecture defines the core capabilities Preservation Operation, and Preservation Planning including Monitoring [22]. The SCAPE project is taking this further by implementing appropriate scalable systems that support these capabilities.

A key activity in preservation planning is systematic testing of preservation software. The quality of preservation actions such as tools for migration, but equally of emulators, has to be determined to be able to reach an informed decision for a specific action. The ISO standard 25010 - 'Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models' [16] has its roots in the the earlier ISO 9126 family and defines a hierarchy of high-level quality attributes. It combines characteristics relating to the outcome of interaction when the software product is used in a specific context ("quality in use") and characteristics relating to static properties of software and dynamic properties of the computer system ("product quality") [16]. The ISO 25010 quality model has been adopted in preservation planning to classify decision criteria [12].

Digital objects have certain significant properties that need to be preserved for the objects' performance to be deemed authentic. Significant properties have been extensively analysed in the InSPECT project[23], which has provided a detailed analysis of significant properties of different types of digital objects such as vector images, moving images, and software [11].

Since preservation is a continuous process, a preservation system needs to be capable of monitoring aspects that influence the preservation process. A preservation watch component is designed in [4] for monitoring internal (systems and operations in place, assets and activities) and external (e.g. user communities, technologies, available solutions) influence factors.

The SCAPE project[24] is focusing its work on scalable operations to enable the preservation of large sets of digital information [8]. The components developed in the project use open APIs to enable communication between e.g. plan-

[18] http://dbpedia.org/
[19] http://isda.ncsa.uiuc.edu/NARA/CSR
[20] http://www.lds3.org/
[21] http://www.planets-project.eu
[22] http://shaman-ip.eu
[23] http://www.significantproperties.org.uk/
[24] www.scape-project.eu

ning, watch and repositories. Open APIs make interfaces available to the public and thus enable continuous growth of systems by community participation. Standardisation in this area however needs to go one step further and enable semantic interoperability of components. Information exchanged between these components also needs to be opened up to the community to build synergies, enable knowledge discovery, and move from static to dynamically growing information sources.

2.4 Interoperability

Each of the information sources described above has been developed for particular intended users, types of objects, platforms, and with specific domain and project needs in mind. Hence the way they structure data internally and provide it to users vary. Standard metadata schemes are often adopted to facilitate interoperability of systems. The *Preservation Metadata Implementation Strategies (PREMIS)*[25] working group has produced a technically neutral scheme for preservation matadata. It links intellectual entities, objects, rights, events, and agents to provide a data dictionary. One of the most prominent and commonly used metadata schemes is *Dublin Core (DC)*[26]. DC metadata terms describe resources of various types to enable discovery.

Many systems in the digital preservation domain, including PRONOM and UDFR, adopt linked data techniques to share their data and make them re-usable. At the core of this effort is the Resource Description Framework (RDF)[27]. RDF is a standard model to enable the representation of data and metadata that essentially allows for the expression of subject-predicate-object triples. The Web Ontology Language (OWL)[28] provides further mechanisms for the description of vocabularies or ontologies that define classes and properties. These can be used to annotate, describe and define resources. OWL has a well-defined semantics that facilitates the use of reasoning, supporting ontology management and querying of data. Collections of RDF statements (RDF graphs) can be serialised using a variety of concrete formats including RDF/XML and N3[29], while SPARQL[30] provides a language for querying and manipulating RDF graphs.

3. A PRESERVATION ECOSYSTEM

We observe that many different systems exist that support in digital preservation efforts, and many information sources and tools exist that are directly relevant to the preservation efforts of dedicated systems. Not all of these information sources and tools originate from the digital preservation domain. Components in an open preservation ecosystem need to use standards and appeal beyond digital preservation to enable growth and community participation. They should be built around a simple core instead of aiming for being all-encompassing and overwhelming. It becomes clear that the goal should be to connect and enable rather than impose and restrict. The key domains of the ecosystem in focus are the following.

1. **Organisation.** – The organisation operates an information system, e.g. a *repository*, concerned with the preservation of digital information over time. People acting on behalf of the organisation adopt a number of *tools* in the process of preserving the organisation's digital holdings. These include tools for identification, characterisation, migration, and emulation. The organisation formulates and makes available its *goals and policies* that guide operations.

2. **Solution components.** – This domain includes software tools, platforms, and services addressing real needs of the organisation. These components are developed, maintained, and distributed by commercial or non-commercial solution providers concerned with providing solutions according to market needs. The main building blocks include software tools for identification (e.g. *DROID*[31], and the Linux command *file*), characterisation and validation (e.g. *FITS*), migration (e.g. ImageMagick *convert*), emulation, and quality assurance.

3. **Decision support and control.** – Systems and tools in this domain support the decision making process in preservation planning and exerting control over operations. They are capable of analysing digital objects and providing descriptive information about these objects, monitor changes in the technical environment, and support in the decision making for a specific preservation action. The main building blocks in focus of this paper include *Plato*, *c3po*, and *Scout*.

4. **Community environment.** – Individual people as well as organisations and institutions with a particular concern develop and populate systems that drive the preservation process. These systems contain essential information on aspects relevant to preservation. The main building blocks in this domain include technical registries such as *PRONOM*, but increasingly extend to environments not originally emerging within digital preservation, such as the workflow sharing platform *myExperiment* or public open source software repositories.

Each software system requires information about certain domain entities. For example, *c3po* needs to describe objects it analyses, and *preservation tools* need to report measures. The planning tool *Plato* needs to discover preservation actions, evaluate actions, and describe plans. *Scout* needs to collect measures on all these entitites, detect conditions, and observe events. Finally, decision makers need to describe their goals and objectives in a way understandable by the systems, so that decision support can provide customized advice and support that befits their specific policies and constraints.

4. A COMMON LANGUAGE

4.1 Requirements

From the discussion of the preservation ecosystem and its building blocks it becomes evident that a common language is required to achieve semantic interoperability. The

[25] http://www.loc.gov/standards/premis/
[26] http://dublincore.org/
[27] http://www.w3.org/RDF/
[28] http://www.w3.org/TR/owl2-overview/
[29] http://www.w3.org/TeamSubmission/n3/
[30] http://www.w3.org/TR/sparql11-overview
[31] https://github.com/digital-preservation/droid

expected benefits include the ability to communicate about shared concepts, i.e. query across organisational information, policies, monitoring requests, preservation plans, and preservation components using a single framework. To further align with requirements for preservation systems, such a common language needs to fulfill three key objectives.

1. The vocabulary and instances need to cover elements from different domains and make meaningful connections.
2. The model and its representation need to be accessible to both people and software tools.
3. The model should be based on open standards and Linked Data principles.
4. It should be modular and easily extensible, while scaling freely.

The vocabulary described in this article strives to achieve these objectives by building on a simple core model and applying Linked Data principles. By providing a permanently linked core ontology applying across domains and the ability to extend it continuously, it should provide the appropriate support for an evolving ecosystem. The next sections will describe the core domains of the initial model, while Section 5 shows how the existing models are used across the SCAPE ecosystem to improve information sharing, reasoning and discovery.

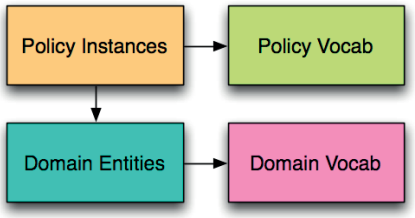

Figure 1: Models

4.2 Control policies

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"[18]. A control policy contains quantified, precise statements of facts, constraints, objectives and directives about these entities and their properties. Such policies are not directly enforceable. They contain statements that can be fully represented in a machine-understandable model, but the policies are often not directly actionable in that it does not make sense to directly execute them. For example, multiple control statements may contradict each other. A decision making process such as preservation planning translates these policies into a specified set of rules in a plan. This rule set is then actionable 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.

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 illustrates these interactions. Figure 2 provides a high-level overview of the policy model including the classes and properties discussed.

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 [26], 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*" [16, 15]. 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.

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.

Figure 3 shows a set of triples describing a concrete measure for determining the degree of adoption of a certain file format.

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

`m` A measure pertaining to an authenticity, access, action, or representation instance objective.

`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, should not).

A sample preservation case is shown in Figure 4. This case relates to a newspaper collection at the Austrian State Archives which is mainly accessed by researchers. It includes an example of a concrete policy statement from this case stating that the degree of adoption of file formats should be ubiquitous.

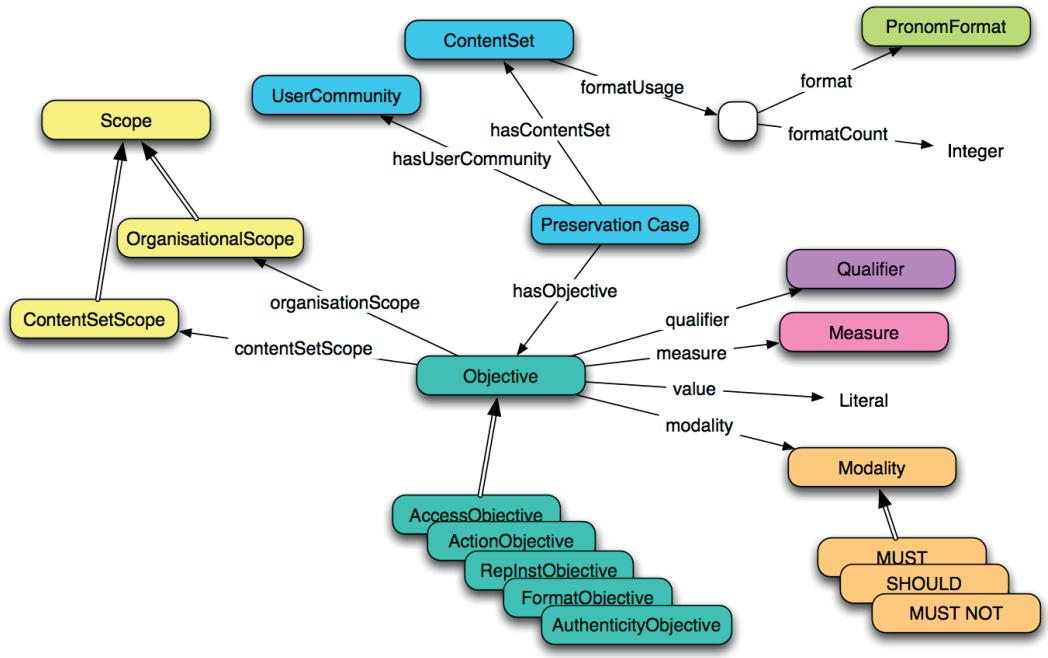

Figure 2: Core model of control policies

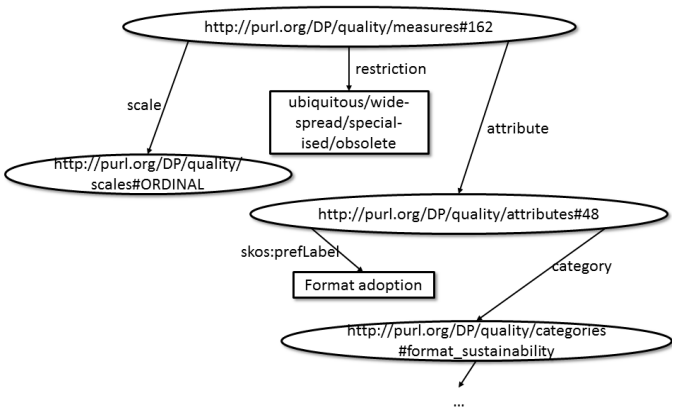

Figure 3: A concrete measure described by triples

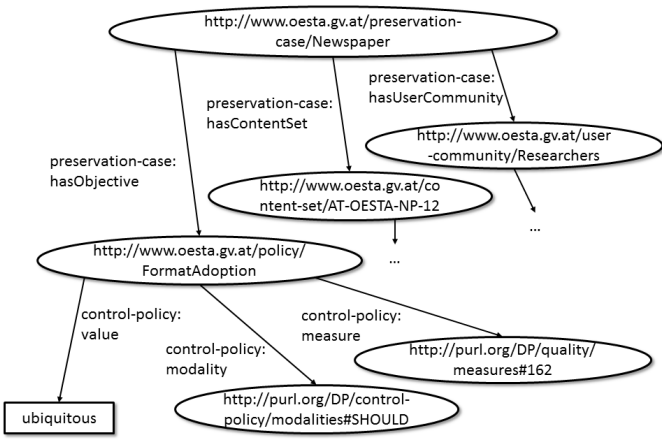

Figure 4: Sample preservation case described by triples

4.3 Domains

4.3.1 Preservation Case

The preservation case documents the particularities 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. The time horizon will often be determined by legal requirements and contextual issues. To a large extent, access requirements are derived from knowledge on the user community and their used technology. Considering Figure 4 the content set and user community elements provide the extension points to further describe the preservation case.

4.3.2 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:

- **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 Objectives** describe objectives referencing a property that representations of content, such as files and bytestreams, should or must have. This includes aspects such as compression, encryption, size, or validity.
- **Access Objectives** 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 Objectives**, finally, describe constraints on the preservation action process, such as the maximum time or memory resources available or a restriction on allowed licensing.

4.3.3 *Quality*

Preservation cases are associated to objectives, and each objective references a particular aspect of quality in objects, representations, formats, or actions. One of the key activities in preservation planning is the assessment of such quality. Hence, attributes and measures are required to be capable of evaluating preservation solution components and their ability to achieve objectives and minimize risks. Examples include "Format shall be standardised by ISO", and "Image size must be retained".

4.3.4 *Solution*

Software components deployed in the preservation ecosystem require standardised descriptions to enable automation and scalability. For example, planning and monitoring need to discover, evaluate and compose components with minimal manual effort. This will be described in Section 4.4.

4.4 Component profiles

Software tools play a key role in preservation systems. They are deployed for tasks such as format identification, characterisation, migration, or quality assurance. The result of the decision making process in preservation planning is a concrete preservation action to be applied to an identified set of digital objects, including mechanisms for validating the result. Figure 5 shows a high-level view of an executable plan as Taverna workflow with different types of activities (e.g. red circles represent invocation of external tools). Hence, the plan deployed by operations needs to make use of this diverse set of tools and services. Running these tools often requires technical knowledge and expertise in the digital preservation domain. The output generally is (semi-)structured data that neither has a standardised format nor follows a common vocabulary.

To overcome these shortcomings and reduce the overall effort in preservation operations and decision making, an analysis was conducted that identified the following requirements.

1. **Publishing** of components is necessary to allow tool developers and preservation experts to share solution components and expertise needed to create preservation components and enable reuse by others in the community.
2. **Discoverability** of such components is required to enable preservation planning to find the most relevant published components. This allows reuse during planning experiments and in operational plan execution.

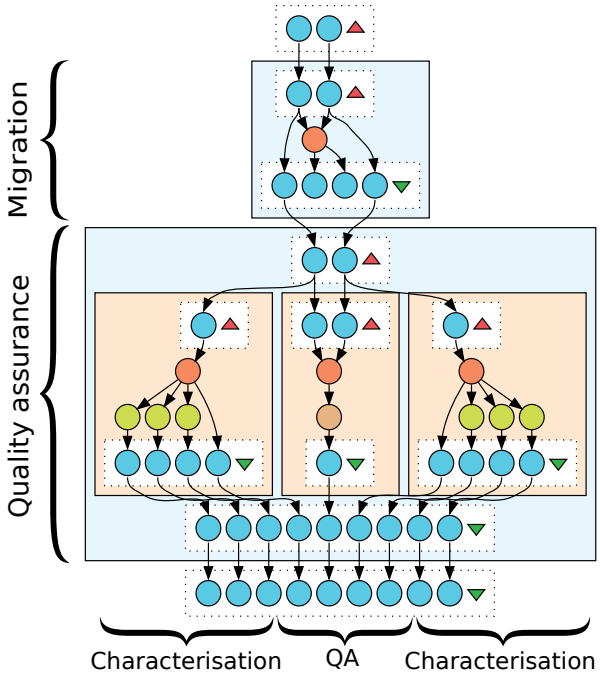

Figure 5: An executable preservation plan

3. **Automated execution** is required to increase the scalability of operations and preservation planning by allowing to run automated planning experiments on representative samples of the content set subject to preservation. This reduces the preservation effort by automating the execution of preservation actions on a large number of tool and parameter combinations. Additionally, it enables automated characterisation and quality assurance of action results.
4. **Reproducibility** is key requirement for trustworthy, evidence-based preservation. Experiments conducted in the course of preservation planning need to be reproducible. Hence, a thorough description of the requirements and dependencies of these tools is required. This is an essential part of the evidence that a preservation plan needs to provide, but equally important for operational deployment.
5. **Standardised output** is required to enable comparability of measures provided by the diverse set of available tools and services. Therefore, the output of components must be well-defined and follow a common vocabulary. This not only allows automated evaluation of experiments in preservation planning, but also enables collecting real world data on tool usage and quality of tool results across organisational boundaries [4].
6. **Composition** is required to allow different components to be combined in an executable plan that can be executed in a repository environment. This should be as easy and automated as possible.

Three main component types of digital preservation tools are in focus:

1. **Migration** components support migrating files to different formats. They must specify supported migration paths.

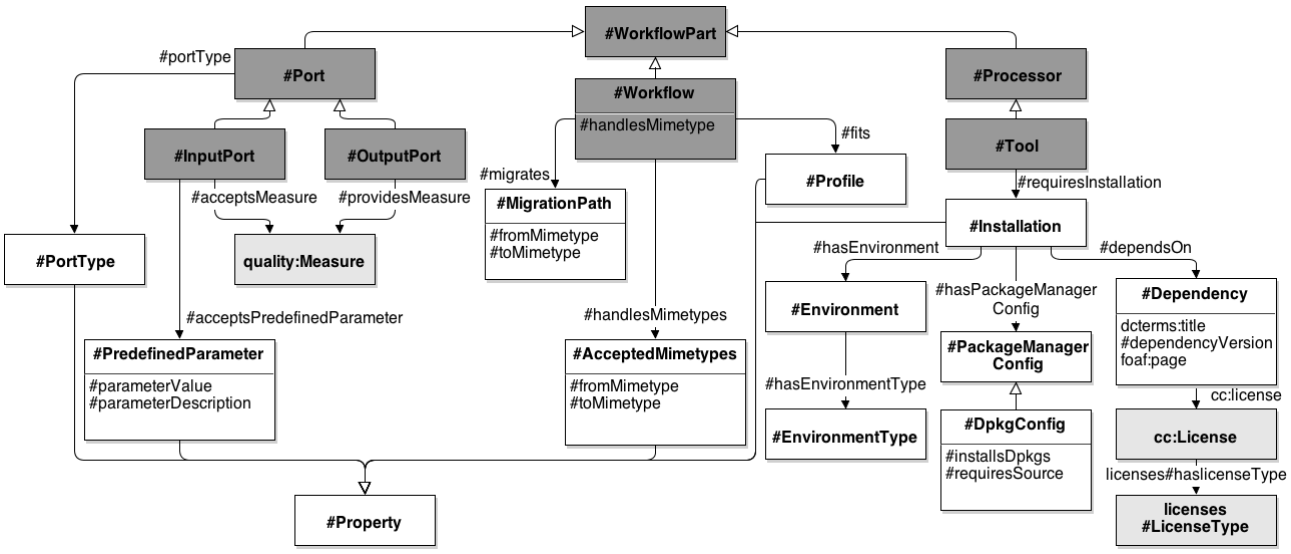

Figure 6: Overview of the ontology http://purl.org/DP/components

2. **Characterisation** components characterise an object and provide specific measures as output. Their specification must contain supported input formats as well as the measures they provide.

3. **Quality assurance** components provide measures that can be used to assess authenticity, validity, cost, and risk. They are split into three subtypes: *Object comparison components* accept two objects as input and provide measures about the degree of similarity. The components need to specify format pairs they support. *Property comparison components* focus on comparing measures and report on the degree of similarity. *Validation components* are used to validate one object against measures. Thus they have to provide supported formats and measures they provide.

In principle, emulation components can easily be added to this ontology. The current focus of workflow development, publication and discovery, however, is on migration and associated characterisation and quality assurance.

Taverna workflows provide a common, platform independent language to execute command line tools and other services and perform pre- and postprocessing on data. All components have to specify the environment they require as well as dependencies needed to execute. Taverna workflows and contained workflow parts can be annotated with human-readable free-text annotations[32]. More complex metadata can be added as semantic annotations based on RDF. The workflow sharing environment myExperiment[33] is an established platform for publishing and discovering workflows and supports querying by annotations.

Preservation components are built on top of Taverna workflows. The Taverna Workbench supports creating components according to component profiles and publishing them to a component catalogue. It also provides basic validation against profiles. Component profiles allow to define the interface and required metadata of workflows as XML documents[34]. As part of the metadata specification, they also define the ontologies used for semantic annotations.

For preservation components, the new ontology `http://purl.org/DP/components` provides a vocabulary to annotate workflows with necessary metadata. Figure 6 shows its classes and properties. The ontology contains classes for the workflow parts that can be annotated. The ports, the workflow itself and associated processors, in the common case preservation tools, each have properties that link them to annotations. For example, a workflow fits a specific profile (such as migration), hence supports a certain set of migration paths, and handles specific mimetypes. Input and output ports are linked to measures in the quality ontology. Annotations can either be literals, individuals already defined in the ontoloy, or more complex RDF graphs from the ontology.

All components must be annotated with the profile they fit. If external tools are used in the component, it must provide the metadata needed to enable execution. These dependencies are modeled as *Installations*. Installations can be used in an *environment* and describe their dependencies, including the license. Further configuration for package managers can be provided to allow automated installation of the dependencies.

5. SUMMARY AND APPLICATIONS

The last section introduced a controlled vocabulary for preservation cases and associated objectives, quality, and solution components. This common language enables interoperability between the building blocks of the preservation ecosystem. A pictorial view of the ecosystem and its building blocks is shown in Figure 7. These include the software systems *Plato*, *Scout*, *c3po*, and *myExperiment* platform which are key elements of SCAPE. The policy vocabulary we proposed is the connecting element between these software systems. The organisation specifies control policies for a spe-

[32] `http://dev.mygrid.org.uk/wiki/display/taverna/Annotations`

[33] `http://www.myexperiment.org/`

[34] `http://ns.taverna.org.uk/2012/component/profile/ComponentProfile.xsd`

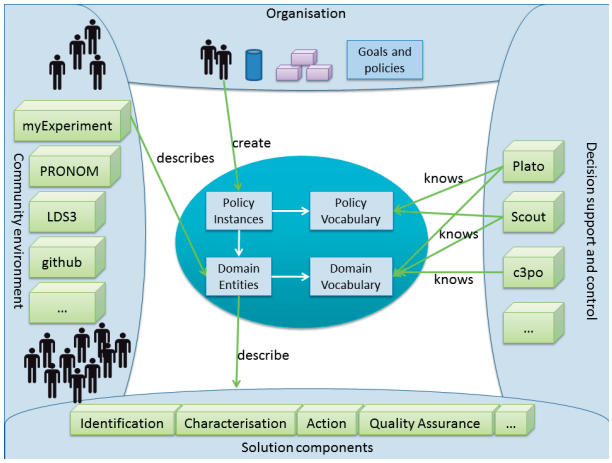

Figure 7: The SCAPE Preservation Ecosystem

cific preservation case, i.e. an identified set of objects that is intended to be preserved for a specified user community. This enables Scout to detect violations and trigger Plato to create a preservation plan for the identified content set. Plato implements the policy vocabulary and relevant domain vocabularies. Hence, Plato can directly incorporate the organisation's objectives, constraints, and directives into planning. c3po creates content profiles using elements from relevant domain vocabularies to describe the digital objects in the content set. This content profile constitutes an essential part of the plan. The component profile allows Plato to query for relevant components on the platform myExperiment, but it also allows standardized specification of executable preservation workflows and their deployment onto target repository environments.To illustrate how the common language is used to improve the effectiveness and efficiency of planning, monitoring and operations, this section discussess several use cases in turn.

Creating policies. To specify control policies, the decision maker leverages the existing vocabulary of domain and policy constructs. Most common policies on this contextual level are until now implicit in the organisational context and used to be discovered in tedious activities within preservation planning [6, 17]. Making these goals and constraints explicit on a higher level with standardised vocabulary enables the decision support tools to offer much more effective support. Tool support for the formulation of policy statements is currently being developed to guide decision makers through a progression of statements that comprise a preservation case. These policies can then be stored in the planning component Plato and the monitoring component Scout, both of which are making use of this organisational context in specific ways:

Automated detection of policy violations. Scout is able to correlate statements in a preservation policy model with the information obtained about the state of affairs in a repository. This most importantly includes the content profile created by *c3po*, which can be queried for violations of specific objectives. For example, the existence of encrypted or compressed files may be not desired. Detecting the existence of such a mismatch causes a notification event to be raised to the attention of the responsible decision maker.

Objective tree construction for evaluation. Upon detection of a non-conforming state or a risk, a mitigation strategy can consist of developing a preservation plan using *Plato*. This in turn is greatly eased by the policy awareness of *Plato 4*, which is able to derive the entire tree of objectives, and measures used for evaluating alternative actions from the control policy model.

Discovery of action components in Plato 4 is enabled through the myExperiment site, where applicable components can be queried, downloaded and executed in a test environment, using the dependency specification to automate installation of required packages. Experimental information that is gathered about the behaviour of tools in real environments on the actual data is associated to the well-defined measurement ontology, which enables cross-linking of cases within an organisation, but also across organisations. Aggregate statistics will in the future be published and can be monitored in Scout, which in turn will enable proactive recommendation of likely successful candidates based on the policies the decision maker's organisation.

For a thorough evaluation of improvements achievable by the integration of the policy vocabulary into Plato 4, we want to refer to a recent controlled case study we carried out with the State and University Library Denmark [17].

6. CONCLUSIONS AND OUTLOOK

This article discussed the information requirements of key building blocks in a preservation ecosystems and showed how controlled vocabularies and ontologies can be leveraged to connect these systems in a loosely-coupled manner to improve knowledge discovery and automation.

We outlined the key systems Plato, c3po, Scout, myExperiment, and Taverna and introduced a common language as connecting element. To enable a loosely-coupled preservation ecosystem where the preservation community can use continuously maturing software tools and collaborate efficiently and effectively, the common language facilitates the systems to be linked on the syntactic and semantic level. We introduced a policy vocabulary based on open standards including RDF and OWL, which enables the ecosystem building blocks to be linked. Concrete policy instances expressed using the policy vocabulary link entities from other domains. Scout can detect policy violations and trigger planning for a specific content set. Decision makers act upon this notification and create a preservation plan.

The current vocabulary presents an important milestone. Current work is geared towards linking in additional existing ontologies to include aspects such as software properties covered in the the Software Ontology (SWO)[35]. On the other hand, we are developing higher level ontology concepts closely linked to preservation intent statements [26]. This aims at dramatically reducing the level of detail required to define objectives related to preservation cases. For example significant properties making up the "appearance" of digital documents can be identified and grouped. An ontology pulling together these properties could reduce the effort of curators to defining "*Appearance must be preserved*" instead of having to deal with the individual technical properties. The decision support system can then derive the set of measures required to assess the authenticity with respect to appearance of specific documents.

Finally, current implementation work on Plato and Scout

[35] `http://theswo.sourceforge.net`

is focused on leveraging this language further.

- Publication of quality assurance components using annotations that specify standardised measures enables Plato to integrate automated evaluation in the experiment workflows and include service-level agreement (SLA) specifications in the generated preservation plans.
- The execution of these generated plans can then be monitored for compliance to the SLAs specifications expressed using the domain vocabulary.
- Additionally, experience sharing on public data endpoints will enable the monitoring of risks and opportunities connected to components and quality measures.

Acknowledgements

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

7. REFERENCES

[1] S. L. Abrams. Establishing a global digital format registry. *Library Trends 54 (1) Summer 2005*, pages 125–143, 2005.

[2] B. Aitken, P. Helwig, A. N. Jackson, A. Lindley, E. Nicchiarelli, and S. Ross. The planets testbed: Science for digital preservation. *Code4Lib*, 1(5), June 2008. See http://journal.code4lib.org/articles/83.

[3] N. Beagrie, N. Semple, P. Williams, and R. Wright. *Digital Preservation Policies Study: Final Report*. HEFCE, October 2008.

[4] C. Becker, K. Duretec, P. Petrov, L. Faria, M. Ferreira, and J. C. Ramalho. Preservation watch: What to monitor and how. In *9th International Conference on Preservation of Digital Objects (IPRES 2012)*, 2012.

[5] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries (IJDL)*, December 2009. `http://dx.doi.org/10.1007/s00799-009-0057-1`.

[6] C. Becker and A. Rauber. Preservation decisions: Terms and Conditions Apply. Challenges, Misperceptions and Lessons Learned in Preservation Planning. In *Proc. JCDL 2011*, June 2011.

[7] A. Dappert and A. Farquhar. Modeling organizational preservation goals to guide digital preservation. In *The Fifth International Conference on Preservation of Digital Objects (iPRES 2008)*, 2008.

[8] O. Edelstein, M. Factor, R. King, T. Risse, E. Salant, and P. Taylor. Evolving domains, problems and solutions for long term digital preservation. In *Proc. of iPRES 2011*, 2011.

[9] L. Faria, C. Becker, P. Petrov, K. Duretec, M. Ferreira, and J. Ramalho. Design and architecture of a novel preservation watch system. In *14th International Conference on Asia-Pacific Digital Libraries (ICADL 2012)*, 2012.

[10] M. Ferreira, A. A. Baptista, and J. C. Ramalho. An intelligent decision support system for digital preservation. *International Journal on Digital Libraries*, 6(4):295–304, July 2007.

[11] S. Grace, G. Knight, and L. Montague. *InSPECT Final Report*. InSPECT (Investigating the Significant Properties of Electronic Content over Time), December 2009. `http://www.significantproperties.org.uk/inspect-finalreport.pdf`.

[12] M. Hamm and C. Becker. Impact assesment of decision criteria in preservation planning. In *Proc. of IPRES 2011*, 2011.

[13] D. Hull, K. Wolstencroft, R. Stevens, C. A. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(Web-Server-Issue):729–732, 2006.

[14] ISO. *Space data and information transfer systems - Audit and certification of trustworthy digital repositories (ISO/DIS 16363). Standard in development*, 2010.

[15] ISO/IEC. *Software engineering – Software measurement process (ISO/IEC 15939:2002)*. International Standards Organisation, 2002.

[16] ISO/IEC. *Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models (ISO/IEC 25010)*. International Standards Organisation, 2011.

[17] H. Kulovits, C. Becker, and B. Andersen. Scalable preservation decisions: A controlled case study. In *Archiving 2013*. Society for Imaging Science and Technology, 2013.

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

[19] Object Management Group. *Business Motivation Model 1.1*. OMG, May 2010.

[20] J. Rothenberg. Ensuring the longevity of digital documents. *Scientific American*, 272, 1995.

[21] D. D. Roure, C. A. Goble, and R. Stevens. The design and realisation of the $\text{my}_{\text{experiment}}$ virtual research environment for social sharing of workflows. *Future Generation Comp. Syst.*, 25(5):561–567, 2009.

[22] SHAMAN. Shaman reference architecture v3.0. Technical report, SHAMAN project, 2012.

[23] M. Smith and R. W. Moore. Digital archive policies and trusted digital repositories. *International Journal of Digital Curation*, 1(2), 2007.

[24] D. Tarrant and L. Carr. Lds3: applying digital preservation principals to linked data systems. In *9th International Conference on Preservation of Digital Objects (IPRES 2012)*, 2012.

[25] D. Tarrant, S. Hitchcock, and L. Carr. Where the semantic web and web 2.0 meet format risk management: P2 registry. *The International Journal of Digital Curation*, 1(6):165–182, June 2011.

[26] C. Webb, D. Pearson, and P. Koerbin. "oh, you wanted us to preserve that?!" statements of preservation intent for the national library of australia's digital collections. *D-Lib Magazine*, 19(1/2), January/February 2013. `http://www.dlib.org/dlib/january13/webb/01webb.html`.