# Scalable decision support for digital preservation

Christoph Becker
*Faculty of Information, University of Toronto, Toronto, Canada*
Luis Faria
*Innovation, KEEP SOLUTIONS, Braga, Portugal, and*
Kresimir Duretec
*Information and Software Engineering Group, Vienna University of Technology, Vienna, Austria*

Received 16 July 2013
Accepted 21 August 2013



## Abstract

**Purpose** – Preservation environments such as repositories need scalable and context-aware preservation planning and monitoring capabilities to ensure continued accessibility of content over time. This article identifies a number of gaps in the systems and mechanisms currently available and presents a new, innovative architecture for scalable decision-making and control in such environments.

**Design/methodology/approach** – The paper illustrates the state of the art in preservation planning and monitoring, highlights the key challenges faced by repositories to provide scalable decision-making and monitoring facilities, and presents the contributions of the SCAPE Planning and Watch suite to provide such capabilities.

**Findings** – The presented architecture makes preservation planning and monitoring context-aware through a semantic representation of key organizational factors, and integrates this with a business intelligence system that collects and reasons upon preservation-relevant information.

**Research limitations/implications** – The architecture has been implemented in the SCAPE Planning and Watch suite. Integration with repositories and external information sources provide powerful preservation capabilities that can be freely integrated with virtually any repository.

**Practical implications** – The open nature of the software suite enables stewardship organizations to integrate the components with their own preservation environments and to contribute to the ongoing improvement of the systems.

**Originality/value** – The paper reports on innovative research and development to provide preservation capabilities. The results enable proactive, continuous preservation management through a context-aware planning and monitoring cycle integrated with operational systems.

**Keywords** Preservation planning, Scalability, Monitoring, Preservation watch, Digital libraries, Repositories

**Paper type** Research paper

## 1. Introduction

Digital preservation aims at keeping digital information authentic, understandable and usable over long periods of time and across ever-changing social and technical environments (Rothenberg, 1995; Garret and Waters, 1996; Hedstrom, 1998). The

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

OCLC Systems & Services: International digital library perspectives
Vol. 30 No. 4, 2014
pp. 249-284

1065-075X
DOI 10.1108/OCLC-06-2014-0025

challenge of keeping digital artifacts accessible and usable while assuring their authenticity surfaces in a multitude of domains and organizational contexts.

While digital longevity as a challenge is increasingly encountered in domains as diverse as high-energy physics and electronic arts, the repository is still the prototypical scenario where the concern of longevity is of paramount importance and libraries continue to play a strong role in the preservation community. Repository systems are increasingly made fit for actively managing content over the long run so that they can provide authentic access even after the availability of the original creation context, both technical and social.

In this process, they have to address two conflicting requirements:

(1) the need for trust, a fundamental principle that is indispensable in the quest for long-term delivery of authentic information; and

(2) the need for scalability, arising from the ever-rising levels of digital artifacts deemed worthy of keeping.

Systems that address aspects of preservation include repository software, tools for identification and characterization of digital artifacts, tools for preservation actions such as migration and emulation, systems to address aspects of analysis and monitoring and preservation planning. It is understood today that automating most aspects of an operational preservation system is a crucial step to enable the scalability required for achieving longevity of digital content on the scales of tomorrow. Such automation is not only required within components of a complex system but also needs to address systems integration, information gathering and, ultimately, decision support.

The core capabilities that an organization needs to establish cover:

- preservation operations, i.e. preservation actions such as emulation, virtualization and migration of digital objects to formats in which they can be accessed by the users, but also object-level characterization, quality assurance and metadata management;
- preservation planning, i.e. the creation, ongoing management and revisions of operational action plans prescribing the actions and operations to be carried out as means to effectively safeguard, protect and sustain digital artifacts authentically and ensuring that the means to access them are available to the designated community; and
- monitoring as a sine-qua-non of the very idea of longevity. Most of the risks that need to be mitigated to achieve longevity stem from the tendency of aspects in the socio-technical environment to evolve and sometimes change radically. Without the capability to sustain a continued awareness of a preservation system and its environment, preservation will not achieve its ultimate goal for long.

Monitoring focuses on analyzing information gathered from different sources, both internal and external to the organization, to ensure that the organization stays on track in meeting its preservation objectives (Becker *et al.*, 2012). Such awareness needs to be based on a solid understanding of organizational policies, which provide the context for preservation. In general terms, it can be said that policies "guide, shape and control" decisions taken within the organization to achieve long-term goals (Object Management Group, 2008; Kulovits *et al.*, 2013b).

Monitoring, policy-making and decision-making processes are guided by information on a variety of aspects ranging from file format risks to user community trends, regulations and experience shared by other organizations. Sources that provide this kind of information include online registries and catalogues for software and formats or technology watch reports of recognized organizations. These are increasingly available online, but the variety of structures, semantics and formats prohibit, so far, truly scalable approaches to utilizing the knowledge gathered in such sources to provide effective decision support (Becker *et al.*, 2012).

However, the key challenge confronting institutions worldwide is precisely to enable digital preservation systems to scale cost efficiently and effectively in times where content production is soaring, but budgets are not always commensurate with the volume of content in need of safeguarding. Recent advances in using paradigms such as MapReduce (Dean and Ghemawat, 2004) to apply distributed data-intensive computing techniques to the content processing tasks that arise in repositories show a promising step forward for those aspects that are inherently automated in nature. But, ultimately, for a preservation system to be truly scalable as a whole, each process and component involved needs to provide scalability, including business intelligence and decision-making. Here, the decision points where responsible stakeholders set directions and solve the tradeoff conflicts that inevitably arise need to be isolated and well-supported.

Planning and monitoring as key functions in preservation systems have received considerable attention in recent years. The preservation planning tool Plato has shown how trustworthy decisions can be achieved (Becker *et al.*, 2009). Its application in operational practice has advanced the community's understanding of the key decision factors that need to be considered (Becker and Rauber, 2011a), and case studies have provided estimates of the effort required to create a preservation plan (Kulovits *et al.*, 2013a). Finally, the systematic quantitative assessment of preservation cases can provide a roadmap for automation efforts by prioritizing those aspects that occur most frequently and have the strongest impact (Becker *et al.*, 2013).

However, creating a preservation plan in many cases still is a complex and effort-intensive task, as many of the required activities have to be carried out manually. It is difficult for organizations to share their experience in a way that can be actively monitored by automated agents and effectively used by others on any scale. Automated monitoring, in most cases, is restricted to the state of internal storage and processing systems, with little linking to preservation goals and strategies and scarce support for continuously monitoring how the activities in a repository and its overall state match the evolving environment. Finally, integrating whatever solution components an organization chooses to adopt with the existing technical and social environment is difficult, and integration of this context with strategies and operations is challenging (Becker and Rauber, 2011c).

This article presents an innovative architecture for scalable decision-making and control in preservation environments, implemented and evaluated in the real world. The SCAPE Planning and Watch suite builds on the preservation planning tool Plato and is designed to address the challenges outlined above. It makes preservation planning and monitoring context-aware through a semantic representation of key organizational factors, and integrates this with a sophisticated new business intelligence tool that collects and reasons upon preservation-relevant information. Integration with

repositories and external information sources provide powerful preservation capabilities that can be freely integrated with virtually any repository or content management system. The new system provides substantial capabilities for large-scale risk diagnosis and semi-automated, scalable decision-making and control of preservation functions in repositories. Well-defined interfaces allow a flexible integration with diverse institutional environments. The free and open nature of the tool suite further encourages global take-up in the repository communities.

The article synthesizes and extends a series of articles reporting on partial solution blocks to this overarching challenge (Becker and Rauber, 2011c; Becker *et al.*, 2012; Faria *et al.*, 2012, 2013; Petrov and Becker, 2012; Kulovits *et al.*, 2013a, 2013b; Kraxner *et al.*, 2013). Besides pulling together the compound vision and value proposition of the integrated systems and providing additional insight into the design goals and objectives, we conduct an extensive evaluation based on a controlled case study, outline the interfaces of the ecosystem to enable integration with arbitrary repository environments, discuss implications for systems integration and assess the improvement of the provided system over the state of the art in terms of efficiency, effectiveness and trustworthiness.

The article is structured as follows. The next section illustrates the state of the art in preservation planning and monitoring and highlights the key challenges faced by repositories to provide scalable decision-making and monitoring facilities. Section 3 presents the key goals of our work and the main conceptual solution components that are developed to address the identified challenge. We outline common vocabularies and those aspects of design pertaining to future extensions of the resulting preservation ecosystem. Section 4 presents the suite of automated tools designed and developed to improve decision support and control in real-world settings. Becker *et al.* (2015) will discuss the improvements of the presented work and identified limitations based on a quantitative and qualitative evaluation of advancing the state of art including a case study with a national library.

## 2. Digital preservation: background and challenges

### *2.1 Digital preservation and repositories*

The existing support for active, continued preservation in the context of digital repositories can be divided into several broad areas: repository software, tools for identifying and characterizing digital objects, tools for preservation actions (migration and emulation) and quality assurance and systems for preservation monitoring and preservation planning.

The main goal of repository software is to provide capabilities of storing any type of content and accompanying metadata, managing those data and providing search and retrieval options to the user community. Rapidly growing demands for storing digital material are shifting development trends toward more scalable solutions aiming to provide a fully scalable system capable of managing millions of objects. Even though it is a very important aspect, scalability is only one of the dimensions which need to be effectively addressed in digital repositories.

Many repositories are looking for ways to endow their systems with the capabilities to ensure continued access to digital content beyond the original creation contexts. Replacing an entire existing repository system is rarely a preferred option. It may not be affordable or sustainable but also will often not solve the problem, as the organizational

side of the problem needs to be addressed, as well and preservation is inherently of continuous nature. In a recent survey, a majority of organizations looking for preservation solutions stated that they are looking for mix-and-match solution components that can be flexibly integrated and support a stepwise evolutionary approach to improving their systems and capabilities. There is a strong preference in the community for open-source components with well-defined interfaces, fitting the approach preferred by most organizations (Sinclair *et al.*, 2009).

The importance of file properties and file formats in digital preservation resulted in broad research and the development of tools and methods for file analysis and diagnosis. According to their functionality, such tools can be divided into three categories: identification (identifying the format of a file), validation (checking the file conformance with the format specification) and characterization (extracting object properties) (Abrams, 2004). Probably the best-known identification tool is the Unix command file. Further examples are the National Archives' DROID[1] tool and its siblings such as fido[2].

Characterization tools differ in performance characteristics as well in feature and format coverage. Some of the most used and cited examples are the JSTOR/Harvard Object Validation Environment JHove[3] and its successor JHove2, and the eXtensible Characterization Languages (XCL) (Thaller, 2009). Apache Tika[4] combines fast performance with a coverage that extends beyond mere identification to cover extraction of various object features. Acknowledging that a single tool cannot cover all formats and the entire feature space, the File Information Tool Set (FITS)[5] combines other identification and characterization tools such as DROID and JHove and harmonizes their output to be able to cover different file formats and have a richer feature space as a result.

Some efforts have been reported on aggregating and analyzing such file statistics for preservation purposes. Most approaches and tools demonstrated thus far are often focused solely on format identification (Knijff and Wilson, 2011; Hutchins, 2012). Brody *et al.* (2008) describes PRONOM-ROAR, an aggregation of format identification distributions across repositories.

Today, automatic characterization and metadata extraction is supported by numerous tools. The SCAPE project is packaging such components into discoverable workflows and, by that, providing a possibility to automatically discover, install and run those tools[6]. This encompasses migration actions, characterization components and quality assurance. The latter refers to the ability to deliver accurate measures about the quality of digital objects, in particular to ensure that preservation actions have not damaged the authenticity of an object's performance (Heslop *et al.*, 2002). The SCAPE project is addressing that question by providing a number of tools for image, audio video and web quality assurance (Pehlivan, 2013). Furthermore, it is packaging those components into discoverable workflows and by that providing the facilities to discover and invoke these tools.

The metadata collected by different identification and characterization tools will help with managing the objects more efficiently and effectively. The real power of such data becomes especially visible when visual aggregation and analysis methods are used. Jackson (2012) presented a longitudinal analysis of the format identification over time in the UK web. Even though this study was only using identification data, the resulting statistics of the evolution of different formats over time can yield significant insights

into format usage trends and obsolescence. There is a clear need of a broad systematic approach to large-scale feature-rich content analysis to support business intelligence methods in extracting important knowledge about the content and formats stored in a repository.

This is also a key enabler for successful preservation planning, one of the six functional entities specified in the OAIS model (CCSDS, 2002). Its goals are to provide functions for monitoring internal and external environments and to provide preservation plans and recommendations which will ensure information accessibility over a longer period of time. Viewed as an organizational capability, the two main sub capabilities are Operational Preservation Planning and Monitoring (Antunes *et al.*, 2011).

The planning tool Plato (Becker *et al.*, 2009) is up to now the best known implementation of an operational preservation planning method. It provides a 14-step workflow that guides a preservation planner in making decisions about the actions performed on a digital content. The result of a planning process is a trustworthy and well-documented recommendation which identifies the optimal action from a defined set of alternatives according to specified objectives and requirements. These plans are not strategic plans guiding the organization's processes and activities, but operational specifications for particular actions to be carried out with exact directives on how they shall be carried out. Even though Plato offers a great deal of automation, some steps in the workflow require significant manual work. Kulovits *et al.* (2009) showed that in 2009, a typical use case involved several people for about a week, including a planning expert to coach them.

Preservation monitoring shows a comparable gap of automated tool support. Current activities usually result in technical reports, as Lawrence *et al.* (2000), DigiCULT[7] and Digital Preservation Coalition periodic reports[8] or file format and tool registries (PRONOM[9], Global Digital Format Registry[10], Unified Digital Format Registry[11], the P2 registry[12], and others). Technical reports function on a principle of periodically publishing documents about available formats and tools. They are meant for human reading and support no automation. Registries such as PRONOM are shared and potentially large, but very often do not provide in-depth information. They have difficulties in ensuring enough community contributions and, where those contributions exist, they are often sparse and dispersed in different registries. Moderation of such contributions through a closed, centralized system has proven notoriously difficult, which has led to increasing calls for a more open ecology of information sources (Becker and Rauber, 2011a; Pennock *et al.*, 2012)[13].

An early attempt to demonstrate automation in preservation monitoring was PANIC (Hunter and Choudhury, 2006). The goal was to provide a system which will periodically combine the metadata from repositories with the information captured from software and format registries to detect potential preservation risks and provide recommendations for possible solutions. The initiative to develop an Automatic Obsolescence Notification Service (Pearson, 2007) aimed at providing a service that would automatically monitor the status of file formats in a digital repository against format risks collected in external registries. Unfortunately, the dependency on external format registries to provide information for a wider range of file formats was a limitation for automated obsolescence notification system (AONS), which caused it to monitor only a limited amount of information.

Preservation actions need to be carefully chosen and deployed to ensure they, in fact, address real issues and provide effective and efficient solutions. There is an increasing awareness and understanding of the interplay of preservation goals and strategies, tools and systems and digital preservation policies. Policies are often thought to provide the context of the activities and processes an organization executes to achieve its goals, and hence the context for the preservation planning and monitoring processes described. Yet, in Digital Preservation, the term "policies" is used ambiguously; often, it is associated with mission statements and high-level strategic documents (Becker and Rauber, 2011c). Representing these in formal models would lead to only limited benefit for systems automation and scalability, as they are intended for humans. On the other hand, models exist for general machine-level policies and business policies. However, a deep domain understanding is required to bring clarity into the different levels and dimensions at hand. This should be based on an analysis of the relevant drivers and constraints of preservation. A driver in this sense is an "external or internal condition that motivates the organization to define its goals" (Object Management Group, 2010), while a constraint is an "external factor that prevents an organization from pursuing particular approaches to meet its goals" (Object Management Group, 2010).

Common examples for preservation policies are on the level of statements in TRAC (OCLC and CRL, 2007), ISO 16363 (International Standards Organisation, 2010) or statements in Beagrie *et al.* (2008). These are well known, but their impact is not always well understood, and operations based on these can be quite complex to implement. Moreover, there is no recognized model for formalizing preservation policies in a standard way. Providing such context for preservation planning, monitoring and operations, however, is key to successful preservation. So far, context has been provided implicitly as part of decision-making, adding a burden on decision makers and threatening the quality and transparency of planning and actions.

These policies correspond to what the object management group (OMG) standards call "business policies". The OMG has been active in modeling and standardizing this concept for many years and produced, in particular, two valuable standards: the Business Motivation Model (Object Management Group, 2010) and the Semantics of Business Vocabulary and Business Rules (SBVR) (Object Management Group, 2008). According to these, policies are non-enforceable elements of governance that guide, shape and control the strategies and tactics of an organization. An element of governance is an "element of guidance that is concerned with directly controlling, influencing, or regulating the actions of an enterprise and the people in it". Enforceable means that "violations of the element of governance can be detected without the need for additional interpretation of the element of governance" (Object Management Group, 2008).

There are various levels of policy statements required in a digital preservation environment. While the digital preservation (DP) community has specified criteria catalogs for trustworthy preservation systems, these fail to separate concerns and distinguish between strategic goals, operational objectives and constraints, and internal process metrics. The relationship between these is often vague. Compliance monitoring in operational preservation systems is restricted to generic operations and does not align well with the business objectives of providing understandable and authentic access to information artifacts. The lack of clarity, separation of concerns, formalism and standardization in regulations for DP compliance means that operationalizing such

compliance catalogs is very difficult, and verification of compliance is manual and either limited to abstract high-level checks on a system's design or inherently subjective.

*2.2 On trust and scalability*

Preservation planning methods and tools such as Plato have evolved considerably from their origins (Strodl *et al.*, 2006). It is worth recalling here the two fundamental dimensions along which such evolution could take place – dimensions set by the decision space in which these methods are designed to operate. The key requirements, not at all compatible at first sight, are trust and scalability.

Trust as a requirement hardly disputed mandates organizations to strive for transparency, accountability, and traceability, traits evidently recommended by standards such as the Repository Audit and Certification checklist (International Standards Organisation, 2010). Achieving trust requires a carefully designed environment that promotes transparency of decisions, ensures full traceability of decision chains and supports full accountability. Scalability, on the other hand, is mandated by the sheer volumes of content pouring into repositories in the real world, and calls for automation, reduced interaction, simplified decisions and the removal of human interaction wherever possible.

Scalability calls for automated actions applied in standardized ways with minimized human intervention. Trust, on the other hand, mandates that any automated action is fully validated prior to execution, providing an assessment trail against the objectives specified by the organization which is supported by real-world evidence.

Preservation planning methods and tools such as Plato come a long way along the path of trustworthy decision-making, but, by the very nature of the task, have difficulties in making progress on the dimension of scalability. Considerable effort is commonly required for taking trustworthy decisions, as well as for creating, structuring and analyzing the underlying information that is the input for the decision-making process. Until now, this often means that organizations fail to move from hardly trustworthy *ad hoc* decision-making to fully trustworthy, well-documented preservation planning (Becker and Rauber, 2011c; Kulovits *et al.*, 2013a).

*2.3 Challenges and goals*

The preservation of digital content requires that continuous monitoring, planning and the execution of corrective actions work together toward keeping the content authentic and understandable for the user community and compatible with the external environment and restrictions. However, many institutions carry out these processes in a manual and *ad hoc* way, completely detached from the content life cycle and without well-defined points of interoperability. This limits the ability to integrate and scale preservation processes to cope with the escalating growth of content volume and heterogeneity, and it undermines the capacity of institutions to provide continued access to digital content and preserve its authenticity.

We observe that there are a number of gaps in the means currently available to institutions:

- Business intelligence mechanisms are missing that address the specific needs of preservation over time and enable organizations to monitor the compliance of their activities to goals and objectives, as well as risks and opportunities. Similarly, organizations lack the scalable tools to create feature-rich profiles of their holdings

to support this monitoring and analysis process. There are no accepted ways to address the need for continuous awareness of a multitude of key factors prone to change, including user communities, available software technology, costs and risks, to provide a unified view on the alignment of an organization's operations to goals and needs. While the community is eager to share the burden and promote collaboration, it is notoriously difficult for organizations to effectively do so.

- Knowledge sharing and discovery at scale is not widely practiced, as there is no common language, no effective model and little clarity as to what exactly can and should be shared and how. Hence, sharing is practiced on an *ad hoc* and peer-to-peer basis, with little scalable value for the wider community.
- Decision-making efficiency needs to be improved without sacrificing transparency and trustworthiness. This requires not only more efficient mechanisms built into decision-making tools but also a more explicit awareness of an organization's context.
- Preservation policies are a key factor to achieve this and have been notoriously difficult to pin down. In this context, it is important to understand policies as "elements of guidance that shape, guide and control" (Object Management Group, 2008) the activities of an organization, so that the core aspects can be formalized and understood by decision support and business intelligence systems.
- Systems integration, finally, is chronically difficult and only successful where modular components with clearly defined purpose and well-specified interfaces are provided in the place of monolithic, custom-built solutions.

It becomes clear that establishing such capabilities cannot simply be solved by introducing a new software tool, but requires careful consideration of the socio-technical dimensions of the design problem. Designing a set of means to address these issues requires a solid understanding of socio-technical environments and a flexible suite of methods and tools that can be customized, integrated and deployed in a real-world context to address the issues pertaining to a particular situation.

The following section will discuss each of the design challenges in turn and derive a set of overarching design goals. Based on these, we will present the main concepts and solution components that form the main contribution of our work and discuss how they can be used in isolation or conjunction to improve the state of art in scalable decision-making and control.

## 3. Scalable, context-aware preservation planning and watch

### *3.1 Overview*

Based on the observations outlined above, this section derives a number of design goals to be addressed to enable scalable decision-making and control for information longevity, while further advancing the progress made on the path of trust, in a form that can make substantial real-world impact for a variety of organizations. Based on a new perspective that emphasizes the continuous nature of preservation, we describe an architectural design for trustworthy and scalable preservation planning and watch. Section 4 discusses the implementation of the architecture in the SCAPE Planning and Watch suite.

*3.2 Design goals*

Systematic analysis of digital object sets is a critical step toward preservation operations and a fundamental enabler for successful preservation planning. Without a full understanding of the properties and peculiarities of the content at hand, informed decisions and effective actions cannot be taken. While large-scale format identification has been in focus for a while and tools for in-depth feature extraction exist, little work has been shown that combines in-depth analysis and large-scale aggregation into content profiles that are rich in information content and large in size.

*G1*. Provide a scalable mechanism to create and monitor large and rich content profiles.

For successful preservation operations, a preservation system needs to be capable of monitoring compliance of preservation operations to specifications, alignment of these operations with the organization's preservation objectives and associated risks and opportunities that arise over time. Achieving such a business intelligence capability for preservation requires linking a number of diverse information sources and specifying complex conditions. Doing this automatically in an integrated system should yield tremendous benefits in scalability and enable sharing of preservation information, in particular risks and opportunities.

*G2*. Enable monitoring of operational compliance, risks and opportunities.

The preservation planning framework and tool Plato provide a well-known and solid approach to create preservation plans. However, a preservation plan in Plato 3 is constructed largely manually, which involves substantial effort. This effort is spent in analyzing and describing the key properties of the content that the plan is created for; identifying, formulating and formalizing requirements; discovering and evaluating applicable actions; taking a decision on the recommended steps and activities; and initiating deployment and execution of the preservation plan. When automating such steps, trustworthiness must not be sacrificed for efficiency. Still, the efficiency of planning needs to be improved to the point that creating and revising operational plans becomes an affordable and normal, routine part of organizations responsible for safeguarding content and is understood well enough so that it can potentially be offered as a service.

*G3*. Improve efficiency of trustworthy preservation planning.

For decision support and monitoring systems to be truly useful, they need to be aware of the context in which they are operating. That includes an awareness of the organizational setting and the state of the repository so that they can assess risks and identify issues that need intervention, but it extends to an awareness of the world outside the repository to ensure these systems can provide this assessment also with respect to the larger context in which the repository operates. So far, it has been very difficult to make the organizational context known to the systems in a way that enables them to act upon it. The planning tool Plato 3, for example, requires the decision-makers to model their goals and objectives in a tree structure; but it is not directly aware of other organizations' goals and objectives. Similarly, the context awareness of systems such as PANIC is very limited.

Most importantly, hence, preservation systems need to be endowed with an awareness of the context in which they shall keep content alive. This includes the organizational goals and objectives, constraints, and directives that shape and control the preservation operations of a repository. Such an awareness of the context requires a formalized representation of organizational constraints and objectives and a controlled vocabulary for representing the key entities of the domain. Given the evolutionary nature of the world in which preservation has to operate, such a vocabulary needs to be permanent, modular and extensible.

*G4*. Make the systems aware of their context.

Preservation planning focuses on the creation of preservation plans; Preservation Watch focuses on gathering and analyzing information; operations focus on actual processing of data and metadata. These methods and tools will, in general, be deployed in conjunction with a repository environment. This requires open interfaces and demonstrated integration patterns to be useful in practice. We hence need a system architecture that is based on open interfaces, well-understood components and processes, open data and standard vocabularies but also able to be mixed and matched, extended and supportive of evolution over time.

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, with the goal to connect and enable rather than impose and restrict. The preservation community is painfully aware how important sustainable evolution is for their systems, as emphasized by a recent discussion[14]. Correspondingly, the ecosystem in question should be built with sustainability in mind.

*G5*. Design for loosely coupled preservation ecosystems.

Clearly, addressing the sum of these goals requires a view on the preservation environment that focuses on the continuous, evolving nature of information longevity as a sustained capability rather than a one-time activity. The following section presents such a view, focusing on the preservation lifecycle and its key components.

### *3.3 The preservation lifecycle*

Figure 1 shows a view on the key elements in a preservation environment that relates the key processes required to successfully sustain content over time to each other. The preservation lifecycle naturally starts with the repository and its environment and evolves around the continuous alignment of preservation activities to the policies and goals of the organization. The Repository is an instance of a system which contains the digital content and may comprise processes such as ingest, access, storage and

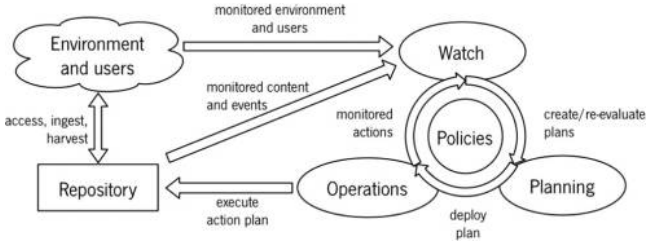


**Figure 1.**
Digital preservation lifecycle

metadata management. The Repository may be as simple as a shared folder with files that represent the content, or as complex as dedicated systems such as DSpace[15], Eprints[16] and RODA[17].

The Repository refers not only to the generic software system but also its instantiation within an institution, related to an institutional purpose guided and constrained by policies that define objectives and restrictions for its content and procedures. In the context of this article, the preservation policies drive how the Repository must align to its context, environment and users, guiding digital preservation processes such as Watch, Planning and Operations.

The alignment of the content and the activities of a repository to its context, environment and users is constantly monitored by Watch to detect preservation risks that may threaten the continuous and authentic access to the content. This starts by obtaining an understanding of what content the repository holds and what the specific characteristics of this content are. This process is supported by the characterization of content and allows a content owner to be aware of volumes, characteristics, format distributions and specific peculiarities such as digital rights management issues and complex content elements. The characterization process feeds the aggregated set of key characteristics of the monitored content, i.e. the content profile, into the Watch process. This is depicted as the "monitored content" in Figure 1. Repository events such as ingest or download of content are monitored by the Watch process, as they can be useful for tracking producer and consumer trends and uncover preservation risks.

The Watch process cross-relates the information that comes from internal content characterization and repository events with the institutional policies and the external information about the technological, economic, social and political environment of the repository, allowing for the identification of preservation risks and opportunities. For example, checking the conformance of content with the owner's expectations or policies, identifying format or technological obsolescence in content or comparing the content profile with other repositories can reveal possible preservation risks but also opportunities for actions and possibilities to improve efficiency or effectiveness.

These possible risks and opportunities should be analyzed by Planning to devise a suitable response. The Planning process carefully examines the risks or opportunities, considering the institution's goals, objectives and constraints. It evaluates and compares possible alternatives and produces an action plan that defines which operations should be implemented and which service levels have been agreed on, and documents the reasoning that supports this decision (Becker *et al.*, 2009).

This action plan is deployed to the Operations process that orchestrates the execution of the necessary actions on the repository content, if necessary in large-scale distributed fashion, and integrates the results back to the repository. These operations can include characterization, quality assurance, migration and emulation, metadata and reporting.

The Operations process should provide information about executed actions such as quality assurance measurements to the Watch process to be sure that the results conform to the expectations set out in the action plan. All the conditions about internal and external information considered as a decision factor by Planning should be continuously monitored so that the organization knows where active plans remain aligned and valid over time. Once a condition is detected that may invalidate a plan, Planning should be called upon to re-evaluate the plan.

This perspective on digital preservation as a set of processes or capabilities that interact with each other to achieve the digital preservation objectives has evolved considerably over the past decade. From the oft-cited standard model in the domain, the OAIS (CCSDS, 2002), which emphasizes a functional decomposition of elements in an archive, the perspective evolved to the capability-based view of the SHAMAN Reference Architecture (Antunes *et al.*, 2011), which based the model strongly in Enterprise Architecture foundations and thus integrated the domain knowledge of preservation with a holistic view of the organizational dimensions. However, neither presents a specific view on how these processes can align with each other in practice, allowing the flow of information from one process to the next. The streamlined view illustrated in Figure 1 forms a life cycle that ensures digital content on repositories is continuously adapted to the environment, the target users and institutional policies.

Considering the above, it becomes clear that optimization of efficiency (whether of performance and cost or effort) must not only occur within each process, and not only consider scalable processing of data but also at the integration points between each of the processes and in the decision functions themselves, so the whole preservation life cycle becomes efficient and sustainable. Finally, many of the activities in these processes require sophisticated tool support to be applicable in a real-world environment.

### *3.4 An architecture for loosely coupled preservation systems*

Achieving a full preservation life cycle requires a set of components that implement the digital preservation processes and interoperate with each other in an open and scalable architecture. Figure 2 shows the set of components that are required to support and partially automate the processes necessary to sustain the preservation lifecycle. These need to be designed to be modular and flexible, have clearly distinguished functionalities and fit the technical specifications of the institution context.

The Content profiler has the function of aggregating and analyzing content characteristics and producing a well-specified content profile that provides a meaningful and useful summary of the relevant aspects of the content. This component has to cope with large amounts of data in the content and support the watch and planning components by summarizing the important aspects to a content profile, exposed via the Content profile interface.

The Watch component has the function of collecting this and other aspects to provide the business intelligence functionalities necessary for monitoring and alignment. By gathering information from diverse sources relevant for preservation, it enables the organization to monitor compliance, risks and opportunities, based on monitoring conditions that can be specified in the corresponding interface. For example, it provides the means to answer questions such as "How many organizations have content in format

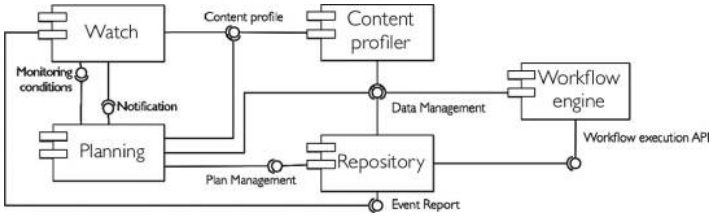


**Figure 2.**
Overall architecture of scalable planning and watch

X?" or "Which software components have been tested successfully in analyzing objects in format Y?" (Becker *et al.*, 2012). The component should be able to raise events when specified conditions are met. Interested clients provide a Notification interface to receive such events.

The Planning component is hence informed about conditions that require mitigation. Its key function is to support the creation, revision and deployment of trustworthy and actionable preservation plans. To achieve this, it needs to retrieve the complete content profile, potentially access data from the repository for sample sets to experiment with for evaluation purposes and will use the Plan Management interface to initiate the execution of the actions specified in the plan.

A Repository should be able to integrate this preservation life cycle architecture by implementing a set of interfaces[18]: Data Management enables basic operations on the data held by the repository to ensure controlled access; Event Reporting ensures that Watch can be informed about the status of operations and repository activities (Becker *et al.*, 2012); and Plan Management provides the facilities to create and update preservation plans and initiate their deployment. To coordinate these sometimes complex activities and processes that are executed by operations, a Workflow engine can be used to execute the preservation action plan, i.e. the set of all actions and quality assurance tasks that compose the execution of a plan on the content. The Data Management interface can also be used to merge the results of executing the action plan back to the repository.

Interoperability between components is achieved via well-defined interfaces that allow the decoupling from the specific implementation of each component and also allow the reuse, replacement and easier maintenance of each of the components. The interfaces are open to allow easy support of different component implementations, in particular different repository implementations. A key goal of such open interfaces is to enable continuous growth of systems by community participation. However, standardization in this area needs to go one step further and support semantic interoperability of the components. Components need to be aware of the context they are operating in, and this context need to be well communicated and mapped between each of components. Information exchanged between these components needs to be opened up to the community to build synergies, enable knowledge discovery and move from static to dynamically growing information sources. The next section will describe the mechanisms designed to support this.

### *3.5 Policies as basis for preservation management*

When endowing components of a context-aware planning and watch system as envisioned here with an awareness of organizational context to create "policy-driven planning and watch", the idea cannot be that entirely non-enforceable elements drive something automatically, as the result would be random behavior. Instead, the idea is to relate non-enforceable high-level policies to practicable policies that are machine-understandable, but usually not specific enough to directly drive operations. The control of operations then is the responsibility of preservation planning, which creates enforceable preservation plans based on practicable policies.

Corresponding to the observation that policies "guide, shape and control" the activities of an organization (Object Management Group, 2010), we distinguish between the following levels.

Guidance policies are non-enforceable governance statements that reside on the strategic (governance) level and often relate several high-level aspects of governance to each other. For example, they express value propositions to key stakeholders, commit to high-level functional strategies, define key performance indicators to be met or express a commitment to comply with a regulatory standard. These policies are expressed in natural language and need to be interpreted by human decision-makers. Automated reasoning on these is not generally feasible. The aspects to be included in such policy statements can be standardized and identified, but the statements can often not feasibly be expressed as machine language to a meaningful extent. In the preservation domain, typical examples can be seen in current regulatory compliance statements (International Standards Organisation, 2010) but also in preservation business policies (Beagrie *et al.*, 2008):

> Control policies, on the other hand, are practicable elements of governance that relate to clearly identified entities in a specified domain model […] [and] constitute quantified, precise statements of facts, constraints, objectives, directives or rules about these entities and their properties. (Kulovits *et al.*, 2013b).

Practicable means that a statement 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 behavior is acceptable or not, or how something is understood (Object Management Group, 2008).

Such policies can be fully represented in a machine-understandable model, but are often not directly actionable in the sense that it does not make sense to directly enforce them in isolation. The exact enactment will depend on the context and the relation of multiple control policies. For example, multiple control policies may be defined in isolation and contradict each other. The resolution of this contradiction in the decision-making process (preservation planning) leads to a specified set of rules in the plan. This rule set is then actionable and enforceable. Some control policies will, on the other hand, be, in principle, enforceable. For example, constraints about data formats to be produced by conversion processes can be automatically enforced in a straightforward way.

Control policies are practicable in the sense of the SBVR, but generally have to be specified by human decision-makers in policy specification processes that refer to the guidance policies and take into account the drivers and constraints of the organization to create control policies. These processes can be standardized to a degree similar to standard business processes. The typical inputs and outputs, as well as the stakeholders responsible, accountable, consulted and informed, can be specified. Yet, it should not be prescribed to a particular organization in which way these policies have to be managed.

By applying these levels, non-enforceable high-level policies can be related to practicable policies that are machine-understandable, but usually not specific enough to directly drive operations. The control of operations then is the responsibility of preservation planning, which creates enforceable preservation plans based on practicable policies. These preservation plans correspond to business rules. We note that if control policies are specified in a formal model, it should be possible to check instances of that model against formal constraints.

An institutions' specific policies should thus be specified following a well-defined vocabulary. To make such policies meaningful, a core set of domain elements has to be

identified and named so that the properties of these concepts can be referred to, represented and measured. This is illustrated in Figures 3 and 4.

Ultimately, a preservation case arises, in analogy to a business case, from the identified value of a set of digital artifacts for a specified, more or less well-defined, set of users, called the user community. A preservation case hence concerns identified content and identified users and specifies the goals that should be achieved by preservation. Practically, the level of detail known in each specific instance about the users' goals and means will vary greatly, but where there is no identified potential value in preserving a set of digital artifacts, it will likely be discarded. The scope of the preservation case thus corresponds closely to the statements of "preservation intent" discussed by Webb *et al.* (2013).

To successfully preserve objects for a set of users, i.e. address a preservation case, goals will be identified and made explicit by specifying objectives. These are more explicit than a general preservation intent and represent the general goals for effective and efficient continued access to the intellectual content of the digital artifacts in precise statements: The objectives specify desirable properties of the objects with regards to authenticity, formats and other aspects of representation (such as compression, codecs, or encryption); desired properties of the formats in which such objects shall be represented; desired properties of the preservation operations carried out to achieve

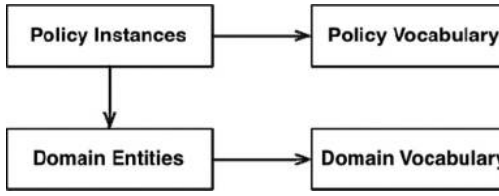


**Figure 3.**
Digital preservation policies need a well-defined domain model (Kulovits *et al.*, 2013b)

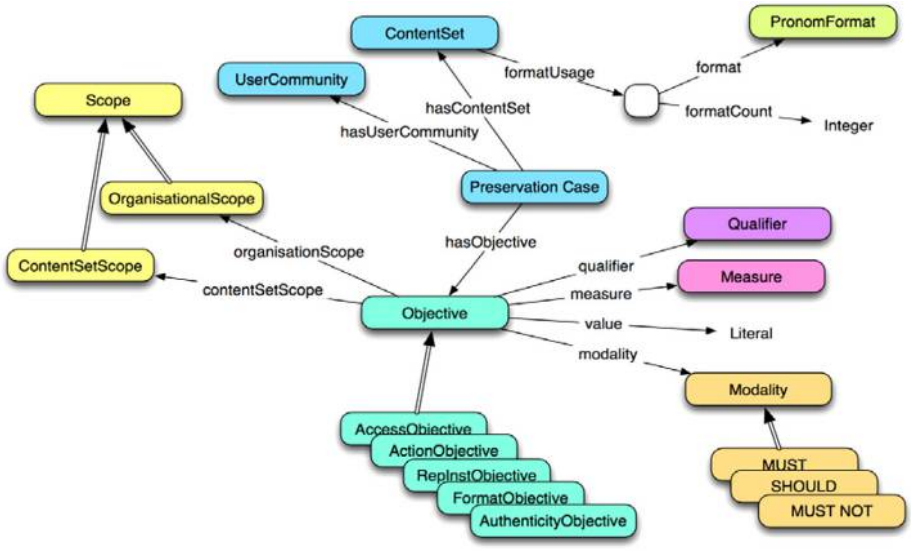


**Figure 4.**
The core set of elements in the vocabulary (Kulovits *et al.*, 2013b)

preservation goals, in particular preservation actions to be applied (such as a preferred strategy of migration or an upper limit on costs); and access goals derived from knowledge about the user community.

It can be seen that the core focus of this model is on continued accessibility and understandability on a logical level, emphasizing the continued alignment that is at the heart of preservation rather than the mere conservation of the bitstreams themselves, which is seen as a necessary precondition to be addressed independently. It is only through access (of whatever form) that preservation results in value; and it is only through a continued process that such understandability and access can be assured. Making the aspects that should be aligned explicit and measurable is the first step toward intelligent detection and reaction. Correspondingly, the core set of control policy elements is shown in Figure 4, taken from (Kulovits *et al.*, 2013b) which describes the controlled vocabularies in more detail.

The ontology of core control policies and the ontology of the domain elements referenced in these statements are permanently accessible on: http://purl.org/DP/control-policy and http://purl.org/DP/quality. While a detailed discussion of these elements is out of the scope of this article, the next sections will show how it enables the components of the implemented software suite to sustain an awareness of an organization's objectives and constraints and monitor the alignment of operations to the preservation goals.

### *3.6 A preservation ecosystem*

The standardization of the policy vocabulary and the domain model allows us to envision a digital preservation ecosystem that brings together the Organization, the Community environment, the Solution components and the Decision support and control tools that make up the loosely coupled system presented in Section 3.4. The vocabularies allow all of these entities to share a common language and be able to interoperate. Figure 5 illustrates how the digital preservation vocabulary connects the ecosystem domains:

- *Organization*: An organization has digital content and internal goals regarding its purpose and delivery which influence decisions on how to curate, preserve and reuse the content over time. People on behalf of the organization manage information systems and define policies that guide and constrain the selection and design of operations to be executed to preserve the content. The formulation of policy instances for the organization can follow a vocabulary that is widely understood by the other parts of the ecosystem.
- *Community environment*: Other organizations with particular concerns, not necessarily to preserve content, develop and populate systems that support various aspects of preservation directly or indirectly. 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[19] or public open source software repositories such as github[20].
- Solution components comprise the services and tools, platforms and infrastructure components that support the necessary operations to address organizations'

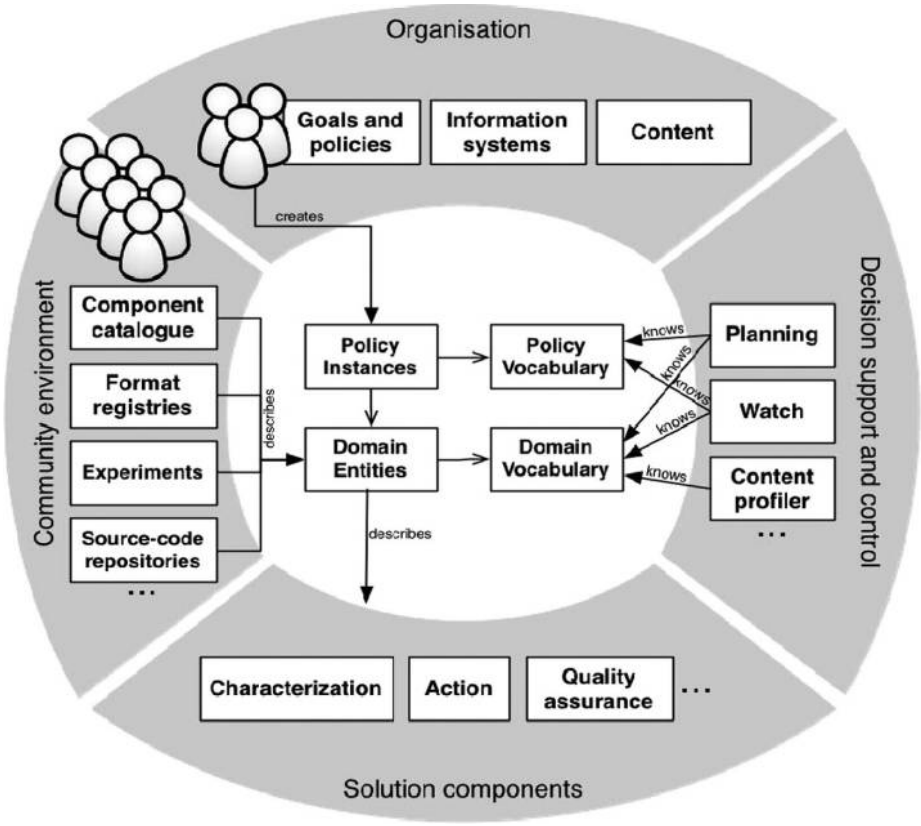


**Figure 5.**
A common language connects the domains of the preservation ecosystem

needs. These components relate to the pieces that need to be put together to allow addressing the organization objectives in specific cases in cost-efficient ways. These tools must be selected considering the organization's policies (criteria and constraints) that define requirements for a solution. The main types of such solution components include software tools for file format identification, feature extraction, validation, migration, emulation and quality assurance. Solution components in this domain are generally developed, maintained and distributed by commercial or noncommercial solution providers trying to meet market needs. Many of them are, in fact, created by members of the preservation community[21].

- Decision support and control, finally, brings together those methods and systems that support the organization in choosing from the solution domain those elements that fit their policies and goals best, ensure most effectively that the content remains usable for the community and support the organization in the continued task of monitoring.

Each software system requires information about certain domain entities. For example, content profiling needs to describe objects it analyzes, and preservation tools need to report measures. Planning needs to discover preservation actions, evaluate actions, and

describe plans. Watch needs to collect measures on all these entities, 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.

The next section will outline how this ecosystem has been implemented and instantiated and show the preservation life cycle in action within the ecosystem. We will outline the solution architecture and discuss the specific components of the architecture in turn, and then return to the preservation lifecycle and how the ecosystem increasingly supports scalable, context-aware preservation planning and monitoring and its integration into repository environments and the community.

## 4. The SCAPE Planning and Watch suite

### *4.1 Overall solution architecture*

The architecture outlined above has been implemented by a publicly available set of components and API specifications that can be freely integrated with any repository system. The suite of components aims to provide the tool support necessary to enable organizations to advance from largely isolated, *ad hoc* decisions and actions reacting to preservation incidents to well-supported and well-documented, yet scalable and efficient preservation management. The following section describes each of the key building blocks of this tool suite, focusing on the core design goals and features and pointing to references for further in-depth information.

Note that the design is not limited to large-scale environments, but understands scalability as a general flexibility with a focus on efficiency and automation. This is relevant in two ways: first, the tools do not require large-scale infrastructure, but are able to leverage it when present. Second, providing a loosely-coupled set of modular components enables organizations to adopt the suite using an incremental approach, without large upfront investments.

Figure 6 depicts the SCAPE software components supporting the preservation life cycle and implementing the components and interfaces described above. The next sections will describe each of these components in turn.

### *4.2 C3PO: scalable content analysis*

Recent advancements in tool development for file analysis resulted in a number of tools covering different functionality such as identification, validation and characterization. A crucial challenge presented by those tools is the variance of coverage in terms of file formats supported and features extracted. The characterization tool FITS addresses the problem of coverage by combining outputs from different identification and

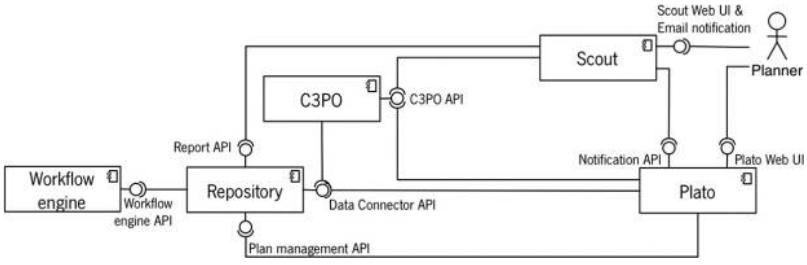


**Figure 6.**
SCAPE software components supporting the preservation lifecycle

characterization tools in one descriptor. This enables a rich characterization of a single file by using only one tool, which will, in fact, run the appropriate identification and characterization tools on the content and normalize the output into a well-defined XML output. While this comes at a performance cost, it is currently the only method that provides reasonable coverage of the feature space, covering both a variety of identification measures such as the PRONOM format ID and mime-types, as well as in-depth feature extraction supported by an array of tools[22].

Using the output of characterization tools such as FITS and Apache Tika, the tool Clever Crafty Content Profiling of Objects (C3PO)[23] enables a detailed content analysis of large-scale collections (Petrov and Becker, 2012). Figure 7 provides a high-level overview of the process, which as a result produces a detailed content profile describing the key distribution characteristics of a set of objects. The process starts with running identification and characterization tools on a set of content. The metadata produced by those tools is collected and stored by C3PO, which currently supports the metadata schemas of FITS and Apache Tika. Support for other characterization tool output formats can easily be added by extending the highly modular architecture, which enables the integration of additional adaptors to support other metadata formats and gathering strategies.

The combination of using multiple metadata extraction tools on the same content will often result in conflicts, a state where two tools provide different values for the same feature. A common example is the file format, when two tools assign different format identifiers to the same file, either because of different interpretation logics or simple because they have a different way of representing the same format. C3PO offers the possibility to add rules which will resolve those conflicts. These rules can range from simple conditions regulating that certain two identifiers represent the same format to complex rules prioritizing certain tools or deriving values based on the presence of other features.

The architecture of C3PO decouples the persistence layer so that a variety of engines can be used. The default database provides strong scalability support by using the open-source highly scalable MongoDB, which supports sharding (Plugge *et al.*, 2010) and map-reduce (Dean and Ghemawat, 2004) natively. This also enables users to

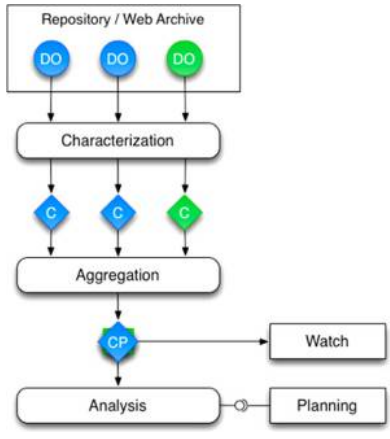


**Figure 7.**
The key steps of content profiling

provide their own analytics on the basis of this platform, similar to the built-in queries that are readily supported through a web user interface.

These standard analytical queries calculate a range of statistics from the size of the collection to the triangular distributions of all numerical features and a histogram of all non-numerical features in the collection. The set of these statistics is the heart of the content profile.

In addition to its processing platform, C3PO offers a simple web interface which allows dynamic exploration of the content. Part of it is shown in Figure 8, displaying a collection with about 42,000 objects and an overall size of approximately 23 GB. Additional diagrams show the distribution of mime-types and formats. The user can create additional diagrams for any feature present in the set to visualize key aspects of the property sets. Advanced filtering techniques enable exploring the content in more detailed fashion. By clicking on a bar representing a certain format in the format distribution diagram, for instance, the user will filter down on the corresponding object set to see details about that part of the collection only. This enables a straightforward drill-down analysis to see, for instance, how many of a set of TIFF files are valid or how many have a certain compression type.

While C3PO can be readily used independently, it integrates with the remaining two components in the Planning and Watch suite, Scout and Plato. The integration with Scout offers the possibility to monitor the feature distributions of any number of collections over time. By creating a historic profile from a collection, its growth and changes in the distributions of key aspects such as formats can be revealed over time. The integration with Plato uses an export of the content profile for the whole or a subset of a collection into a well-defined content profile[24]. This profile identifies and describes the set of objects contained and provides a statistical summary of file format

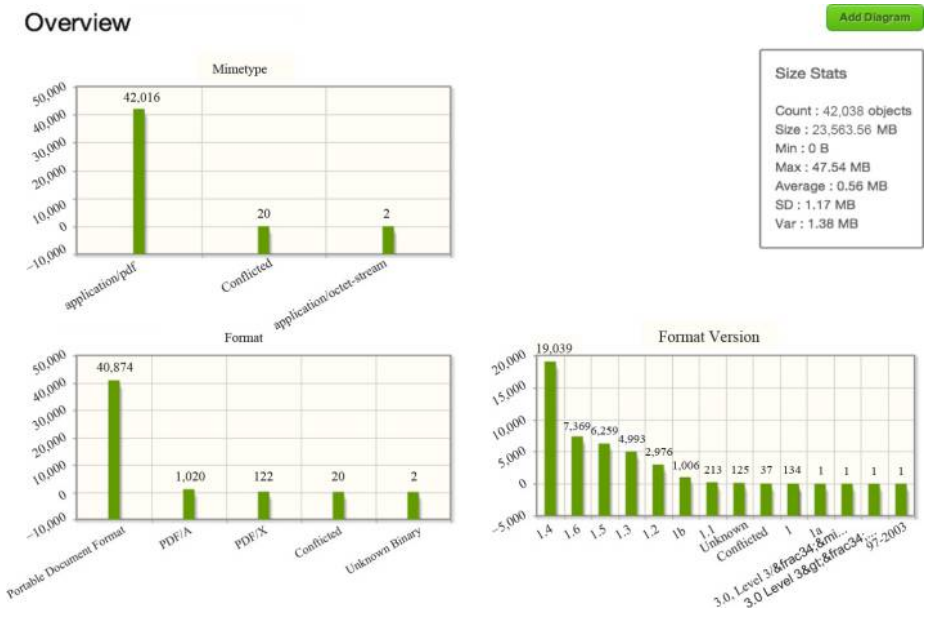


**Figure 8.**
C3PO visualizing a content profile

identification and important features extracted. Plato understands this profile and uses it to obtain statistics about the content set for which a plan is being created.

Finally, this profile can contain a set of objects that are seen as representative for the entire set, to enable controlled experimentation on a realistic subset instead of the entire set of objects. This can yield increased reliability of the sample selection and provide a substantial speedup, as without these heuristics, samples have to be selected by hand, a tedious and error-prone process (Becker and Rauber, 2011c). As with the other modules, the heuristics used to select samples from this multidimensional view on the content set are flexible and configurable, and additional algorithms for sample selection can be added easily.

### *4.3 Scout: scalable monitoring*

Scout[25] is an automated preservation monitoring service which supports the scalable preservation planning process by collecting and analyzing information on the preservation environment, pulling together information from heterogeneous sources and providing coherent unified access to it. It addresses the need to combine an awareness of the internal state of an organization and its systems (internal monitoring) with an awareness of the environment in the widest sense (external monitoring) to enable a continued assessment of the alignment between the two (Faria *et al.*, 2012).

The information is collected by implementing different source adaptors, as illustrated in Figure 9. Scout has no restrictions on the types of data that can be collected. It is built to collect a variety of data from different sources such as format and tool registries, repositories and policies. It already implements source adaptors for the PRONOM registry, content profiles from C3PO, repository events (ingest, access, and migration), policies and other specific adaptors. The combination of content profiles from C3PO with repository events from the Report API provides a complete overview of

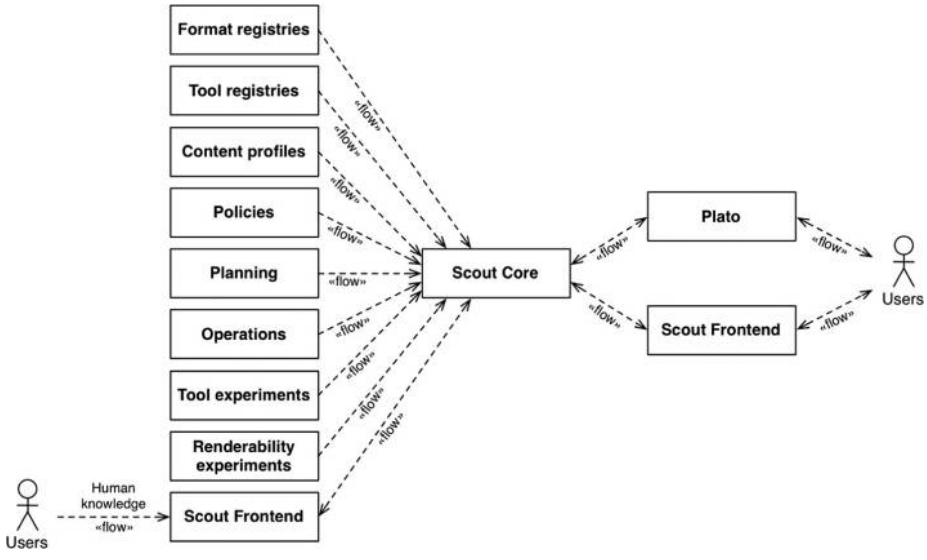


**Figure 9.**
Scout information flows from sources to users

the current content in a repository and shows trends of how the overall set of content is evolving.

Continuous automated rendering experiments[26] can be used to track the ability of viewing environments to display content and verify whether it corresponds to the original performance (Law *et al.*, 2012).

Once information is collected, it is saved in a formally specified and normalized manner to the knowledge base (Faria *et al.*, 2012). Built upon linked data principles, the knowledge base supports reasoning on and analysis of the collected data using standard mechanisms such as SPARQL[27]. Such queries provide the mechanisms for automatic change detection. By registering an interest in a watch condition associated with such a query, the results will be monitored periodically. When the condition is met, a notification is sent to the user. Conditions can cover arbitrary circumstances of relevance in the known domain, ranging from checks on content validity and profile conformance to certain constraints to the question whether any new tools are available to measure a certain property in electronic documents, or whether a Quality Assurance tool that is in use for validating authenticity of converted images is still considered reliable by the community. Upon receiving the notification, the user can initiate additional actions such as preservation planning to address any risks that have surfaced or take advantage of opportunities that have been detected.

Scout has a simple web interface which allows operations such as management, adding new adaptors and triggers and browsing the collected data. This includes dynamically generated visualizations of data over time. By operating over a longer period, Scout is expected to have a valuable collection of historical data. Figure 10 shows an example of evolution of file formats through time. The resulting graph is based on an analysis of approximately 1.4 million files gathered in the period from December 2008 to December 2012 by the Internet Memory foundation[28]. Additional content sets that are gathered for historical analysis and shared publication include a set of over 400 million web resources collected in the Danish web archive over almost a decade and characterized using fits[29].

Other specific adaptors demonstrate the capacity of Scout to incorporate new information and identify new preservation risks. Faria *et al.* (2013) describe a case study that demonstrates how to use information extraction technologies on crawled web content to extract specific domain cases, like publisher–journal relationships, and integrate it with Scout for monitoring producers in journal repositories.

Another specific adaptor feeds large-scale experiments on the renderability analysis of web pages into the knowledge base. Here, image snapshots are taken of pages from web archives with different web browsers, and the result is compared with image quality assurance tools. Expanding the comparison with structural information from the web page and cross-relation with content profiles of the resources used by the page will give further insight into which formats and which of their features are affecting the renderability of pages on modern web browsers.

### *4.4 Plato: scalable decision-making*

Upon discovery of a risk or misalignment between the organization's content and actions and the objectives, a plan is needed to resolve the detected problem and improve the robustness of the state of the repository against preservation threats. Creating such

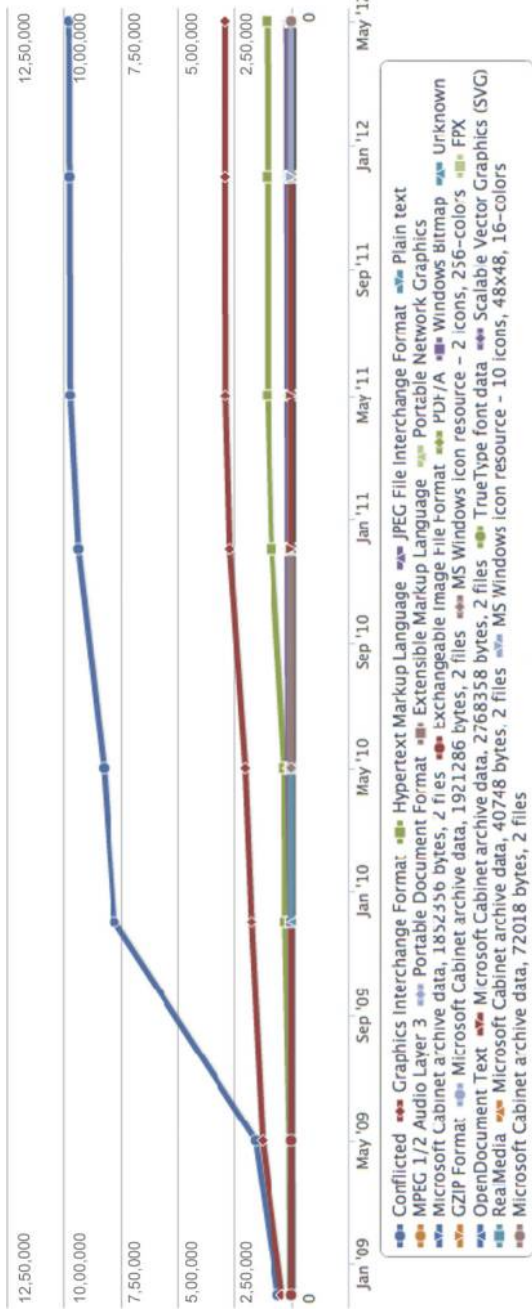


**Figure 10.**
A format distribution of 1.4 M files from the Internet memory archive

a plan is supported by the publicly available open-source planning tool Plato, which implements the preservation planning method described in detail in (Becker *et al.*, 2009). The tool guides decision-makers through a structured planning workflow and supports them in producing an actionable preservation plan for a defined set of objects. In doing so, they use a thorough goal-oriented evidence-based evaluation of the potential actions that can be applied. Controlled experimentation on real sample content is at the heart of the four-phase workflow shown in Figure 11: Testing the candidate actions on real-world content greatly increases the trust that stakeholders put into the actions to be taken and ensures that the chosen steps are not simply taken from elsewhere and applied blindly, but will be effective and fit for the specific situation (Becker and Rauber, 2011c):

- *Define requirements*: In the first phase, the context of planning is documented, and decision criteria are specified that can be used to find the optimal preservation action. The specification starts with high-level goals and breaks them down into quantifiable criteria. The resulting objective tree provides the evaluation mechanism for choosing from the candidate preservation actions. To enable this, the set of objects to preserve is profiled, and sample elements are selected that will be used in controlled experimentation.
- *Evaluate alternatives*: In an experiment step, empirical evidence is gathered about all potential candidate solutions by applying each to the sample content selected. The results are evaluated against the decision criteria specified in the objective tree.
- *Analyze results*: For each decision criterion, a utility function is defined to allow the comparison across different criteria and their measures. This utility function maps

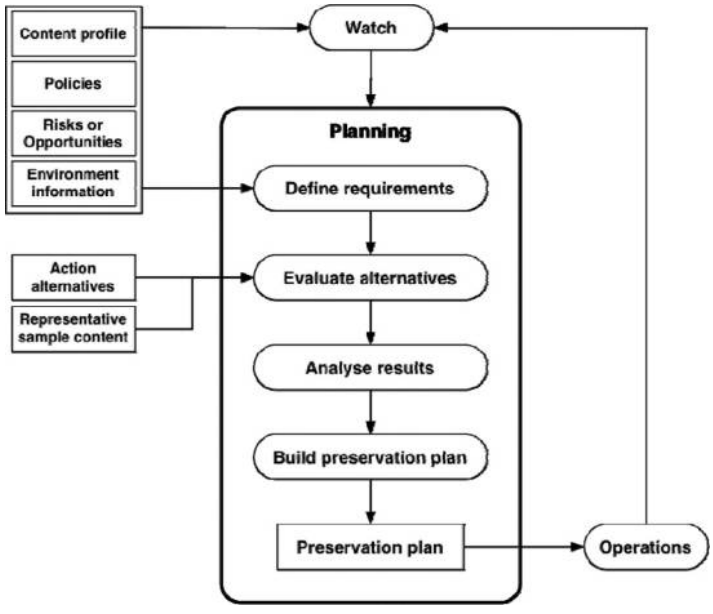


**Figure 11.**
The planning workflow

all measures to a uniform score that can be aggregated. Relative weights model the preferences of the stakeholders on each level of the goal hierarchy. An in-depth visual and quantitative analysis of the resulting score of candidates leads to a well-informed recommendation of one alternative to choose.

- *Build preservation plan*: In this final phase, the concrete plan for action is defined. This includes an accurate and understandable description of which action is to be executed on which objects and how, and specifies the quality assurance measures to be taken along with the action to ensure that the results are verified and correspond to the expected outcomes. Responsibilities and procedures for plan execution are defined. The finished preservation plan drives the activities in operations and Watch and will be reevaluated over time.

Plato has been used for operational preservation planning in different scenarios in recent years. The Bavarian State Library, for example, evaluated the migration options for one of their largest collections of scanned images of sixteenth-century books (Kulovits *et al.*, 2009). A detailed discussion of this and several other case studies is given in (Becker and Rauber, 2011b). At this point, creating a preservation plan still was an effort-intensive and complex task, as many of the required activities had to be carried out manually for each plan. However, the collected set of real-world cases enabled systematic analysis of the variety of decision factors and a systematic categorization and formalization of the criteria used for decision-making (Becker and Rauber, 2011a; Kulovits *et al.*, 2013b). Figure 12 shows Plato visualizing aggregated decision criteria collected in the knowledge base. This is increasingly supporting Plato in becoming context-aware and automating many of the steps that have previously prohibited large-scale, policy-driven preservation planning (Kraxner *et al.*, 2013; Kulovits *et al.*, 2013b).

As part of the tool suite presented here, Plato has been integrated with Scout, C3PO, and an online catalog for preservation components published as reusable, semantically annotated workflows on myExperiment[30]. An actionable preservation plan can contain a complex number of automated processing steps of different kinds of operations, linked through a pipeline of inputs and outputs that is best represented as a workflow as shown in Figure 13. Specifying such a workflow in a standard manner, as opposed to a textual operations manual, greatly reduces the risk of operational errors and streamlines deployment. The integration of Plato with the Taverna workflow engine provides such possibilities (Kraxner *et al.*, 2013).

Plato furthermore is endowed with an awareness of the control policies encompassing objectives and constraints to be followed. This understanding of the drivers and constraints of an organization is provided by an awareness of the semantic policy model which can be shared across members of the same organization. This removes much of the burden of contextual factors needing clarification, which previously accounted for much of the difficulty in starting a planning process (Becker and Rauber, 2011c; Kulovits *et al.*, 2009). Together, this removes much of the effort required for preservation planning: the institutional context is provided by and documented by a semantic model; content statistics, samples and technical descriptors are provided by the content profile; and the available actions that can mitigate risks such as obsolescence can be discovered on myExperiment. Finally, executable workflows can be deployed to the repository, removing risks of misunderstandings and

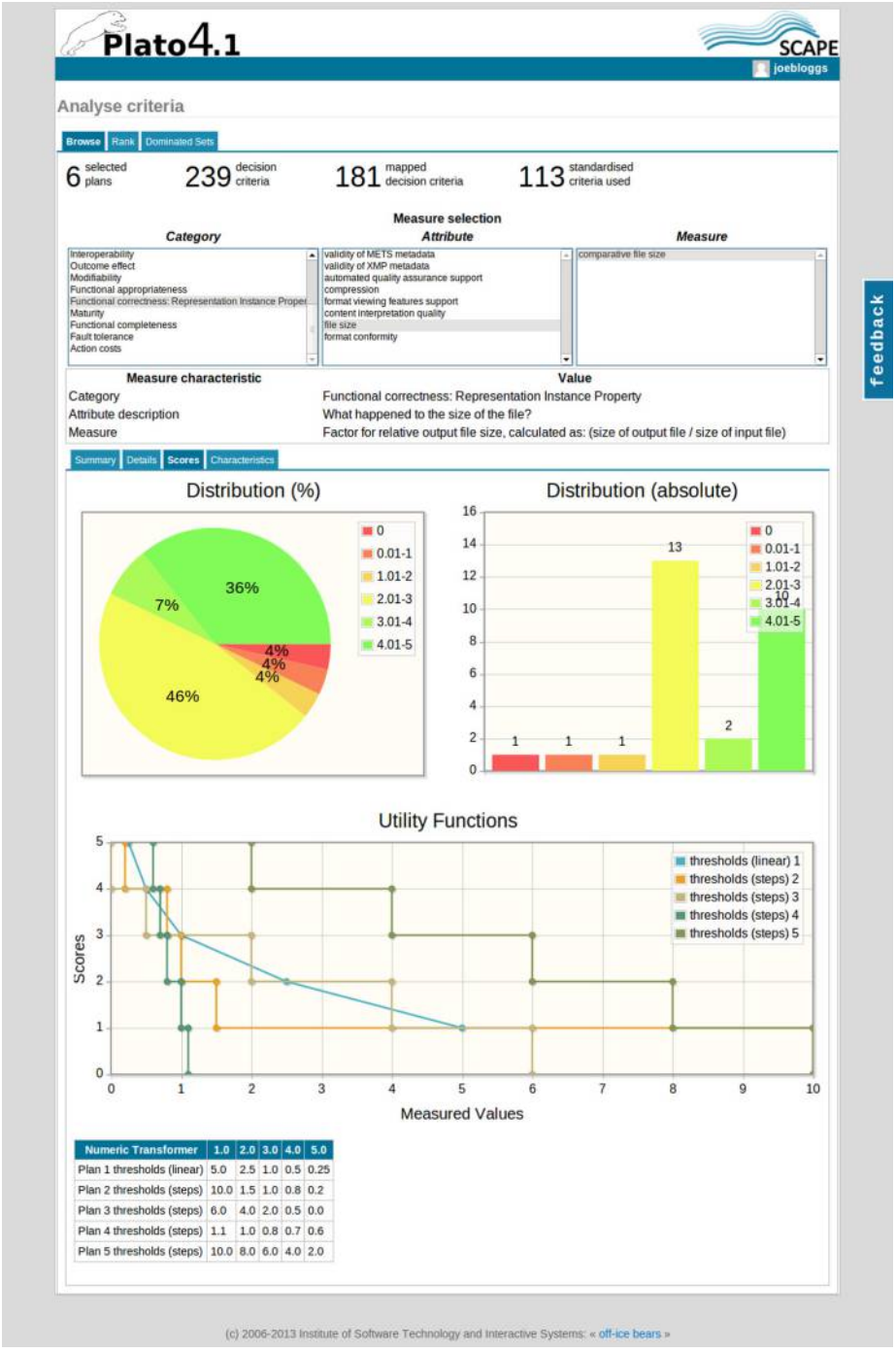


**Figure 12.**
Plato visualizing criteria statistics from its knowledge base (Becker *et al.*, 2013)

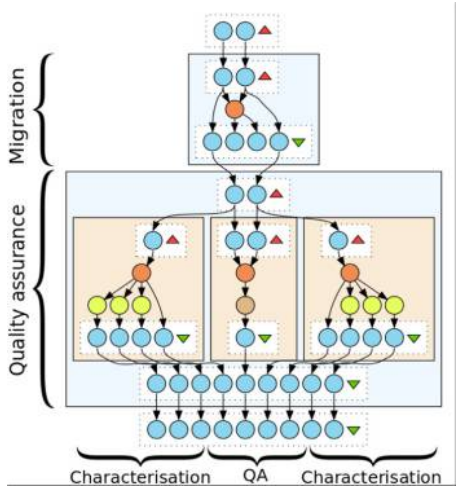


**Figure 13.**
Preservation operations are composed of multiple components (Kulovits *et al.*, 2013b)

misconfigurations and easing the burden of running operations in accordance to specifications (Kraxner *et al.*, 2013).

This awareness and the integration with an open and growing experiment sharing platform plus an open controlled vocabulary provides the basis for continued improvement of operations over time, as organizations can build on each other's work, show quantitative improvement of new solution components over those previously available and discover which solution components are needed most urgently.

As an example, consider the need to verify the quality of migration processes with respect to content authenticity: when converting even seemingly simple artifacts such as digital photographs, many conversion components introduce subtle errors by omitting embedded metadata, misinterpreting white balance and color setting or using lossy compression methods where none was expected. Automated means are required to validate each conversion (Bauer and Becker, 2011), but developing these is a heavy burden for each organization on their own. Instead, by showing that certain quality checks are required by multiple scenarios, efforts can be shared and focused on those aspects that are most frequent and at the same time critical for decision makers. The visual analysis shown in Figure 12 supports this by visualizing the quantified impact of each decision criterion and computing aggregated impact factors for arbitrary sets of criteria and preservation plans (Becker *et al.*, 2013).

### *4.5 Repository*

The repository is defined here as the system that contains and manages content, allowing ingest and access features. A repository may be as simple as a shared folder with files that represent the content, or as complex as dedicated systems such like DSpace or RODA. There are many different types and implementations of repositories, each with different features and a focus on the needs of different types of institutions. Endowing a repository with digital preservation features should therefore be independent on the repository type and implementation. To achieve the integration with the tools described above, that effectively support the digital preservation processes, a

set of repository integration APIs are defined: Data Connector API, Report API and Plan Management API.

*4.5.1 Data connector API.* The Data Connector API is an interface that allows access and modification of content in the repository. Defined as a RESTful web service (Fielding, 2000), it contains methods to:

- retrieve intellectual entities, metadata, representations, files and named bit streams;
- ingest an intellectual entity (synchronously and asynchronously);
- update an intellectual entity, a representation or a file; and
- search intellectual entities, representations or files using Search/Retrieval via URL protocol[31].

The SCAPE Digital Object Model defines how to represent the intellectual entities, metadata, representations, files and named bit streams defined above. It defines a METS[32] profile that uses PREMIS[33] to specify the technical metadata, the rights associated with the object, and the digital provenance metadata.

The Data Connector API specification and SCAPE Digital Object Model is available[34], and the API reference implementations are provided by RODA and Fedora Commons 4.

*4.5.2 Report API.* The report API is an interface that provides access to repository events such as:

- ingest started or finished;
- descriptive metadata viewed or downloaded;
- representation viewed or downloaded; or
- preservation plan executed.

The Report API is defined as an OAI-PMH[35] provider that uses PREMIS metadata to describe the repository events. The PREMIS Agent is used to define who triggered the event, PREMIS Date/time to define when the event has occurred and PREMIS Details is used to describe what has happened. The OAI-PMH protocol allows harvesting of all events and filtering by date and type of event. A Scout Report API adaptor harvests all events and creates aggregations of the events[36]. The Report API specification is available[37] and a reference implementation is available in RODA[38]. A Fedora Commons reference implementation is being developed.

*4.5.3 Plan management API.* This interface provides the facilities to deploy and manage preservation plans in the repository. Defined as a RESTful web service, it contains methods to:

- search and retrieve plans;
- deploy a new plan;
- retrieve or add a preservation execution state (e.g. in progress, success or fail); and
- enable and disable a preservation plan.

The implementation of the Plan Management API (called the Plan Management Component) can use a Workflow Engine such as Taverna, which understands the

workflow language in which the action plan is defined, to execute the workflow and run its preservation actions and quality assurance components. Finally, the Plan Management Component can use the Data Connector API to merge the result of preservation action, such as migration, back into the repository.

The Plan Management API specification is available online[39], and the API reference implementations are being developed by RODA and Fedora Commons 4.

### *4.6 Workflow engine*

Any complex set of operations such as those outlined in Section 4.4 will benefit from a workflow environment to support the coordinated execution on large amounts of content. The system design separates the implementation-level detail of such a workflow engine to enable integration of different platforms. However, there is strong tool support available based on an integration of the workflow engine Taverna and the workflow sharing platform myExperiment, where fully annotated solution components can be published for sharing and discovery. Operational preservation plans can be created and specified as Taverna workflows and published using semantic annotations following the controlled vocabularies described above (Kraxner *et al.*, 2013). Components published using this ontology can be discovered automatically and monitored for specific properties in Scout. The aggregated experience collected on their behavior can support early selection and recommendation of likely fits in the planning process. However, an organization who wishes to support a different workflow engine could replace Taverna with a platform of their own choice.

### *4.7 Automating the preservation lifecycle*

To illustrate how the presented suite of tools can support the preservation lifecycle, consider the following scenario. An institution has a repository with content and policies in place. These policies might not be formalized, and some even only documented implicitly, but they are what represents the intentions of an organization and should guide and constrain all the preservation processes.

A first step requires the institution to define and formalize the purpose of the content and the digital preservation requirements associated with it. This requirements start by a high level definition of the mission and objectives, such as "long-term availability and authenticity", and must iteratively relate to low level requirements that relate to tangible and measurable facts, for example, "no compression allowed". These more specific requirements, i.e. control policies, should be defined using the SCAPE policy model.

By running characterization tools and C3PO on the repository and configuring the Scout adaptors for repository integration, which uses the C3PO and the Report API, Scout will be able to constantly monitor characteristics of the content and the repository events of importance for digital preservation. Scout provides the facility to upload the policies defined in the SCAPE policy model and activate a set of triggers. It will then notify the users when policy conformance is not fulfilled. These triggers might need external information monitored by Scout, such as the content of format and tool registries, different classes of experiments and even manually inserted human knowledge. Scout may, for example, detect that some content uses compression, but that this violates a defined policy, and hence send an email notification to the Planner.

The third step is to decide which actions should be taken to mitigate this problem. The Planner can use Plato to support the creation of a well-described and traceable preservation plan that addresses the detected preservation risk. By knowing the defined preservation policies, Plato can pre-fill many of the necessary contextual bits of required information, supporting the reuse of the institution's objectives definition and greatly reducing the time needed to create a preservation plan (Kulovits *et al.*, 2013a). Furthermore, Plato can automatically find and retrieve solution alternatives by connecting to the myExperiment preservation components catalog. Also, Plato can automatically conduct experiments on all alternatives discovered in myExperiment, applying them to the set of sample objects. The analysis of results is partially supported by quality assurance tools that provide an evaluation of the behavior of each alternative considering the case requirements, which enables the decision-maker to discover the best solution.

The fourth step is to deploy the preservation plan into the repository via the Plan Management API. The Plan Management Component of the repository can use a workflow engine to execute the preservation action, including the quality assurance steps, and use the Data Connector API to merge the action results back into the repository. The results of the preservation action quality assurance step are sent to Scout via the Report API, so that Scout can monitor if the action performed as expected. Finally, the preservation plan contains triggers to be installed in Scout to automatically monitor if the assumptions taken on the decision-making step remain true. If the action plan does not execute as expected or if the preservation plan needs to be reviewed because policies or the environment have changed, then the Planner is again notified to re-evaluate the preservation plan, starting again the cycle.

## 5. Summary

Digital Preservation is the set of activities and processes required to ensure the continued, authentic access to digital content over time. Providing such information longevity across changing socio-technical environments poses a number of challenges, in particular in the light of recent rising content volumes. Scalability for handling large amounts of data can be achieved by state of the art technologies commonly used in the cloud. Additionally, scalable monitoring and decision-making is required to support automated, large-scale operations of systems and tools.

Scaling up decision-making, policy definition and processes for monitoring and actions requires a set of techniques that include scalable in-depth content analysis, intelligent information gathering and efficient multi-criteria decision support. But it also requires loosely coupled systems that are able to interact with each other and the wider preservation context and are capable of evolution over time, and a set of common vocabularies that can be used to publish and discover knowledge about the evolving preservation ecosystems.

This article presented the SCAPE Planning and Watch suite, a new, innovative system for scalable decision-making and control in preservation environments. The Planning and Watch suite builds on Plato and extends it into a loosely coupled, extensible preservation planning and monitoring system that can be integrated with virtually any repository and content management system through open and standardized interfaces. While each of the components can be used and integrated independently from the other components, this article focused on the compound value

contribution that can be obtained by the set of systems and showed how the resulting SCAPE ecosystem can support organizations in managing their holdings more effectively, using policy-driven monitoring and well-supported decision-making systems to provide scalable decision-making and control capabilities in support of digital preservation objectives.

According to the study by Becker *et al.* (2015), we will conduct a systematic assessment of the system based on the design goals outlined in this article. We will discuss the improvements of the presented work and identified limitations, based on a quantitative and qualitative evaluation including a case study with a national library.

## Notes

1. www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm
2. http://github.com/openplanets/fido
3. http://jhove.sourceforge.net/
4. http://tika.apache.org/
5. http://code.google.com/p/fits/
6. www.scape-project.eu/tools
7. www.digicult.info/pages/techwatch.php
8. http://dpconline.org/advice/technology-watch-reports
9. www.nationalarchives.gov.uk/PRONOM/
10. www.gdfr.info
11. http://udfr.cdlib.org/
12. http://p2-registry.ecs.soton.ac.uk
13. http://fileformats.archiveteam.org/ is one example.
14. http://blogs.loc.gov/digitalpreservation/2013/06/why-cant-you-just-build-it-and-leave-it-alone/
15. www.dspace.org
16. www.eprints.org
17. www.roda-community.org
18. https://github.com/openplanets/scape-apis
19. www.myexperiment.org/, www.scape-project.eu/
20. https://github.com/, www.scape-project.eu/
21. www.scape-project.eu/tools
22. https://code.google.com/p/fits/
23. http://peshkira.github.io/C3PO/
24. https://github.com/peshkira/C3PO/blob/master/format/C3PO.xsd
25. http://openplanets.github.io/scout/
26. http://wiki.opf-labs.org/display/SP/Comparison+of+Web+Snapshots
27. www.w3.org/TR/rdf-sparql-query/
28. http://internetmemory.org

29. www.openplanetsfoundation.org/blogs/2013-01-09-year-fits
30. http://myexperiment.org
31. www.loc.gov/standards/sru/
32. www.loc.gov/standards/mets/
33. www.loc.gov/standards/premis/
34. https://github.com/openplanets/scape-apis/
35. www.openarchives.org/pmh/
36. https://github.com/openplanets/scout/tree/master/adaptors/report-api-adaptor
37. https://github.com/openplanets/scape-apis
38. https://github.com/openplanets/roda
39. https://github.com/openplanets/scape-apis



## References

Abrams, S.L. (2004), "The role of format in digital preservation", *VINE: The Journal of Information and Knowledge Management Systems*, Vol. 34 No. 2, pp. 49-55.

Antunes, G., Borbinha, J., Barateiro, J., Becker, C., Proenca, D. and Vieira, R. (2011), "Shaman reference architecture", Shaman Project Report, Version 3.0, available at: www.shaman-ip.eu/sites/default/files/SHAMAN-Reference%20Architecture.pdf

Bauer, S. and Becker, C. (2011), "Automated preservation: the case of digital raw photographs", in *Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation Proceedings of 13th International Conference on Asia-Pacific Digital Libraries (ICADL 2011)*, Springer-Verlag, Beijing.

Beagrie, N., Semple, N., Williams, P. and Wright, R. (2008), *Digital Preservation Policies Study Part 1: Final Report*, Higher Education Funding Council for England, available at: www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport.pdf

Becker, C., Duretec, K. and Faria, L. (2014), "Scalable decision support for digital preservation: an assessment", *OCLC Systems and Services*, Vol. 31 No. 1.

Becker, C., Duretec, K., Petrov, P., Faria, L., Ferreira, M. and Ramalho, J.C. (2012), "Preservation watch: what to monitor and how", in *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES)2012*, Toronto.

Becker, C., Kraxner, M., Plangg, M. and Rauber, A. (2013), "Improving decision support for software component selection through systematic cross-referencing and analysis of multiple decision criteria", in *Proceedings of 46th Hawaii International Conference on System Sciences (HICSS)*, Maui, HI, pp. 1193-1202.

Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A. and Hofman, H. (2009), "Systematic planning for digital preservation: evaluating potential strategies and building preservation plans", *International Journal on Digital Libraries*, Vol. 10 No. 4, pp. 133-157.

Becker, C. and Rauber, A. (2011a), "Decision criteria in digital preservation: what to measure and how", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 6, pp. 1009-1028.

Becker, C. and Rauber, A. (2011b), *Four Cases, Three Solutions: Preservation Plans for Images*, Technical Report, Vienna University of Technology, Vienna.

Becker, C. and Rauber, A. (2011c), "Preservation decisions: terms and conditions apply. Challenges, misperceptions and lessons learned in preservation planning", in *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Ottawa, pp. 67-76.

Brody, T., Carr, L., Hey, J., Brown, A. and Hitchcock, S. (2008), "PRONOM-ROAR: adding format profiles to a repository registry to inform preservation services", *The International Journal of Digital Curation*, Vol. 2 No. 2, pp. 3-19.

CCSDS. (2002), "Reference model for an open archival information system (OAIS)", available at: http://public.ccsds.org/publications/archive/650x0b1.pdf

Dean, J. and Ghemawat, S. (2004), "MapReduce: simplified data processing on large clusters", in *Proceedings of 6th Conference on Symposium on Operating System Design & Implementation*, Berkeley, CA.

Faria, L., Akbik, A., Sierman, B., Ras, M., Ferreira, M. and Ramalho, J.C. (2013), "Automatic preservation watch using information extraction on the Web", in *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPRES)*, Lisbon.

Faria, L., Petrov, P., Duretec, K., Becker, C., Ferreira, M. and Ramalho, J.C. (2012), "Design and architecture of a novel preservation watch system", in *The Outreach of Digital Libraries: A Globalized Resource Network Proceedings of 14th International Conference on Asia-Pacific Digital Libraries (ICADL)*, Taipei, pp. 168-178.

Fielding, R.T. (2000), "Architectural styles and the design of network-based software architecture", *Doctoral dissertation*, University of California, Irvine, CA.

Garrett, J. and Waters, D. (1996), *Preserving Digital Information: Report of the Task Force on Archiving Digital Information*, The Commission on Preservation and Access and Research Libraries Group, available at: www.clir.org/pubs/reports/pub63watersgarrett.pdf

Hedstrom, M. (1998), "Digital preservation: a time bomb for digital libraries", *Journal of Computers and the Humanities*, Vol. 31 No. 3, pp. 189-202.

Heslop, H., Davis, S. and Wilson, A. (2002), "An approach to the preservation of digital records", Green Paper, National Archives of Australia, available at: www.naa.gov.au/Images/An-approach-Green-Paper_tcm16-47161.pdf

Hunter, J. and Choudhury, S. (2006), "PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services", *International Journal on Digital Libraries*, Vol. 6 No. 2, pp. 174-183.

Hutchins, M. (2012), *Testing Software Tools of Potential Interest for Digital Preservation Activities at the National Library of Australia*, Technical Report, National Library of Australia, Canberra.

International Standards Organisation (ISO) (2010), *Space data and Information Transfer Systems – Audit and Certification of Trustworthy Digital Repositories (ISO/DIS 16363)*, International Standards Organisation, available at: www.iso.org/iso/catalogue_detail.htm?csnumber=56510

Jackson, A. (2012), "Formats over time: exploring UK Web history", in *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES)2012*, Toronto.

Knijff, J. and Wilson, C. (2011), "Evaluation of characterization tools", Technical Report, available at: www.scape-project.eu/wp-content/uploads/2012/01/SCAPE_PC_WP1_identification21092011.pdf

Kraxner, M., Plangg, M., Duretec, K., Becker, C. and Faria, L. (2013), "The SCAPE planning and watch suite", in *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPRES)2013*, Lisbon.

Kulovits, H., Becker, C. and Andersen, B. (2013a), "Scalable preservation decisions: a controlled case study", in *Proceeding of Archiving 2013*, Washington, DC, pp. 167-172.

Kulovits, H., Kraxner, M., Plangg, M., Becker, C. and Bechofer, S. (2013b), "Open preservation data: controlled vocabularies and ontologies for preservation ecosystems", in *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPRES)2013*, Lisbon.

Kulovits, H., Rauber, A., Kugler, A., Brantl, M., Beiner, T. and Schoger, A. (2009), "From TIFF to JPEG2000? Preservation planning at the Bavarian state library using a collection of digitized 16th century printings", *D-Lib Magazine*, Vol. 15 Nos 11/12.

Law, M.T., Thome, N., Gançarski, S. and Cord, M. (2012), "Structural and visual comparisons for web page archiving", in *Proceedings of the 2012 ACM Symposium on Document Engineering (DocEng'12)*, New York, NY, pp. 117-120.

Lawrence, G.W., Kehoe, W., Kenny, A.R., Rieger, O.Y. and Walters, W. (2000), *Risk Management of Digital Information: A File Format Investigation*, Council on Library and Information Resources, available at: www.clir.org/pubs/reports/pub93/pub93.pdf

Object Management Group (2008), *Semantics of Business Vocabulary and Business Rules (SBVR)*, Version 1.0., Object Management Group, available at: www.omg.org/spec/SBVR/1.0/

Object Management Group (2010), *Business Motivation Model 1.1*, Object Management Group, available at: www.omg.org/spec/BMM/1.1/

OCLC and Center for Research Libraries (CRL) (2007), *Trustworthy Repositories Audit and Certification: Criteria and Checklist*, OCLC and Center for Research Libraries, available at: www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

Pearson, D. (2007), "AONS II: continuing the trend towards preservation software Nirvana", in *Proceedings of the 4th International Conference on Preservation of Digital Objects (iPRES)2007*, Beijing.

Pehlivan, Z. (2013), "Quality assurance workflow, release 2 + release report", Technical Report, available at: www.scape-project.eu/wp-content/uploads/2013/06/SCAPE_D11.2_UPMC_V1.0.pdf

Pennock, M., Jackson, A. and Wheatley, P. (2012), "CRISP: crowdsourcing representation information to support preservation", in *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES)2012*, Toronto.

Petrov, P. and Becker, C. (2012), "Large-scale content profiling for preservation analysis", in *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES)2012*, Toronto.

Plugge, E., Hawkins, T. and Membrey, P. (2010), *The Definitive Guide to MongoDB: The NoSQL Database for Cloud and Desktop Computing*, Apress, New York, NY.

Rothenberg, J. (1995), "Ensuring the longevity of digital documents", *Scientific American*, Vol. 272 No. 1, pp. 42-47.

Sinclair, P., Billenness, C., Duckworth, J., Farquhar, A., Humphreys, J. and JArdine, L. (2009), "Are you ready? Assessing whether organisation are prepared for digital preservation", in *Proceedings of the 6th International Conference on Preservation of Digital Objects (iPRES)2009*, San Francisco, CA, pp. 174-181.

Strodl, S., Rauber, A., Rauch, C., Hofman, H., Debole, F. and Amato, G. (2006), "The DELOS testbed for choosing a digital preservation strategy", in *Digital Libraries: Achievements, Challenges and Opportunities*, *Proceedings on 9th International Conference on Asian Libraries (ICADL)*, Springer-Verlag, Kyoto, pp. 323-332.

Thaller, M. (2009), *The eXtensible Characterisation Languages – XCL*, Verlag Dr Kovac, Hamburg.

Webb, C., Pearson, D. and Koerbin, P. (2013), "Oh, you wanted us to preserve that?! Statements of preservation intent for the National Library of Australia's digital collections", *D- Lib Magazine*, Vol. 19 Nos 1/2, available at: www.dlib.org/dlib/january13/webb/01webb.html



**Corresponding author**
Christoph Becker can be contacted at: christoph.becker@utoronto.ca