Preservation Watch: What to monitor and how

Christoph Becker, Kresimir Duretec, Petar Petrov
Vienna University of Technology
Vienna, Austria
{becker,duretec,petrov}@ifs.tuwien.ac.at

Luis Faria, Miguel Ferreira
KEEP SOLUTIONS
Braga, Portugal
{lfaria,mferreira}@keep.pt

Jose Carlos Ramalho
University of Minho
Braga, Portugal
jcr@di.uminho.pt

ABSTRACT

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 organisation's preservation objectives, and associated risks and opportunities. This requires linking a number of diverse information sources and specifying complex conditions. For example, the content to be preserved needs to be related to specifications of significant properties, to the file formats used to represent that content, to the software environments available to analyse, render and convert it and the technical environments available to the designated user communities that will consume this content.

This article analyses aspects of interest in a preservation context that call for automated monitoring and investigates the feasibility of drawing sufficient information from diverse sources together and linking it in a meaningful way. We define a number of preservation triggers that lead to preservation planning activities and present the requirements and a high-level design of a preservation watch system that is currently being developed. We demonstrate the value of such a monitoring approach on a number of scenarios and cases.

Categories and Subject Descriptors

H.1 [**Information Systems**]: Models and Principles; H.3 [**Information Systems**]: Information Storage and Retrieval; H.3.7 [**Information Systems**]: Information Storage and RetrievalDigital Libraries; K.6.4 [**Computing Milieux**]: Management of computing and Information Systems—*System Management*

Keywords

Digital Preservation, Preservation Planning, Monitoring, Watch

1. INTRODUCTION

Digital Preservation is in essence driven by change of organisational and technical kind. Aspects of change range from platform technologies and rendering environments to storage media, shifting modes of access and interactivity, and finally, shifts in the semantics of information itself. Any archival information system thus needs to continuously adapt to changing environments to ensure alignment between preservation operations and the goals and objectives of the system.

Monitoring is recognised as a key element of succesful preservation. However, to date it is mostly a manual process that is sporadically initiated as a reaction to urgent questions. At best, technical reports are produced about selected topics and circulated within the community.

More and more cases of successful preservation operations are being developed. The SCAPE project[1] is focusing on scalable operations for preserving massive amounts of information through data-centric parallel execution mechanisms [7]. However, for such operations to be scalable and successful over time, automated mechanisms and processes for control and monitoring need to be designed.

Isolated strands of systematically collecting information that can be used to guide preservation decision making have been developed. Well-known examples include registries of file formats or emulation environments. However, these are far from being complete in the information they cover, and there are few links between the islands of information.

For an organisation responsible for managing a digital repository over time, the corresponding *monitoring capabilities* that are required can be described as

1. *Internal monitoring* of the systems in place, the operations in place, the assets and activities, and
2. *External monitoring* of the world of interest such as user communities, technologies, and available solutions.

Based on these systematic information gathering processes, preservation planning as decision making capability can then act well-informed to ensure that what the organisation does to keep content authentic and understandable is sufficient and optimal. A number of questions arise in this context.

1. Which are the key aspects that need to be monitored? What are the main entities and which properties need to be considered?
2. How can the information be collected? How can it be represented?

[1] `http://www.scape-project.eu`

3. How can the information be linked together so that connections between parts of the data can be made?
4. What properties does a system need to possess to enable automated monitoring and linkage between the relevant aspects?
5. How can this be deployed and used in a way that multiple organisations can mutually benefit from each other's effort and experience?
6. How can we ensure that such a platform will be extensible and create synergies between different players so that the knowledge base grows continously?

In this article, we discuss these questions and propose a design for such a system. We envision a 'Watch component' to collect information from a number of sources, link it appropriately together, and provide notifications to interested parties when specified conditions are satisfied. The motivation of this Watch component is in part driven by the experience gathered in preservation planning: The preservation planning tool Plato[2] provides powerful and systematic decision making support and includes an increasingly formalised model of relevant aspects, entities and properties to be considered in preservation planning [9]. Plato is not designed to provide continuous monitoring capabilities, but preservation plans specified as the result of such planning activities specify which aspects to monitor based on the influencers considered in decision making. It is then the responsibility of the Watch component presented here to continuously monitor the state of the world and of the system in question to determine whether conditions are met that may require an update of plans and operations.

Section 2 outlines the background of monitoring in the context of preservation and illustrates typical information sources that are of relevance. In Section 3 we discuss a model of drivers, events, conditions and triggers for preservation watch. Section 4 describes the various sources of information that are being leveraged. Section 5 summarises the key design goals of a preservation watch system and presents a high-level design of the Watch component. Section 6 illustrates typical conditions and benefits of the approach in a scenario based on a real-world preservation planning case, while Section 7 summarises key risks and benefits and outlines the next steps ahead.

2. BACKGROUND

Monitoring is a common subject in any domain that must cope with the demands of a changing environment. It is a major input for decision-making and ensures that specified plans are continuously adapted to changes in the environment. Monitoring feeds back to decision-making to close a continuous adaptative cycle [4]. In the digital preservation domain, monitoring is especially critical as the domain challenge itself stems largely from rapidly changing environments. The need and concept of monitoring have been identified and discussed before [5, 2]. However, these all focus on a very high-level approach of preservation monitoring and do not define a systematic method or guideline on how to accomplish that capability. Furthermore, the tools presently known to support preservation monitoring are mainly manual, incomplete and used in an ad-hoc fashion. Monitoring now comes in the form of research studies and technical reports, format and tool registers, and generic application catalogues outside of the core preservation domain [8].

An early influential report on file format risks and migration strategies discusses risks that executing or postponing a migration might introduce [13]. Regular Technology Watch Reports of the Digital Preservation Coalition provide focused discussions on emerging topics and include technical investigations.[3] However, none of this is machine-understandable.

Online registries with technical information about file formats, software products and other technical components relevant to preservation have been available for some time. This includes the well-known examples PRONOM[4], The Global Digital Format Registry[5] (GDFR) [1], and the newly released Unified Digital Format Registry[6] (UDFR). Complementary approaches include the P2 registry[7] based on semantic web technologies [15], and the Conversion Software Registry[8]. Unfortunately, these online registries are not yet functioning or are not very complete. For example, relevant risk factors *per format* are only covered for a handful of entries.

Online software catalogues monitor new versions of software for a generic domain use. These sites do not specifically consider digital preservation aspects, but provide comprehensive descriptions and commonly have a social component that can contain interesting information. Some examples of these sites are CNET's `download.com`[9] and `iUseThis`[10]. App stores like Apple's Mac App Store and Ubuntu's Software Center and repositories can also be a good source of information. In the domain of digital preservation, TOTEM - the Trustworthy Online Technical Environment Metadata Database tries to address the gap of linking environments and compatible software, but is limited to emulated environments addressed within the KEEP project.[11]

This overview relates solely to file formats and tools for conversion of file formats or emulation, but many more information sources are required. Furthermore, it has to be noted that sources that focus on digital preservation have a generally very reduced coverage (registries) or machine-readability (reports), while general purpose sources normally cover very limited facets of the information relevant for digital preservation. Finally, none of these sources allows preservation monitoring to be done automatically and alert the user when a preservation risk is identified. However, this step towards automation is crucial: As content grows in volume and becomes increasingly heterogeneous, the aspects of technologies that need to be monitored are by far outgrowing any organisation's manual capabilities.

The OAIS model [5] includes, within the functional entity Preservation Planning, the functional components "Monitor Designated Community" and "Monitor Technology". These provide the core monitoring functions in a repository scenario. Monitoring the user community is meant to focus

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

[3] `http://www.dpconline.org/advice/technology-watch-reports`

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

[5] `http://www.gdfr.info`

[6] `http://www.udfr.org`

[7] `http://p2-registry.ecs.soton.ac.uk`

[8] `http://isda.ncsa.uiuc.edu/NARA/CSR`

[9] `http://download.com`

[10] `http://iusethis.com`

[11] `http://keep-totem.co.uk`

on "service requirements and available product technologies", while technology monitoring includes "tracking emerging digital technologies, information standards and computing platforms (i.e., hardware and software)" [5]. The OAIS also mentions monitoring functions of the Administration entity as well as monitoring archival storage and systems configurations, but these are seen as separate functions.

Historically, the identification of risks by monitoring tools has been delegated into other tools such as file format registries or seen as a manual task such as providing technical reports. A step forward in automation was done in the initiative to create an Automatic Obsolescence Notification Service (AONS) that would provide a service for users to automatically monitor the status of file formats in their repositories against generic format risks gathered in external registries and receive notifications [14]. The system would gather collection profiles from repositories, by using format identification tools on content, and seek obsolescence risk indicators on file format information in external registries (like PRONOM). The system would also allow caching and extending format registries and the creation of new adaptors for new registries. The notification service allows subscription of various events, like end of a repository crawl or change in the information about a format, and send a notification via email, RSS feed and task boxes on the GUI.

AONS was limited to gathering information about file formats, assuming that all other aspects of the world that would be relevant for digital preservation would be gathered, assessed, and represented in a structured way by the external registries. This assumption and the lack of available information in format registries constrained the usefulness of the system. Moreover, not all desired features defined in the concepts could be successfully completed.

The lack of a defined methodology for systematic preservation monitoring and tools that help to enforce and automatize this capability forces content holder institutions to create their own methodology, highly based on manual effort and therefore scattered throughout many departments and different people via responsibility distribution, which results in a partial and stratified view of the properties that condition decisions. Furthermore, most institutions cannot afford the effort to have even this partial view on the environment and thus ignore or postpone efforts for preservation monitoring [6].

In contrast, a well-designed monitoring system would inspect the properties of the world and provide needed knowledge to identify risks. This knowledge requires an integration of several aspects of the world – tools and formats, but also content, empirical evidence, organizational context, technology, user trends, and other aspects. Furthermore, this information needs to be cross-referenced and analytically accessible for automated procedures so that indications of risks and opportunities can be found and deep analysis processes (such as a manual intervention) can be initiated only when needed.

Our investigation of the question *What to monitor?* can build on a number of additional analytical steps that have been taken previously. On the one hand, the Plato preservation planning framework provides a systematic guidance on analysing influence factors. On the other hand, systems-oriented approaches such as SHAMAN provide a categorisation of drivers and constraints [2]. Table 2 classifies key drivers in a DP scenario in internal and external categories.

Table 1: DP drivers according to SHAMAN [2]

Internal	
Business Vision	Goals, Scope of designated community, etc.
Resources	Infrastructure (e.g., operational costs, expertise needed), Hardware (e.g., operational costs, technological capability), Software (e.g., operational costs, technological capability), Staff (e.g., expertise and qualifications, commitment)
Data	Volume, Structure, Representation, Semantics, etc.
Processes	Dependencies, Responsibilities, Alignment, etc.
External	
Producers	Demand satisfactions, Content, Technology, Trust and reputation
User community	Technology, Knowledge, Demand satisfaction, Trust and reputation
Contracts	Deposit, Supplier and service, Interoperability, Access, etc.
Supply	Technology, Services, People
Competition	Overlap of: Services, Content, User community, Producers, Technology, Mandate, Rights, Funding, Capabilities
Regulation and mandate	Regulation/Legal constraints, Embedding organization regulation, Mandate,Rights and ownership, Certification, Funding

Any of these drivers feature conditions that influence decisions and operations for preservation. Section 4 will discuss which information sources we can connect to for gathering information about these drivers.

3. AUTOMATED PRESERVATION WATCH

We envision a *Preservation Watch component* as a system that enables automated monitoring of operational preservation compliance, risks and opportunities by collecting, fusing and analysing information from various sources. From an abstract perspective, the usage of such a Watch component can be reduced to the following steps:

1. An actor has a question about a certain aspect of the world that is of interest to the agent.
2. The actor expresses this interest in the form of a question about a property that represents this interest.
3. The function of Watch then is to find a method to deliver an answer to this question that is timely and reliable.
4. Having received an answer to the question, the actor will want to assess the meaning and impact of this answer. This may require consultation of a decision-making capability.

The relevant aspects of the world, i.e. the entities and their properties about which information should be gathered, are expected to evolve and expand over time. The initial model is focused on the core question of information representation, formats and available environments. These are not the only sources of information, but instead represent the seed of key drivers to be considered.

Figure 1 shows a minimal model of the main entities of interest in the initial phase of Watch. Each of the entities shown has a number of known and named properties of interest. A few additional relationships and entities are omitted here for clarity. Organisations holding ownership

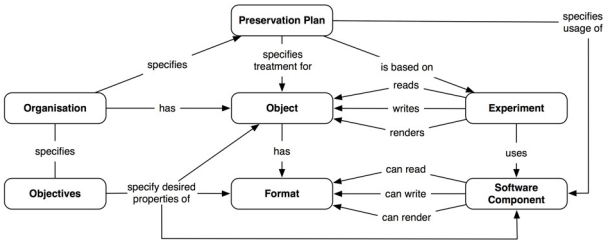

Figure 1: Minimal initial domain– model

and responsibility of objects specify objectives that relate to these objects, the resources required to preserve them, the processes used to preserve and access them and the software components used to run these processes. Objects have a number of properties, including the significant properties that need to be kept unchanged to preserve the authenticity of the object. The format of the representation is a key property and itself has a number of properties, such as the well-known risk factors commonly used to assess formats for their risk and benefits in preservation [15].

Software components for analysing, migrating, rendering and quality assuring content are key elements to ensure continued access. They are evaluated systematically in controlled experiments to provide the evidence base for the decisions specified in preservation plans [3]. These plans specify particular preservation actions to treat content for keeping it understandable and authentic. Systematic experiments are executed to test certain properties of software components on certain test data sets, containing objects. With increasing automation, systematic experiments can be scaled up and systematically conducted on large volumes of content [7]. Considering that such experiments are increasingly common across the DP domain, analysing these can uncover hidden risks and opportunities for operations in related scenarios.

Linking such information *across scenarios* enables us to answer critical questions such as the following.

- How many organisations have content in format X?
- Which software components have been tested successfully in analysing objects in format Y?
- Is the Quality Assurance tool Q, which checks documents for the equivalence of their textual content, reliable for this pair of formats?
- Has anyone encountered problems in terms of stability for this migration component when executed on very large image files?

The questions outlined above refer to known properties of identified classes such as *software components* and their properties [9]. As the preservation community becomes aware of this platform and the mutual benefits to be had from synergetic data collection, we expect this minimal model above to evolve substantially and cover additional entities and properties of interest.

For the recipient, an answer may have a variety of impacts and can overlap with other answers. Decisions are generally taken with a number of influencing factors in mind: For example, a decision to postpone a migration project may be driven by considerations on migration costs, the availability of automated tools to perform quality assurance on conversion processes, storage costs and cost models, and specific rendering environments that are at this point in time able to support content delivery to the user communities [12]. Over time, these drivers can change simultaneously. Each change can be critical, but it is only considering all relevant aspects that informed decisions can be taken.

This means that there may be simple conditions attached to a question. These conditions trigger an event when they are met, for example when the answer changes by more than 5%. The role of an automated watch process is not to assess the cumulative impact of multiple answers and what meaning they have to an external consumer of the answers. Essentially, the Watch component itself should be agnostic of the ultimate effects of changes: Its primary purpose is to make the state of the world available for assessment, not to assess it.

4. SOURCES OF INFORMATION

For any given question, several sources of information will often have to be consulted. This section gives an overview of possible sources in terms of the information they provide and attempts a high-level categorization.

Content profiles. A content profile provides statistical data about digital content of any type and offers an aggregated view of content based on its metadata, in particular detailed technical characteristics. An organisation's own content profile thus provides the basis for in-depth analysis and risk assessment. The quality of any such analysis depends on the richness of information present. While the formats contained in a repository are the first property that comes to mind, it is critical to perform a deeper analysis on other properties to uncover dependencies, feature distributions and hidden risks. By linking information such as the presence of content-specific features, embedded content types or other aspects such as the presence of digital rights management, it becomes possible to monitor often-overlooked preservation issues and collect information that can be meaningfully shared even across organisations.

An entirely different aspect can be covered when considering *others'* content profiles and content profiling on large-scale public content such as web archives. Given the massive data volumes presented there, in-depth profiling of content over time would allow us to provide indicators for file format adoption and impending obsolescence. Specific content profiles of comparable organisations, on the other hand, can enable risk assessment and comparison as well as facilitate collaboration.

Format registries. Changes in the properties of existing formats or the appearance of new formats need to be detected and compared with organisational risk profiles and content profiles. Examples of this type of sources are the PRONOM and P2 registries. However, the crucial point is the coverage of information which current information sources are still severely lacking. Designs for these systems have traditionally relied on inherently closed-world models. Moderated registries such as PRONOM have not shown to be very responsive in capturing the evolving knowledge that is available. The P2 format registry showed the benefits of Linked Data for such format information [15], and increasingly, open information models using RDF and ontologies are leveraged to capture the inherently evolving nature of format properties. This semantic web approach makes efforts such as the new UDFR building on OntoWiki a potentially very valuable source.

Software catalogues. Software components for identification, migration, characterisation or emulation are at the heart of preservation operations. We broadly categorise preservation components into Action, Characterisation and Quality Assurance components. *Action* components perform operations on content or environments, such as migration and emulation. *Analysis* components provide measures of properties in content, such as a format identification or the presence of encryption or compression. *Quality Assurance* components, finally, perform QA on preservation actions, such as algorithmic comparisons of original and converted objects or run-time analysis of rendering quality.

Table 4 lists exemplary change events that can be triggers for preservation activities. Components are continuously developed: New components are published and new versions of components are developed. These components might provide new and better migration paths, new options for performing Quality Assurance, or new and better opportunities for analysing existing content. On the other hand, new knowledge about existing components is gained continuously and, when shared, can provide tremendous value to the community.

Experiments. The role of evidence is central to trustworthy Digital Preservation [**?**]. In addition to collecting declared published information from catalogues, empirical evidence from controlled experiments are a valuable source of information. On the one hand, preservation planning experiments are executed on a subset of a collection to provide manually validated, deep insights into potential alternative actions [3]. These experiments provide valuable knowledge not only for the planning scenario in question but also for future usage. They are executed only on a subset of a whole collection, but processing this subset can still take a significant amount of time. Moreover, the experiment results will often be validated and amended manually and are therefore particularly valuable. Publishing such experimental data so that the results can be accessed can provide significant benefits [11]. On a much larger scale, the Linked Data Simple Storage specification (LDS3)[12] is being positioned to enable large-scale publication of Linked Data sets in digital preservation, describing content statistics, experimental results in content validation and conversion, benchmarks, and other experimental data. This can be used to publish experimental data from any platform and environment, as long as it is properly described.

We note that the combination of the above three information sources goes a long way in answering the questions outlined in Section 3. However, they do not cover the questions of internal systems monitoring and the alignment between a preservation system and its objectives. These are covered by the following sources.

Repository systems. Two aspects about repositories are considered: On the one hand, the state of a repository and the content it is holding is of interest (What is the growth rate of content? Are all objects covered by preservation plans?). On the other hand, repositories perform continuous operations that can provide valuable information to feed into decision making. This includes validity check as well ingest and access operations (What is the average access time using migration upon access? How many access requests have failed?) By specifying a standardised vocabulary for the core set of such events, it becomes possible to monitor whether the performed preservation activities are successful. This also supports anticipation of trends in the repository operations and usage.

[12] `http://www.lds3.org`

Table 2: Examples of software triggers

Event	Example Cause
New software	New migration software for specific formats used in the repository
	New analysis or characterization software for a certain content type or format
	New QA software for a certain content type
	New monitoring service for a question of interest
	New software version release
New knowledge about software	New testing results for a action, analysis, quality assurance or monitoring software used in the content repository
	Change in software dependencies
New repository system or version	New software release, acquisition of a new system
Capabilities	Optimized internal infrastructure leads to new technical opportunities (e.g. faster throughput in organizational SOA)

Organisational objectives. Changes in the organisations strategies and goals may respond to shifts in regulations or changes in priorities. These high-level elements of governance will be reflected in the policies of an organisation and ultimately in the specific objectives for preservation. If such objectives can be formalised, they will provide a critical starting point for monitoring fulfilment of these objectives on specified indicators. Ongoing work is formalising such a model using semantic web technologies. This can be fed into a knowledge base to enable direct queries resolving objectives against the state of the world.

Simulation. Based on trends that can be inferred from gathered data, models can be created to predict future events and the consequences of preservation actions on repositories and other preservation environments [16]. These predictions can be fed back into a watch system as a source of information, allowing the detection of possible risks before they actually happen. This is especially important for large-scale repositories where understanding of storage and computational resources is crucial.

Human knowledge. Finally, human users should be able to insert information about every possible entity (objects, format, tool, experiment, repository status, etc.).

All these sources will evolve and change. They can cease to exist, modify their behaviour or the way information is published, even the type of information or the way it is structured. The monitoring system should be designed to allow for this through a loosely coupled architecture of information adaptors. It allows the update, addition and replacement of sources, so that the configuration of complementary information sources can evolve over time.

We observe that these sources differ in their structure, ranging from highly structured linked databases to operational systems that raise events through log mechanisms. Furthermore, some of these drivers are internal, such as the operations specified by plans and the operational attributes of the system, while others are external. Attributes of the surrounding environment can influence plans, policies and operations [2]. Some sources can be both internal and external, since information internal for one organisation can

(in anonymised form) be of tremendous interest to another. For example, the format profile of a repository is internal to the organisation, but when shared, it can be used to assess whether a format is commonly used and can be considered a *de facto* standard.

5. A MONITORING SYSTEM

To collect, link and analyse information in the way described, a Watch component should aim for the following high-level goals.

1. **Enable a planning component such as Plato to automatically monitor entities and properties of interest.** Plans are created based on an evaluation of specific alternatives against formally modelled criteria [9]. A plan can thus cause a number of questions and conditions that can be tracked continuously to verify the compliance of operations to plans and detect associated risks and opportunities. The Watch component shall enable this tracking and support round-trip evolution of plans. We will discuss this in Section 6.
2. **Enable human users and software components to pose questions about entities and properties of interest.** Components and human users will be able to pose questions to the Watch component and receive answers about the measures. They can also deposit conditions to receive a notification upon significant changes.
3. **Collect information from different sources through adaptors.** Different sources will be relevant for the Watch component. Each source of information provides specific knowledge in different information models that will have to be mapped, normalized, merged and linked.
4. **Act as a central place for collecting relevant knowledge that could be used to preserve an object or a collection.** Information for object/collection preservation shall be collected and linked so that the Watch component provides a uniform reference point for gathering information about a variety of aspects.
5. **Enable human users to add specific knowledge.** While growing automation enables scalable data collection, human input is and will remain a valuable source of information.
6. **Notify interested agents when an important event occurs** through configurable notification channels.
7. **Act as an extensible platform.** This last item is particularly important: The Watch component is intended to function as a platform on which additional information sources can be added and connected easily.

Figure 2 shows the main building blocks of the Watch component, which is currently under development [13]. A number of external sources are monitored through a family of *adaptors* as outlined in Section 4. These correspond to an adaptor interface and deliver measures of defined and named properties of interest. A set of interfaces is defined to allow pulling information from the outside world, specifically used when the relevant sources remain agnostic to the watch service. These also serve to push information into the Watch component, used for example when sources of information need more control over what information is sent and how. Finally, a manual web user interface empowers users to send information when no automatic source is available. The extension of additional adaptors is supported by a dynamic plug-in architecture that relies on automated discovery of applicable information sources based on the questions posed.

Several adaptors can be linked together through a Monitoring Service which configures each adaptor and delegates information collection. Such adaptors extract information from a particular source, analyse and transform the information model if necessary and provide measures of specified properties of interest in a well-defined information model. The monitoring services can thus feed the collected information into the knowledge base.

The knowledge base presents a generic data model that is able to capture different measurements of relevant properties for digital preservation in an internal Linked Data store [10]. The data within the store represent a focused, curated part of preservation-related properties of the world. This can be used by queries and analyzed for occurrences of relevant properties, events and anomalies. All internal components make use of this internal model, which specifies a common language for data exchange [6].

To enable the structuring of information, the knowledge base data model must define two sets of elements. One, managed administratively, describes which model of the world is covered and accepted by defining which types of entities can exist and which properties are defined for them. Another, managed by the sources of information, describes instances of these entities and the values of their properties. The data model also keeps a register of the information provenance and history of changes to allow the traceability of information, which will transitively improve the traceability of the decision making process.

A key design decision concerns the representation of the collected information and how it should be linked. It is clear that the data layer of the Watch component must support the flexible model described for the knowledge base and provide the reasoning and querying features needed to answer complex questions about the world. The most appropriate technology for the representation of this model is clearly Linked Data implemented as a Triplestore. Relying on semantic models has strong benefits: It allows flexible integration and extension of the data model, while at the same time supporting inference and reasoning and ensures the extensibility of the underlying ontologies. An in-depth analysis of scalability issues of Linked Data concluded that the current state of the art in query performance is more than adequate for the estimated volume and throughput needed by the Watch component data layer [6].

6. MONITORING SCENARIOS AND EVENTS

To illustrate the effect of creating a component that addresses the goals envisioned, consider an organisation running a large digital repository. The organisation has connected the repository to a central deployment of the component (or created its own deployment based on the openly available code base). We can assume that the Watch component constantly monitors sources like format registries and component catalogues allowing users to pose questions about

[13] `https://github.com/openplanets/scape-pw`

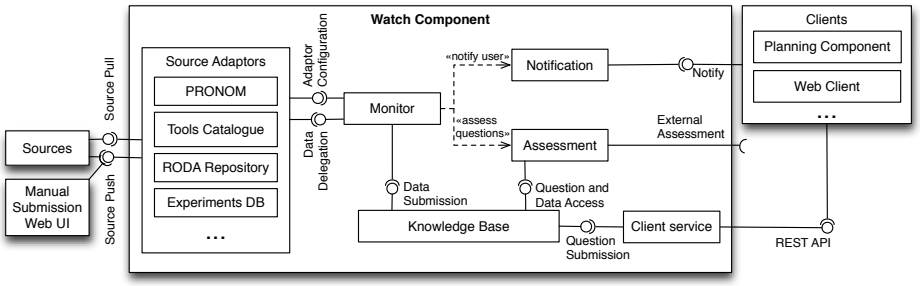

Figure 2: High-level architecture of the Watch component

different format or component properties.

There are two primary situations when a user would pose a question to the Watch component.

First, a typical case is the completion of a preservation plan as described in [3] and illustrated in [12]. A plan specifies a number of decision criteria, including format objectives, preservation action constraints and goals, and metrics specifying how to verify and validate authenticity of content upon execution of a preservation plan (such as a large-scale migration). The plan selects and specifies the best-performing action based on its real performance against these criteria. Upon deployment and operational execution of that plan, the organisation needs to verify that large-scale operations perform according to specifications (operational compliance). It also benefits from automated monitoring of potential risks and opportunities. For example, a large-scale experiment conducted by another organisation on content in the same format may uncover a systematic bias of a quality assurance tool when measuring image quality in TIFF-to-JP2 conversions with embedded color profiles. Such a bias is a potential risk to the authenticity of content that should be raised by a monitoring system. On the other hand, risk factors of format will change over time and should lead to appropriate events.

Second, a key feature of the proposed design is the fact that organisational policies and objectives can be linked to the information collected in monitoring, so that automated queries are able to resolve the question to which degree the current state of a repository matches the goals of an organisation. Hence, a set of standard conditions can be activated to perform such automated monitoring. As soon as organisational objectives change and are fed into the knowledge base through an appropriate adaptor, an automated compliance check can be performed.

What happens now in the event of a change detected by this monitoring process? As outlined earlier, an event raised by the Watch component can be assessed for its significance internally, according to its conditions, and externally, in its contextualised impact on preservation operations. Such an external assessment may consist in recalculating scores for alternative preservation strategies with updated information, which is fully supported by the utility approach followed by Plato [3]. If the assessment concludes that a change in operations is advisable, an iteration of the planning workflow will lead to a revision of the corresponding preservation plan and an update of the associated monitoring conditions. This leads to a continuous monitoring lifecycle of evolving plans and operations.

Table 3 summarizes these and other potential events. While space prohibits an in-depth discussion of all drivers, indicators and conditions, it can be seen that the fusion and interlinking of such diverse, yet related sources provides a powerful mechanism for monitoring. An in-depth discussion on these triggers and a full data model for specifying questions, conditions and triggers can be found in [6].

7. DISCUSSION AND OUTLOOK

Monitoring the preservation environment is a crucial part of the long-term viability of a system and the data it preserves. Monitoring supports the planning process with continuous tracking of the suitability of the decisions, delegating the risk assessment of the perceived significant changes back to planning. Such a monitoring system is essential for continued digital preservation.

So far, manual processes are used to track all the environment variables that might affect the multitude of object file formats within a repository, with all their different characteristics and contexts. Currently, no tool or service exists that could properly provide this function in a scalable way.

This document delineates the design and development of such a system, named the Watch component, based on the knowledge of past research and experience and going forward by defining new concepts and strategies. Based on real-world scenarios and an analysis of drivers and possible sources of information, we outlined the key sources to be monitored and specified the requirements and high-level design of a Watch component that is currently under development. This architecture enables organisational capability development through flexible and extensible services. The presented design supports automated elements of preservation management without constraining organisations to follow a specific model in their deployment of the planning capabilities. By modelling and implementing watch mechanisms, triggers, and suitable actions to be taken for each trigger, this system supports closed-loop preservation processes in which automated monitoring of collections, actions, plans, systems, and the environment triggers appropriate diagnosis and reaction.

Current work is focused on developing and connecting information source adaptors and providing API specifications that allow additional adaptors to be connected easily. Furthermore, the planning component Plato is being extended to support evolving lifecycles of preservation plans and provide automated assessment of accumulated changes against organisational objectives.

Table 3: From drivers to information sources: Exemplary preservation triggers[6]

Driver	Questions	Indicators	Example conditions	Sources
Content	Is there corrupted content?	Completeness validation fails	Log contains validation failure	Repository
		Access fails	Access failure event reported	Repository, User
	Is any content being ingested into the repository?	Ingest activity notices new content	Format of ingested content is different from content profile	Repository, Ingest, Collection profiler
	Is the content volume growing unexpectedly?	Rate of growth changes drastically in ingest	Growth of collection X exceeds threshold	Repository, Ingest, Collection profiler
New content	Which text formats appear in collection X?	New acquisition activity, new ingest activity	Mode (format) changes	Collection profile, Repository
Operations	Do we have plans defined for all collections?	Mismatch between content profile and set of plans	Exists collection with size greater than threshold defined in policy model without a plan	Repository, plans, policies
	Are our content profiles policy-compliant?	Mismatch between content profile and format/representation objectives	Content exhibits properties that are not acceptable (e.g. encryption)	Content profiles, policy model
Producers and Consumers	Are there any new or different producers?	New producer uses ingest process, new producers send new material	Exists new producer	Repository
	Are there any new or changed consumers?	New consumers use access process	Exists new consumer	Repository
Policies	What is the current valuation of collection X?	Change in policy model	Valuation changes	Policy model
Software	Which experiments have tested migration from TIFF to JPEG2000?	Evaluation data in an experiment platform	Exists new experiment with specified entities and properties	Experiment results
	Is there any software for converting TIFF to JPEG2000 that is applicable to our server environment?	New software matching certain criteria is tested within the experiment database	Matching migration component is tested on >10K objects on Linux platform without crashes or corrupt output	Experiment results, Component catalogue
Storage	What will be the content volume in 18 months? What will be the monthly storage costs?	Simulator prediction	In time X, the total size will be above a certain threshold	Simulator
Format	What is the predicted lifespan of format X?	Simulator prediction	The obsolescence time is below threshold	Simulator
New format risk	What is the risk status for the format X	New risk defined in the policy model	There is a new control policy about required format properties	Policy model

Acknowledgements

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

8. REFERENCES

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

[2] G. Antunes, J. Barateiro, C. Becker, J. Borbinha, D. Proença, and R. Vieira. Shaman reference architecture (version 3.0). Technical report, SHAMAN Project, 2011.

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

[4] S. Beer. *Brain of the Firm*, volume 1st ed. John Wiley & Sons, 1981.

[5] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1, 2002.

[6] K. Duretec, L. Faria, P. Petrov, and C. Becker. *Identification of triggers and preservation Watch component architecture, subcomponents and data model. SCAPE D12.1*, 2012.

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

[8] M. Ferreira, A. A. Baptista, and J. C. Ramalho. A foundation for automatic digital preservation. (48), July 2006.

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

[10] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*, volume 1 of *Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool, 2011.

[11] W. Kilbride. Preservation planning on a spin cycle. *DPC What's New*, 28, 2010.

[12] H. Kulovits, A. Rauber, M. Brantl, A. Schoger, T. Beinert, and A. Kugler. From TIFF to JPEG2000? Preservation planning at the Bavarian State Library using a collection of digitized 16th century printings. *D-Lib*, 15(11/12), November/December 2009.

[13] G. W. Lawrence, W. R. Kehoe, O. Y. Rieger, W. H. Walters, and A. R. Kenney. Risk management of digital information: A file format investigation. Technical report, Cornell University Library, 2000.

[14] D. Pearson. AONS II: continuing the trend towards preservation software 'Nirvana'. In *Proc. of IPRES 2007*, 2007.

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

[16] C. Weihs and A. Rauber. Simulating the effect of preservation actions on repository evolution. In *Proc. of iPRES 2011*, pages 62–69, Singapore, 2011.