Design and architecture of a novel preservation watch system

Luis Faria[1], Petar Petrov[2], Kresimir Duretec[2], Christoph Becker[2],
Miguel Ferreira[1], and Jose Ramalho[3]

[1] KEEP SOLUTIONS, Braga, Portugal,
{lfaria,mferreira}@keep.pt,
[2] Technical University of Vienna, Austria
{petrov,duretec,becker}@ifs.tuwien.ac.at,
[3] University of Minho, Braga, Portugal
jcr@di.uminho.pt

Abstract. Successful preservation of content requires sophisticated mechanisms for collecting, tracking and analyzing information about a multitude of relevant aspects. This is not limited to content itself, but also tracking of available software, other organization's content, usage statistics and trends, format risks, systems operations and many more. Such tracking requires a flexible system that supports evolution over time and provides an extensible platform for scalability.
This article presents a novel approach towards automated monitoring of preservation-related information. We discuss the challenges and information sources that need to be covered and outline the key design features of a novel preservation watch system. We discuss how this system addresses critical information needs for informed preservation management and outline next steps ahead.

Keywords: digital preservation, monitoring, watch, system

1 Introduction

Digital assets are continuously endangered by preservation threats that can hinder user access or even cause irreparable loss to the correctness or authenticity of valuable content. These threats belong to many distinct domains, from technological to organizational, economical and political, and can relate to the holder of the content, the producers or to the target communities to which the content is primarily destined for.

In digital preservation, monitoring is a key capability that enables the early detection of these threats [1]. However, as the volume and heterogeneity of assets increase, it becomes unfeasible to manually monitor all aspects of the world that may hinder their preservation. Furthermore, monitoring should not only detect symptoms of preservation risks but also opportunities (e.g. cost reduction) and ensure that preservation actions, as defined by decision-making processes, continue to meet goals and fulfill expectations.

Considering the problem scale, the automation of the monitoring process becomes a necessary step to ensure proper digital preservation. To enable the monitoring capability of an organization, there is a need to ensure automation through a system that gathers digital preservation information from several sources and curates it, i.e. maintains the information normalized and cross-referenced. The system deployment could be centralized to allow information sharing between organizations and create new synergies. This article will describe the key goals of such a system and define the architecture and underlying data model.

The next section will outline related background in monitoring. Section 3 describes the key goals to be achieved by a preservation watch systems and derives the main system requirements. Section 4 outlines the architecture of the system and discusses the data model and interfaces as well as information sources. Finally, Section 5 describes current work and outlines future steps.

2 Related work

The concept of preservation watch, or monitoring, has been presented in several reports [2,3,1]. However, there is a lack of research on how to properly achieve this capability and what methodology to use in order to systemize and automatize the monitoring processes. Currently, monitoring is done mainly by manual and ad-hoc procedures, depending greatly on tacit knowledge that the institution's teams are implicitly expected to have.

In the Open Archival Information System (OAIS) functional model, the monitoring functions are divided into *monitoring the designated community* - to detect and track changes in consumers and producers requirements, technologies and trends - and *monitor technology* - to detect "emerging digital technologies, standards, computing platforms" (software and hardware) and detect "technologies which could cause obsolescence" in digital archives [2]. Subsequent literature tries to categorize the information, i.e. the properties of the world, that are needed to watch for successful digital preservation. Previous reports have defined the domains of the needed information [3] - Administration, User Community, Organization, Producer, Technical Environment - and defined monitoring as a capability [1] which can be divided into internal monitoring - the ability to look towards inside, to the system and the operations specified be the preservation plans - and external monitoring - the ability to look towards outside, to the external influencers that might introduce risks.

Others have described information, tools and systems to aid the monitoring process [4], but these are limited to technical reports, format and tool registries, and format obsolescence methodologies and systems that try to identify what preservation risks are generically associated to file formats. For example, *Risk Management of Digital Information: A File Format Investigation* is a report on the impact assessment on file format migration and identification of the risks[5]. The *INFORM methodology* is a model to predict file format obsolescence by discovering threats to preservation and their possible impact on preservation decisions[6]. Other examples are *File Format and Tool registries*, which are online registries focused on digital preservation information about file formats, software products and other technical components relevant to preservation (like PRONOM[4], UDFR[5] and the Conversion Software Registry[6]). Finally, the *Automatic Obsolescence Notification Service (AONS)* is a software tool that provides a service for users to automatically monitor the status of file formats identified in their repositories against generic file format risks gathered from format registries and receive notifications [7].

These tools are limited by their lack of coverage and their focus on file format obsolescence, which is a subset of the preservation risks that might afflict digital content. Even in the limited set of information which focuses on file formats and tools that render, produce or convert them, other relevant information can be found in online file format and software versioning catalogues like FILExt[7], alternativeTo[8] and iUseThis[9]. These have less conventional information (in terms of digital preservation) but compensate with other types of information which can be relevant to digital preservation, e.g. social information. Furthermore, such general-domain communities are much larger than a preservation audience, which greatly improves the coverage of information and may reduce bias.

3 Automated Preservation Watch

To substantially improve automated support for effective digital preservation watch, the state of the art must evolve in two ways. Firstly, a tool that enforces a systematic monitoring methodology must be created and it must be able to scale to millions of objects. Secondly, the scope of collected information for digital preservation must be expanded to information from broader domains (organizational, financial, economical, technological, etc.) and any location or source, by automatic or manual means. At the same time, this requires support for normalization and structuring rules on the gathered information. Having information from different domains then allows a user to pose questions that embodies his own interpretation of preservation risks and opportunities of interest. Examples of these questions are:

- Are there any new tools that can render the format X?
- Is there any content in my repository in a file format that is not kept in any other available repository?
- Has the content volume to be ingested by my repository suddenly increased?
- Is there negative user feedback on object renderability in my repository?
- Are there experiences with tool X that detected an unsatisfactory behaviour?
- What is the current price per GB of data storage?

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

[5] http://udfr.org

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

[7] http://filext.com

[8] http://alternativeto.net

[9] http://iusethis.com

- Is there a lossless file format for images for which experiments show it needs less storage space than the one I currently use?

Such a preservation watch system should strive for the following goals:

1. Collect information from different sources automatically with adaptors;
2. Enable human users to manually add specific knowledge;
3. Act as a central place to collect knowledge relevant for digital preservation;
4. Enable humans and software components to pose questions on entities or properties of the world;
5. Notify interested agents (human or software) of significant events;
6. Enable the monitoring of the validity or appropriateness of defined preservation plans by detecting significant changes on the properties of interest in which the decisions where based upon;
7. Allow to extend functionality by easily adding support for new sources of information and new possible questions.

The system would allow the identification of risks and opportunities by assessing centrally gathered information. This enables optimization of the of needed deep assessment, made by specialized tools, like Plato [8], which can explicitly identify risks and opportunities, discard false positives, and select actions to mitigate risks and take advantage of opportunities.

From defined use cases and scenarios [9], different classes of information sources where identified to be integrated with the proposed system. Other classes of sources might be added in the future as necessity arises.

Format registries are the foremost regarded information in current preservation watch and will continue to be gathered with this system. For example, the PRONOM adaptor is already under development.

Software catalogues are online services with information on tools that render, migrate, analyze and compare files of diverse file formats. This catalogues can be digital preservation specific, like the TOTEM[10] and the SCAPE Component Catalogue (under development), or generic-domain software catalogues.

Digital repositories have information on producer activity and the user community access preferences and problems, which certainly concern the repository owner. Also, when aggregated and viewed as a whole, this information can provide insight on the global tendencies and reveal *de facto* standards.

Content profiles provide an aggregate view on the content characteristics and metadata, specially of the technical type. When applied to digital repositories and web archives, content profiling can provide precious information needed for preservation. If this information is shared with the community and viewed as a whole, it allows valuable insight on the wider state of curated content and can serve as an up-to-date indicator revealing technology and usage trends.

Experiments with tools, like migrators, assessing their behavior, reliability, completeness and quality are usually made as part of the preservation planning process. The outcome of these experiments can be of much interest to other users that are considering using that tools with similar objectives.

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

Organizational objectives. On-going work on formalizing preservation control policies and their relationship with organization strategies and goals will enable monitoring of some of the organizational objectives and check the repository compliance.

Simulation. Based on the data gathered by watch, models of digital repositories can be created to predict the consequences of preservation actions [10], which allows the inclusion of forecasts into the knowledge base and be alerted of preservation risks before they actually happen.

Human knowledge. Human users should be able to add information of any domain into the knowledge base. Some of the knowledge needed for digital preservation is still tacit or unstructured. Having humans as a source of information would allow the watch system to act as the central place for any kind of knowledge relevant for digital preservation to be gathered and formalized, even when there is no other specialized platform to support it.

4 A system design for preservation watch

A proposed three-tier architecture for such a system is depicted in Fig. 1. Information comes from the outside world via source adaptors on the interface tier (top layer). Pull Source Adaptors fetch information from external sources of information and filter, normalize, structure, aggregate and anonymize the gathered information. Adaptors can also run in the external source, pushing the information into watch via the Push Source Adaptor API. New source adaptors can be added to the system at any time. Every source adaptor structures the gathered information to the way it is modeled in the Knowledge Base and delegates is to the Data Enrichment Service on the business logic tier (middle layer).

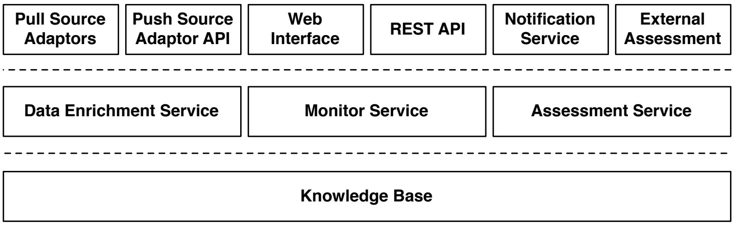

Fig. 1. High-level architecture of the Watch component

As different sources might refer to the same entities and properties of the world, incoherences and incompatibilities between different sources of information are to be expected. For example, format registries that sometimes serve contradictory information about file formats. The Data Enrichment Service solves this by merging the data and treating incoherences before updating the Knowledge Base, in the data tier (bottom layer). This service also cross-references the information of different entities and properties. Cross-linking is very important to enable question that relate and interlink several entities. For example, the file format distribution of a repository, given by its content profile adaptor, should be linked with the file format entities given by the format registry adaptors.

On the bottom, the Knowledge Base defines the data tier that keeps structured information about entities and properties of interest of the world. The Knowledge Base allows cross-reference of information and provides a query engine that is expressive enough to cross-examine the different properties and define conditions that identify symptoms of risks or opportunities.

Back to the top layer, the Web Interface serves as the human interface for manually adding information or to allow humans to browse the knowledge, pose Questions and create Conditions so they are notified when significant events occur. The REST API allows the same functionalities to software components.

These Questions and Conditions are kept in the Knowledge Base. The Monitoring Service determines when Questions and Conditions need to be re-assessed, by scheduling frequent checks or monitoring the state of the Knowledge Base.

The assessment is done by the Assessment Service. If the result contradicts the conditions the interested parties are alerted via the Notification Service.

Additionally, the External Assessment allows for more complex and deep assessment systems, like the Plato planning component, to be directly used in the assessment phase, allowing for a even greater optimization of the system by minimizing false alerts (false-positives).

The next sections describe the details of how the information flows through the various components and how it is kept in the Knowledge Base.

4.1 Knowledge base

As information about the world comes from different external sources it may be represented and structured in various ways. But, to allow a consistent and useful Knowledge Base, the restrictions on the way the data is formatted and structured must be enforced by the system. This requires that a model of interesting aspects of the world be created within the Knowledge Base, forcing all added information to conform to this model. Furthermore, the aspects that are of interest may change in time, so the model must be adaptive and able to grow.

To be able to represent any structure of information, the most suitable model is the ontological, where an aspect of the world can be represented by an **Entity**, i.e. something with distinct and independent existence, described by a series of Properties.

The proposed data model restricts the created Entities to belong to a certain **Entity Type**, which defines the class or facet of the world the related Entities describe, grouping Entities together and constraining the domains they belong to. Examples of Entity Types might be File Formats, Tools, Experiments, etc.

The Entity Types restrict the Properties that a Entity might define as the Entity can only define a **Property Value** for a **Property** defined on its type. A Property is, therefore, the generic definition of an attribute that an Entity of a determined type can be described with.

This method of restricting the data model allows for the model of the world to be constantly updated, increasing its expressive power whenever it is needed for answering questions about properties that become relevant.

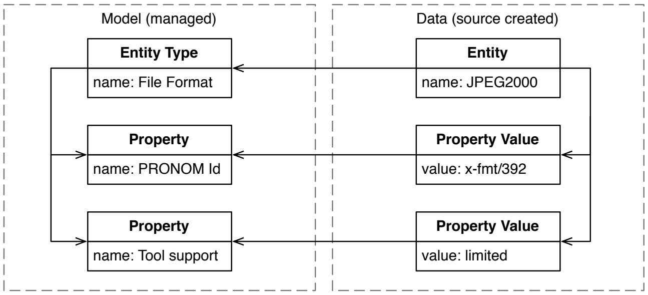

Fig. 2. Example of managed model of the world and source created information

Figure 2 illustrates a simplified example of entries in the Knowledge Base. There can be an administratively created 'File Format' entity type, which defines 'PRONOM Id' and 'Tool support' properties. This means that any entity of the 'File format' type can define the 'PRONOM Id' and 'Tool support' properties[11]. The 'File Format' type is related to a 'JPEG2000' entity, which 'PRONOM Id' is 'x-fmt/392' and 'Tool support' is 'limited'. Several other instances of the 'File Format' type might exist and may be created different source adaptors, but all must conform to the same controlled set of properties and all are inter-connected.

But having the current state of the world is not enough as values change in time and their history can reveal trends. Therefore, it is important to keep the different values that a property takes throughout time and to register the moment when the property was measured. Furthermore, to allow traceability of the information, the provenance of values must also be registered, defining which source adaptor and external source took the measurement.

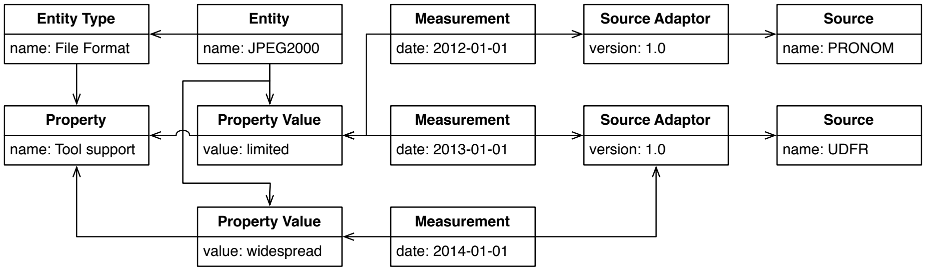

Fig. 3. Example of Knowledge Base with history and provenance

An example of the knowledge base with value history and provenance is depicted in Fig. 3. The value of 'Tool support' property changes throughout time, being 'limited' in 2012 and 2013 but becoming 'widespread' in 2014. The measurements are taken from different sources but they are represented the same way. The measurement of 'limited' tool support was first taken in 2012 by PRONOM and then confirmed in 2013 by an UDFR. In 2014, the same UDFR adaptor detects a change of the tool support property to 'widespread'.

The current Knowledge Base already defines a model of the world with many diverse entities and properties of interest, not listed here for sake of brevity.

[11] Properties can prescribe the format of the value, ignored here for sake of simplicity

4.2 Questions, conditions, notifications and triggers

A Question is a query on the information gathered in the knowledge base that tries to identify symptoms of preservation risks or opportunities. These questions relate to entities and properties and may need to cross-reference them. Section 4 example questions use cross-referenced information between entity types (like formats, tools and experiments) and make queries to entities or their properties. Also, the questions are about the current state of the entities and properties or about how they evolve throughout time (e.g. sudden growth in content volume).

Not all changes in the output of the questions are important. One would not want to be alerted when, for example, the price per GB of storage mildly fluctuates some cents. So, the limits of when an event becomes significant, and therefore needs attention, must be defined via Conditions. Conditions are pieces of algebraic or boolean logic (e.g. thresholds) that define when the result of a question, i.e. an event, becomes significant and needs attention.

When a significant event is detected all defined interested parties should be notified. A Notification defines an alert that is sent to an external user or software component (e.g. email, HTTP API). Types of notification are extensible so it easily adapts to the user community needs.

In sum, a Question gives a result which is evaluated by a Condition that, when it becomes true, sends Notifications. The Monitoring Services can frequently reassess the question and condition, with a scheduler, or react only when the entity types and properties that might affect the question are updated.

4.3 Information source adaptors

Any source of information can be integrated with the watch system by the creation of an adaptor. This Source Adaptor will transfer the information from the external source into the watch system, transforming the information in the process so it fits the watch model of the world.

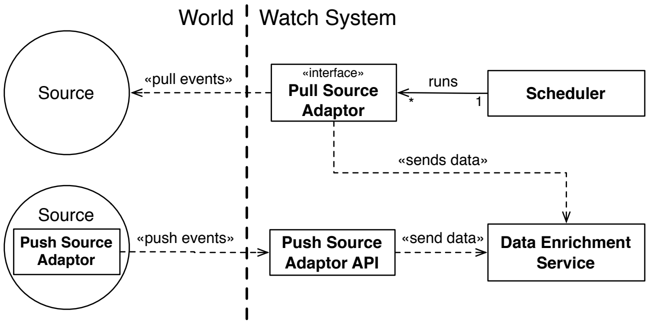

Fig. 4. Source Adaptor model

A Source Adaptor is a software component that can be deployed in two ways: inside the watch system, frequently polling the external sources for relevant information (Pull Source Adaptor), or inside the external sources (Push Source Adaptor), pushing information into the watch system via the Push Source

Adaptor API (Fig. 4). The Pull Source Adaptor must conform to a defined interface and will be run automatically by a Scheduler. The Push Source Adaptor freely runs on the external source but must conform with the API specification. The push and pulled data is finally sent to the Data Enrichment Service to be merged and cross-linked before going to the Knowledge Base.

The source adaptor must filter the data of interest of the external source, aggregate data, analyze and infer new knowledge when possible, map and normalize data to the model specifications and anonymize personal data. Depending on the external source, a push or pull source adaptor might me more adequate. For example, if the external source needs to be agnostic of watch, like a file format registry or a generic software catalogue, then a pull source adaptor must be used. If, on the other hand, there are privacy issues on the data, as is common in repositories, a push source adaptor is more appropriate as it will allow to anonymize the information before sending it to watch.

For human knowledge, a specialized component will be created to allow users to add knowledge on a web page, which will serve as a push source adaptor.

5 Conclusions and outlook

Preservation watch is a crucial part of the digital preservation process and, at the same time, is one of the most underdeveloped processes in this domain. Current practice is mainly manual and ad-hoc and cannot keep up with the exponential growth of content in volume and diversity. One limitation is the focus on technological obsolescence, ignoring risks and opportunities that can stem from other domains, like organizational or economical. Also, there is a lack a defined approach on how to accomplish the monitoring capability. Finally, no platform or system currently exist to assist with the monitoring process.

This document describes a novel approach for preservation watch and a system that applies it. This system is currently under development as a open-source project[12]. The approach consists in gathering information from various external sources of diverse domains, creating a centralized knowledge base with information of interest for preservation. Then express preservation risks and opportunities as questions to this knowledge base. Finally, monitor the result of question assessment to reveal significant events that indicate the existence of the defined risks and opportunities. The system focuses on the automation of the approach, so it scales in terms of volume and heterogeneity.

One early result of the system is the generation of large-scale content profile statistics of 80 million objects from the Danish web archive to create statistics that allow insight into format and feature distribution and trends over time. This will be interlinked with extensible format information. At the same time, we are developing semantic models for organizational objectives that can be directly fed into the knowledge base to allow direct application of query technologies such as SPARQL on a growing knowledge base of linked data elements. Also, repository

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

integration is being developed for RODA[13], Rosetta[14], eSciDoc[15] and any other that follows the defined API. Tools are being developed to allow scalable content profile of web archives and their integration with the watch system. The integration with Plato is currently being designed, so questions are automatically derived from plans, allowing for the preservation actions to be continuously monitored to ensure they continue to meet expectations. Finally, the web interface will allow any user to query and monitor all published information, define their interests and be notified when risks or opportunities appear.

Acknowledgements

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

References

1. Antunes, G., Barateiro, J., Becker, C., Borbinha, J., Proença, D., Vieira, R.: Shaman reference architecture (version 3.0). Technical report, SHAMAN Project (2011)
2. Consultative Committee for Space Data Systems: Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1 (2002)
3. Sierman, B., Hofman, H., Thaller, M.: Report on the Planets Functional Model. Technical report, Planets Project Deliverable PP/7-D3+D4 (2009)
4. Ferreira, M., Baptista, A.A., Ramalho, J.C.: A foundation for automatic digital preservation. Ariadne (48) (July 2006)
5. Lawrence, G.W., Kehoe, W.R., Rieger, O.Y., Walters, W.H., Kenney, A.R.: Risk management of digital information: A file format investigation. Technical report, Cornell University Library (2000)
6. Stanescu, A.: Assessing the durability of formats in a digital preservation environment: The INFORM methodology. OCLC Systems & Services **21**(1) (2005)
7. Pearson, D.: AONS II: continuing the trend towards preservation software 'Nirvana'. In: Proc. of IPRES 2007. (2007)
8. Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., Hofman, H.: Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. International Journal on Digital Libraries **10**(4) (2009) 133–157
9. Duretec, K., Faria, L., Petrov, P., Becker, C.: Identification of triggers and preservation Watch component architecture, subcomponents and data model. SCAPE D12.1. (2012)
10. Weihs, C., Rauber, A.: Simulating the effect of preservation actions on repository evolution. In: Proc. of iPRES 2011, Singapore (2011) 62–69

[13] http://roda-community.org

[14] http://www.exlibrisgroup.com/category/RosettaOverview

[15] http://www.escidoc.org