



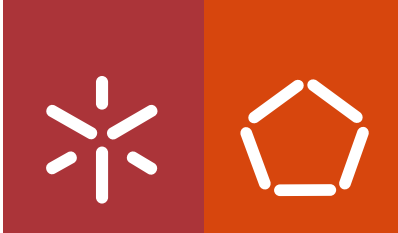
**Universidade do Minho**  
Escola de Engenharia

Luis Francisco da Cunha Cardoso de Faria **Automated Watch for Digital Preservation**

Luís Francisco da Cunha Cardoso de Faria

## **Automated Watch for Digital Preservation**





**Universidade do Minho**  
Escola de Engenharia

Luís Francisco da Cunha Cardoso de Faria

## **Automated Watch for Digital Preservation**

Doctorate Thesis  
Doctoral Program on Informatics

Work under the supervision of  
**Dr. José Carlos Leite Ramalho**  
**Dr. José Miguel Araújo Ferreira**

October 2017

## DECLARAÇÃO

Nome: Luis Francisco da Cunha Cardoso de Faria

Correio eletrónico: lfaria@keep.pt

Título Tese: Automated Watch for Digital Preservation

Orientadores: Professor Doutor José Carlos Leite Ramalho e Professor Doutor José Miguel Araújo  
Ferreira

Ano de conclusão: 2017

Designação do Doutoramento: Programa Doutoral em Informática

É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA TESE APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO QUE A TAL SE COMPROMETE.

Universidade do Minho, 26/10/2017

Assinatura:                     Luis Faria

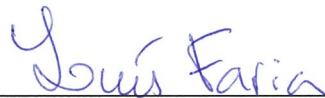
## STATEMENT OF INTEGRITY

I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration. I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, 2017-10-26

Full name: Luís Francisco da Cunha Cardoso de Faria

Signature:



---



# Acknowledgements

Here I leave my deepest gratitude to all that have contributed to this thesis:

To my supervisors, for their support on these long years of work and for their contribution to the thesis and to my personal growth. To Miguel Ferreira, my mentor and role model, for the projects we have worked side-by-side, for all the stories we have shared, for all that he has taught me in all aspects of life, I cannot thank enough. To José Carlos Ramalho, my cherished professor, for his continuous guidance and support, my many thanks.

To KEEP SOLUTIONS, the company that co-funded the presented work and allowed me to follow my research objectives. To my colleagues, Hélder Silva, Rui Castro and Sébastien Leroux, which work has contributed much to this thesis and whose support allowed me to follow it through. To all my company colleagues, for their support and their fellowship.

To the SCAPE project, and the European Union, who also co-funded the presented work. To all SCAPE partners, who have taught me plenty. And a very special thanks to all colleagues in the Technical University of Vienna, to Christoph Becker, Krešimir Duretec, Petar Petrov, Artur Kulmukhametov and Michael Kraxner, no one could ask for better research partners, nor for better friends, to you all my deepest gratitude. I would also like to thank Alan Akbik from the Technical University of Berlin, for his contribution on the use of information extraction technologies, to Barbara Sierman and Marcel Ras from the National Library of the Netherlands, for their insight into the e-journal archiving issues, and to Per Møldrup-Dalum from the Danish State and University Library, for all his great work on web archiving deep characterisation and for being such a nice guy.

Finally, I would like to thank to Andreia, my dear wife, the joy of my life, that has endured so much while I have worked on this thesis, and who also contributed greatly in making it better. To Tiago, my brother in law, who relentlessly reviewed this thesis, and who is a source of inspiration for his pursuit of perfection. To Nuno Vale, my oldest friend, for all his effort on teaching me applied statistics. To Susana and Diogo Pinheiro, for warning me of the caveats of a doctoral thesis and for their friendship. To my parents, for supporting me on following my dreams, and to my family, my sisters and my nephews, for always being there for me.

*To all of you, I am forever grateful,  
Luís Faria*





# **Abstract**

## *Automatic Watch for Digital Preservation*

The current extensive growth of digitally created documents is an obvious effect of the global tendency towards the digital technology. Replacing paper with digital documents has become a common activity in all kinds of public institutions and businesses, on which many already completely eradicated the use of paper and other analogue media. European policies, such as eGovernment, urge for the public administration to cease the use of analogue media and provide all services and documentation in digital form.

But documents in digital form are much more perishable than their paper counterparts and it is not obvious for the users that keeping a digital document accessible for several decades is a very different task than safekeeping paper-based documents. Furthermore, some aspects that a user will consider maintained when keeping the physical form of the paper do not behave the same way when the information is in digital form. Authenticity is one of these aspects, and it is crucial in some domains where the information has no value to be kept if the power to serve as evidence is lost. The digital preservation field tries to tackle all these problems.

The main problem in digital preservation relates with the ever-changing technological environment with which the documents must be compatible to be accessible by users. Part of the solution must include the detection of these changes by continuously monitoring the documents, their users, and the technological, organisational, financial, legal, social and even political environment, to detect preservation threats. However, the volume and heterogeneity of documents make manual monitoring of all relevant aspects of the world infeasible. Moreover, current practice is limited to monitoring influencers of a very reduced set of domains, specifically file format obsolescence, ignoring others that might introduce threats.

This work focuses on creating automatic and systematic ways to monitor the environment on a wide set of domains and provide a valuable input for digital preservation threat detection and assessment. It does so by inquiring the community about their view on the preservation threat importance and the methods used to effectively detect and monitor the threats. Then, an approach for automatic threat detection is laid out and implemented, focusing on the most important and neglected threats. Finally, the approach is validated against real world problems, proving to be a successful approach for scalable and automatic preservation watch.



# Resumo

## *Automatização da Vigilância de Preservação Digital*

O elevado crescimento do número de documentos criados digitalmente nos últimos anos, é um claro efeito da atual tendência global para a utilização da tecnologia digital. A substituição do papel pelo formato digital nas instituições e empresas tornou-se comum, sendo certo que algumas delas erradicaram já a utilização do papel e outros suportes analógicos do seu quotidiano. As políticas europeias que têm vindo a ser adotadas — tal como o Governo Eletrónico — incentivam a administração pública a abandonar a utilização de suportes analógicos, substituindo-os pelo formato digital, tendo as entidades públicas passado a prestar os seus serviços e a disponibilizar a documentação de forma eletrónica.

Todavia, os documentos em formato digital são mais efémeros que os seus análogos em papel, não sendo, contudo, óbvio para o utilizador comum que a preservação de um documento digital por dezenas de anos seja uma tarefa muito diferente da conservação de documentos em papel. Na verdade, algumas características que o utilizador comum considera implicitamente preservadas ao conservar a forma física do papel, não permanecem do mesmo modo quando a informação é mantida em formato digital: a autenticidade é uma dessas características, sendo crucial nos domínios em que o valor da informação é proporcional ao seu valor probatório.

O principal problema da preservação digital prende-se com a volatilidade do ambiente tecnológico e com o qual os documentos necessitam de manter compatibilidade. Parte da solução deverá passar pela monitorização destas mudanças, através da vigilância contínua dos documentos, dos seus utilizadores e também do ambiente tecnológico, organizacional, financeiro, legal, social e até político, de modo a detetar quais as ameaças à preservação dos documentos.

Contudo, o volume e heterogeneidade dos documentos digitais tornam impraticável a monitorização de todos os fatores externos relevantes para a preservação digital. Aliás, atualmente a monitorização é limitada à análise de influências pertencentes a um conjunto reduzido de domínios, especialmente ligados à obsolescência de formatos, ignorando outros que podem revelar ameaças à preservação dos documentos.

O presente trabalho dedica-se ao estudo de mecanismos sistemáticos e automáticos de monitorização do ambiente num conjunto alargado de domínios e de modo a fornecer a informação necessária para a deteção e avaliação das ameaças à preservação digital. Primeiramente, a comunidade é inquirida sobre a sua perspetiva quanto à importância das várias ameaças à preservação digital e quais os métodos utilizados para detetar se tais ameaças afetam o conteúdo digital. De seguida, é apresentada uma nova abordagem para deteção automática de ameaças à preservação do conteúdo digital, focalizada nas ameaças mais importantes e negligenciadas. Finalmente, a nova abordagem é validada perante cenários reais, provando, assim, ser uma proposta viável de monitorização automática para a preservação digital.



# Contents

<b>1 Introduction</b>	<b>1</b>
1.1 Motivation	1
1.2 Objectives and contributions	2
1.3 Document structure	3
<b>2 Digital preservation</b>	<b>5</b>
2.1 Structure of a digital object	6
2.2 Preservation of the physical level	8
2.3 Preservation of the logical level	9
2.4 Preservation of the conceptual level	13
<b>3 Preservation watch</b>	<b>15</b>
3.1 Watch concept across domains	15
3.2 Watch in digital preservation	16
3.3 Watch and risk management	19
<b>4 Context and methodology</b>	<b>23</b>
4.1 SCAPE: Scalable Preservation Environments	23
4.2 SCAPE preservation life-cycle	25
4.3 SCAPE preservation suite	26
4.4 Research questions	29
4.5 Approach	30
<b>5 Survey on preservation watch concerns</b>	<b>31</b>
5.1 List of questions	31
5.2 Results	35
5.2.1 Sample	35

5.2.2	Preservation threats	37
5.2.3	Preservation threats detection methods	40
5.3	Analysis and discussion	44
5.3.1	Sampling bias	44
5.3.2	Preservation threats priority	45
5.3.3	Answering the first research question	48
5.3.4	Gaps on the preservation threats detection methods	48
5.4	Final remarks	51
<b>6</b>	<b>SCOUT - A preservation watch system</b>	<b>53</b>
6.1	Requirements	55
6.2	Architecture	55
6.2.1	Knowledge base	56
6.2.2	Information sources	59
6.2.3	Information source adaptors	61
6.2.4	Extensibility via plugins	62
6.2.5	Collecting information	63
6.2.6	Optimising value and measurement representation	64
6.2.7	Questions, conditions, notifications and triggers	65
6.2.8	Assessing triggers to identify threats	66
6.3	Implementation	67
6.3.1	Monitoring content	67
6.3.2	Monitoring the environment	76
6.3.3	Setting up information source adaptors	79
6.3.4	Adding institutional control policies	79
6.3.5	Detecting and monitoring threats	81
6.3.6	Planning and operations	83
6.3.7	Monitoring preservation actions	89
6.4	Final remarks	90
<b>7</b>	<b>Evaluation</b>	<b>91</b>
7.1	Experiment 1: SCAPE preservation suite	91
7.1.1	Scenario	91
7.1.2	Execute experiment	93

7.1.3 Results	97
7.2 Experiment 2: Checking external references	100
7.2.1 Scenario	100
7.2.2 Execute experiment	101
7.2.3 Results	103
7.3 Experiment 3: Web archive deep characterisation	106
7.3.1 Scenario	106
7.3.2 Execute experiment	106
7.3.3 Results	111
7.4 Experiment 4: e-Journal archive services completeness	115
7.4.1 Scenario	115
7.4.2 Execute experiment	117
7.4.3 Results	120
7.5 Final remarks	122
<b>8 Conclusion and future work</b>	<b>125</b>
8.1 Summary	125
8.2 Conclusions	127
8.3 Contributions	130
8.4 Future work	132
<b>References</b>	<b>134</b>
<b>A Appendix</b>	<b>141</b>
A.1 Survey questionnaire	141
A.2 Scout knowledge base complete UML class diagram	153
A.3 Taverna workflow for a preservation action plan	154
A.4 Report API output examples	155





# Chapter 1

## Introduction

This chapter presents the research motivation, the objectives it aims to achieve and the structure of the rest of the document.

### 1.1 Motivation

Advances on digital technology have driven governments, institutions and businesses to encourage the creation and usage of assets in a digital form. Nowadays, most assets are born digital and many older ones are being digitised for preservation and access reasons (Joint, 2008), while others can only be represented in digital form (e.g. computer games, 3D media, etc.). These assets belong to a wide spectrum of domains, from key business information to irreplaceable cultural heritage, and present many preservation challenges. Problems can be related to hardware (e.g. degradation of digital storage media), technological obsolescence, loss of social and cultural context, and the capability of the institution to maintain the assets may be afflicted by economical, organisational or even political issues.

Digital preservation refers to the sum of activities (procedures, standards, best practices and technologies) necessary to ensure the long-term access to digital information. Generally, preservation threats can occur whenever an internal or external factor hinders the correct access to the information. Digital Preservation processes try to detect these threats, plan a course of action and act in order to mitigate the problem. This is a continuous process because the environment is continuously changing and the factors that influence the course of action can vary, giving rise to new threats or altering the adequacy and efficiency of the previously chosen actions. Therefore, there is a need to frequently monitor the internal and external factors that can affect the correct and continuous access to the digital information.

However, as the volume and heterogeneity of assets increases, as well their related threats, it becomes impractical to manually monitor all aspects of the world that may jeopardize their preservation. Considering the scale of the problem, the automation of the monitoring process becomes a necessary step to ensure proper digital preservation. To enable the preservation monitoring capability of an organisation, there is a need to ensure automation through a system that gathers digital preservation information from several sources and curates it.

Moreover, monitoring for successful digital preservation involves different capabilities and roles, from the most abstract organisational level to the furthest concrete technological level. Having such a wide and thorough view of the environment manually is impractical but crucial for decision-making processes on preservation planning. Being able

to monitor the suitability of the actions selected by decision-making processes is critical to maintain the continuous improving cycle of adaptation to the exterior environment problems and needs, preserving the continuous access and value of the digital information.

## 1.2 Objectives and contributions

The main objective of this work is to model and implement an approach for effective preservation monitoring. This approach should address the community current needs and allow to find preservation problems in digital content to initiate the digital preservation processes that will mitigate them.

This research creates an approach to preservation monitoring with the following drivers:

- Minimise human intervention via automation of processes;
- Improve scalability in terms of volume and heterogeneity of documents and external influencers;
- Allow common and neglected monitoring threats to be detected;
- Provide input for external risk assessment and decision-making processes.

This research identifies that the previous approaches for automatic preservation monitoring (such as AONS) failed because they tried to automatically gather preservation risks (Pearson, 2007), where no complete source for digitally formalised preservation risks currently exists. Some preservation risks may be found on format registers like PRONOM, but they are quite scarce. Also, some preservation risks can be found in technical reports, but they are not formalised in a way they can be automatically processed. Therefore, this research proposes a new approach: to gather enough knowledge about the world and formalise it, enabling risk assessment to be done by human users or software tools.

This input to risk assessment can be represented as the identification and detection of digital preservation threats, i.e. potential events with negative consequence on the preservation<sup>1</sup> of digital content. Examples of the digital preservation threats are:

- A fire, flood, hurricane, bomb, or any other catastrophe;
- Storage media degradation, bit rot, accidental or voluntary deletion or overwrite, cyber attack, etc.;
- Currently available tools no longer support rendering of a format;
- There is content in the repository in niche file formats (e.g. not supported by any other available repository);
- Content volume to be ingested by the repository has suddenly increased beyond capabilities;
- There is negative user feedback on object renderability in the repository;
- There are experiences with a used tool that detected an unsatisfactory behaviour.

---

<sup>1</sup>In this context, the term preservation refers to the content owner's preservation objectives, which are some times defined as the long-term access service provided to a target community.

A great number of digital preservation threats might exist, therefore this work focuses on identifying the most important and neglected ones and discovering whereas the above approach can be applied to them. This can be formalised on the following research questions:

**Question 1.** *What are the most important and neglected digital preservation threats?*

**Question 2.** *Can these threats be detected using a formal representation of the information about the world that is automatically monitored and collected?*

More information about the research questions and the approach to solve them is available in chapter [4](#)

This work aims to contribute to the digital preservation research field by improving current monitoring methodologies to solve the challenges posed by the increasing number of digital assets, their heterogeneity and their related external influencers. This work may also influence other research areas where changes in the external environment have an impact on day to day business operations.

## 1.3 Document structure

This document is organised in eight chapters:

The first chapter presents the introduction to the thesis, briefly describing the motivation for the work and the objectives and contributions it aims to achieve. This chapter identifies the problems that drive the research and sets up the goals to be accomplished on this work.

The second chapter provides context by briefly explaining several concepts on the subject of digital preservation necessary for the understanding of the presented work. This chapter explains why digital preservation is important, what impact does it have in the world, what problems does it try to solve and in which way. It also describes the current state of the art and the current gaps, specially focusing on preservation watch.

The third chapter delves into the subject of preservation watch, explaining how the concept applies across domains and how it has evolved on the digital preservation literature. It further describes the current state of the art approaches to the subject and their main disadvantages, gaps and reasons for failure. This chapter also provides a proposal for an alignment of the concepts with the risk management domain.

The SCAPE project, where much of the work presented on this thesis was done, is presented in chapter [4](#). The chapter presents the project mission, the partners, and the SCAPE preservation life-cycle, an architecture that integrates the presented work on an encompassing solution for digital preservation. It also presents the methodology by identifying the research questions and defining the approach to tackle them. The approach includes the survey presented on chapter [5](#), the software artefact named Scout presented in chapter [6](#), and the evaluation of Scout by means of real-world experiments presented in chapter [7](#).

Finally, the document ends with a description of the conclusions that can be extracted from this work, the main contributes it provides, and some remarks and future work, in chapter [8](#).



# Chapter 2

## Digital preservation

Although the digital preservation research field is already decades old and focuses on very common problems of the present world, the term is a bit obscure when not familiar with the information management field. Digital preservation gravitates around the **digital object**, which is *an information object, of any type of information or any format, that is expressed in digital form* (Thibodeau, 2002). Here are examples of digital objects in many domains:

- In science there are journal articles and the associated research data, necessary for reproducibility of the results;
- In medicine there are X-rays, 3D computer axial tomography scans and even the patient records;
- In engineering there are 2D and 3D models of buildings, diagrams of electric components, source code, models of cars, planes and ships;
- In art there are books, audio masters, digital photography and multimedia art;
- In industry there are movies and computer graphics animation, telecommunication, radio, television, bank records;
- In digital heritage and public record institutions there are digitisation of cultural artefacts, public administration databases, political campaign web sites and the prime minister social network web page;
- In the domestic context there are wedding videos, children pictures and vacation mementos that are often priced as irreplaceable. There is also personally acquired and copy-protected digital music and movies.

There are obvious reasons for the success of digital media: it makes it easy to create, edit, copy, transmit, and publish information but requires little physical space to be stored. However, digital information has a terrible disadvantage, it needs a proper technological context to be correctly consumed. While an analogue object can be readily consumed, for example a paper journal can be picked up by anyone and directly read and understood<sup>1</sup>, the digital object needs

---

<sup>1</sup>Although to read a paper journal one would need to be physically capable of reading the text, understand the language in which is written and having some basic knowledge on the subject of the journal to be able to understand its content. But these are all problems that are shared between the analogue and digital objects.

a complete technology stack so it can be properly rendered. An online version of the same journal needs some hardware device which contains a screen and can form images that can be physically perceived by the human eye. The information of the screen is provided by a software reader, for example a web browser, that understands the format in which digital object is encoded and interprets it. This software reader needs a whole operative system to be able to correctly operate the hardware. But the digital version of the journal might not be originally on the rendering system, it may have been brought unto it via the internet, a very complex and intricate set of technologies that include software and hardware infrastructures scattered around the world. Ultimately, the source digital object would be on a data centre, on one or several servers, and lays on one or more hard drives in a physical arrangement too intricate to detail here. Due to this complexity, a digital object is often divided into three levels (Thibodeau, 2002):

1. **Physical:** The representation of the object as an inscription of signs on some physical medium, e.g. a hard-drive, a DVD or a flash drive.
2. **Logical:** The binary coding of the information, the file format, which is interpreted by a software reader; e.g. a file in Portable Document Format (PDF)
3. **Conceptual:** The tangible unit of information that can be recognised and understood by a person, e.g. a journal, a book or a photo.

If any of the levels is damaged, the access to the next is endangered and we can loose access to the conceptual object. For example, if a bit on a hard-drive randomly swaps due to bit rot (Baker et al., 2005), the capability of the software reader to correctly interpret the file is endangered and the conceptual object can be lost. The tolerance for failures depends on the file format, the software reader and even on the conceptual object. If information entropy<sup>2</sup> is large, and therefore redundancy is low, such as in compressed data, then the probability of information loss is high (Shannon, 1948). Also, if the software reader is very strict on the interpretation of the file format, the probability of read failure is also greater. In like manner, the conceptual object may be so intellectually dense that the meaning cannot be inferred when losing access to a relatively small part of it. All these variables weight on the ability to access and fully understand the digital object.

## 2.1 Structure of a digital object

Figure 2.1 illustrates several components that structure a digital object and relates them to the three levels defined above:

- **Intellectual entity:** A set of content that is considered a single intellectual unit for purposes of management and description: for example, a particular book, map, photograph, or database. An intellectual entity can include other intellectual entities; for example, a Web site can include a Web page; a Web page can include an image. An intellectual entity may have one or more digital representations. (PREMIS Editorial Committee, 2015). The intellectual entity belongs to the conceptual level.

---

<sup>2</sup>In information theory, entropy is the average amount of information contained in each message received. [Shannon] entropy provides an absolute limit on the best possible average length of lossless encoding or compression of any communication.

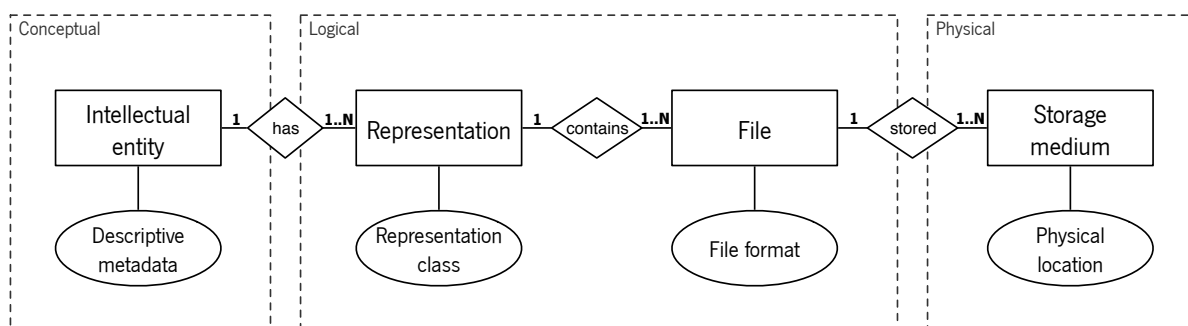


Figure 2.1: Digital object structure (using Chen's Entity-Relationship Diagram)

- **Descriptive metadata:** provides information about the intellectual or artistic content of an object (i.e. intellectual entity) and may also contain data describing the physical attributes of the object (in case there is an analogue counterpart). Descriptive metadata supports specific user tasks, such as discovery and identification of content. In libraries, this category is sometimes called bibliographic metadata.
- **Representation:** The set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity. For example, a journal article may be complete in one PDF file; this single file constitutes the representation. Another journal article may consist of one HTML<sup>3</sup> file and two image files; these three files constitute the representation. A third article may be represented by one TIFF<sup>4</sup> image for each of the 12 pages plus an XML<sup>5</sup> file of structural metadata showing the order of the pages; these 13 files constitute the representation. (PREMIS Editorial Committee, 2015). The Representation belongs to the logical level.
- **Representation class:** a grouping of representations based on technical characteristics for digital preservation purposes. For example: web pages, videos, raster images, etc.
- **File:** a named and ordered sequence of bytes that is known by an operating system. A file can be zero or more bytes and has a file format, access permissions, and file system characteristics such as size and last modification date. (PREMIS Editorial Committee, 2015). File belongs to the logical level.
- **File format:** A set of syntactic and semantic rules for mapping between an information model and a serialized bit stream. Many formats can be grouped into loose categories, or families, sharing a general set of encoding rules that are further restricted or extended for the specific format. Definition taken from Digitization Guidelines Initiative glossary<sup>6</sup> and JHOVE2 glossary<sup>7</sup>.
- **Storage medium:** The physical medium on which the object, represented by its files, is stored (e.g. magnetic tape, hard disk, DVD) (PREMIS Editorial Committee, 2015). The storage medium belongs to the physical level.
- **Physical location:** The particular place, position or address of the storage medium.

<sup>3</sup>HTML means Hypertext Markup Language

<sup>4</sup>TIFF means Tag Image File Format

<sup>5</sup>XML means Extensible Markup Language

<sup>6</sup><http://www.digitizationguidelines.gov/term.php?term=fileformat>

<sup>7</sup><https://bitbucket.org/jhove2/main/wiki/Glossary>

## 2.2 Preservation of the physical level

Strategies to maintain the physical level of digital objects, also called bit-level preservation techniques, are well known on information technology, and are employed in many domains. These strategies include, but are not restricted to, storage media refresh, point-in-time backup, online data redundancy (e.g. RAID), data federation (e.g. LOCKSS, Grid, HDFS) and periodic file fixity checks, which are commonly understood and used by information management experts. Albeit a commonplace for non-experts to forget how unreliable the physical storage mediums are, e.g. 22% of hard drives fail on the first 4 years<sup>8</sup>, there is a growing awareness on the subject by the general public<sup>9</sup>. (Shroeder and Gibson, 2007; Yim et al., 2014)

But these technologies do not solve all the problems that come from keeping information for long periods of time. Digital technology is constantly evolving, changing, adapting to new realities, and retro-compatibility is not always one of the objectives as it is an obstacle to fast development. The hardware devices used by the community to access their information are constantly evolving, getting thinner and faster, as seen on the rapid development of smartphones and tablets. The way information is consumed is changing, multitouch and voice-capable features are mandated by the consumer. Software readers adapt to this new reality, and so do file formats so they can also provide the needs for new features and requirements. Operative systems change frequently and technologies are phased-out, abandoned or replaced (e.g. Adobe Flash).

This brings about what is called of **technological obsolescence**, which is *the condition that happens to a product when it ceases to be useful, even it being in perfect state of functioning, due to the development on a new and more technologically advanced alternative*. Technology obsolescence may happen in many levels. For example, at the hardware level we can see that desktop computers sold today no longer have a diskette drive, and many laptops are sold without any optical disk drive (for CDs, DVDs or Blu-rays)<sup>10</sup>. Formats are replaced by non retro-compatible newer versions, or substituted by different formats. On 2013, CNET advertised the top 5 technologies soon to be obsolete<sup>11</sup>:

1. **Optical disks:** like CDs, DVDs and Blu-rays, optical disks are already being replaced by internet music and video services just like Spotify and Netflix.
2. **Portable cameras:** as smartphones become more common and contain an acceptable camera, they are slowly replacing portable cameras.
3. **Hard disk drives:** although they are currently cheap and have large storage volume, hard disks are being substituted by solid-state drives as these become cheaper.
4. **Keyboard and mouse:** with the advent of smartphones and tablets boom, the keyboard and mouse are being substituted by touch screens, software keyboards and voice recognition technologies.
5. **Home printer:** the home printer is another prey of the ubiquity of screen devices brought upon the smartphone and tablet boom, and also by electronic paper technologies like e-ink (e.g. in Kindle, Kobo or Nook).

---

<sup>8</sup><https://www.backblaze.com/blog/how-long-do-disk-drives-last/>

<sup>9</sup>Backup/recovery software revenue grew 6.8% in 2013, reaching 4.7 billion U.S. dollars.

<sup>10</sup>CD (Compact Disk), DVD (Digital Versatile Disk) and Blu-ray are all digital optical disc storage formats

<sup>11</sup><http://www.cnet.com/news/top-5-soon-to-be-obsolete-technologies/>



With the obsolescence of technology, action must be taken to ensure information doesn't become locked into unsupported or inaccessible hardware or software. If optical disks or hard disk drives do indeed become obsolete, steps must be taken to ensure that information is migrated to newer storage mediums before access is lost to these dying technologies. And the same with the obsolescence of the keyboard and mouse, which will potentially have an impact on hardware devices, operative systems, software readers and, consequently, file formats.

## 2.3 Preservation of the logical level

One of the main objectives of logical-level preservation is to mitigate the effects of technological obsolescence, ensuring that the target community can continuously access and render the information on digital objects. The first step to acquire this capability is to define the policies that will guide the preservation efforts. These policies must identify what are the information assets of interest to preserve (to guide selection and appraisal) and define what is the target preservation level, e.g. if maintaining the physical object integrity is sufficient, or if there are also objectives to maintain continuous access of the digital assets by a target community and to maintain authenticity of the digital information.

On some domains, policies that focus on bit-level preservation, accepting the risk of neglecting the above levels, are currently a trend. The rationale behind this approach is that there is more information being lost on the physical level than at the logical and conceptual levels and if in the future the access to information is lost due to format obsolescence (which is found to be more rare than assumed) then all efforts will be made to make coherence of the information and update it to a format compatible with the state of the art<sup>12</sup>, given the information is of enough importance (Rosenthal, 2010). On this scenario, aggressive harvesting and bit-level preservation techniques are normally used to capture all possible digital objects in scope and maintain their integrity, but the lack of appraisal and selection on the results of the harvest process will result on a large amount of storage being required, spending much of the resources and consequently neglecting the tasks needed to enable continuous access and reuse by the target community.

Other policies define requirements that consider the risks that afflict the logic level as not acceptable. These policies define objectives like *providing continuous access to consumers and enabling the re-use of information* and require other digital preservation strategies. Currently well accepted strategies are:

- **Hardware and software museums:** Preserve the whole technology stack needed to render the original content. The main advantage is the reproduction accuracy by the ability to render information on the exact way it was meant when created. The disadvantages are the great difficulty to maintain the hardware operational on the long-term, the restrictions on the access to information as users have to physically be on the hardware museum, and the need for users to understand how to operate long gone systems.
- **Emulation:** Use of a software system that allows to emulate the behaviour of an older hardware and/or software platform within a newer one. This method is similar to the hardware and software museums, but some of the components of the technology stack are replaced by emulation software. The main advantage is the reproduction accuracy while not having to maintain the emulated hardware. A disadvantage is the complexity to develop emulators for niche hardware or software components, as they depend on the documentation the

---

<sup>12</sup>Rosenthal's view is that pre-emptive format migration does not solve format obsolescence, knowing a format is obsolete can only occur when is too late to migrate, format specifications are not enough to create new renderers or converters, and only by keeping the last working renderers (specially if open-source) one can ease the future work of migrating or emulating an obsolete format.

hardware vendor supplies. Another disadvantage is the difficult reuse of information, as users need to operate obsolete technologies, which hinders the creation of new value. Finally, the emulator itself can become obsolete, in which case we might need to use an emulator to run another emulator, making the technology stack needed to render increasingly complex and more difficult to maintain.

- **File format migration:** Transfer of digital information from one hardware and software configuration into another. More specifically, convert information encoded in a file format, tied into an obsolete technology stack, into another more current or better suited for long term preservation. The main advantage of this strategy is the dissemination of digital assets better suited for the target community to read and re-use, providing the best benefit to the user community. This strategy also removes the need of preserving original hardware and software technology stacks. One disadvantage of this strategy is the possible data loss during conversion, which can be mitigated with proper quality assurance. Also, the need of continuous work to monitor the changes on the technological context of the target community and enact new file format migrations is costly in the long run.
- **Encapsulation:** Keep files together with all necessary documentation needed for future development of emulators, file format migrators or software renderers. This technique allows to postpone preservation actions that can be costly, but is restricted to digital objects that will only be accessed on a far away future. Another disadvantage is the difficulty of gathering specifications of complex objects and closed file formats. Finally, it is difficult to ensure the quality and completeness of the documentation without hindsight, and an incomplete documentation can break down the whole technique.

To better guide the preservation efforts using these strategies, policies can be defined at various levels: Guidance, Procedure and Control and which range from the top-level organisation mission to the most technical control detail (Sierman et al., 2013):

- **Guidance policies:** General description of the organisational goals for long-term preservation of digital collections. Example of a guidance policy: the authenticity all digital objects must be preserved.  
Guidance policies define on a general level what information the institution intends to gather or accept, for how long it will preserve it, what properties of the information it will preserve (e.g. integrity, security, authenticity), and for who will it preserve it (e.g. scientists, general public), considering also what level of service it would strive to provide for this target community (e.g. open-access).
- **Procedure policies:** Description of the approaches the organisation intends to undertake to achieve the goals phrased on the guidance policies.  
Procedure policies give insight on the methods used by the organisation and connect the high-level objectives defined in guidance policies with the measurable indicators defined in control policies. Furthermore, procedure policies should detail the approaches, for example how will it gather information (e.g. web harvest, producer direct contact) how will it keep it (e.g. bit-level preservation, file format migration, emulation), and what service it will provide to the target community (e.g. dark archive, derivatives optimised for access, search portals). Example of a procedure policy: Document all relevant processes within a digital object life-cycle so that current and future users can understand their causes and consequences (towards the objective of keeping authenticity).
- **Control policies:** Definition of the rules and requisites that allow to verify if defined approaches are being correctly followed and therefore the objectives are being achieved.

The control policies define the low-level technical and measurable objectives that must be achieved to consider an approach as successful. For example, for bit-level preservation a measure objective would be that all digital objects must maintain the same file checksum. Another example, for file format migration and emulation, should be that all digital objects must maintain their *significant properties*, even while changing representations.

Significant properties are the characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record (Wilson, 2008; Hedstrom and Lee, 2002). These characteristics need to be defined in the policies to guide strategies like format migration and the following quality assurance processes that ensure digital objects don't lose their intellectual value during migration. The significance of properties may be different for each stakeholder, therefore their definition must be done in light of the institutional objectives (Dappert and Farquhar, 2009).

For the objective of authenticity, which relates to the capability to prove (or vouch) that the digital object is according to the original, the requirements defined on the policies must become more complex as they need to ensure that a repository is trustworthy and certified, that all decision making procedures were correctly followed and documented, and that all preservation actions defined by decision making processes were correctly validated and documented.

The credibility of the digital object authenticity is endowed by the trustworthiness of the digital repository and the institution that supports it. This trustworthiness is a consequence of the institution honourability and credibility and is further improved on the repository by having transparency on the mission, policies and procedures in place for digital preservation, being rigorous on their application and being able to prove, based on evidence, that the defined policies and procedures are correctly followed. To achieve this objective there are tools for repository audit and certification:

- **Data Seal of Approval (DSA):** a certification provided by the experts in digital preservation that are part of the DSA council, with headquarters in The Hague, Netherlands. It certifies that a repository can preserve research data and includes 16 requisites. To obtain the seal, evidences of the requisites achievement must be provided, but there is no formal audit. The result of the remote audit is published online on the DSA site<sup>13</sup>.
- **Nestor seal for trustworthy digital archives:** a certification provided by the preservation experts that belong to nestor (German Network of Expertise for Digital Preservation) and that bases its procedures on the DIN 31644 standard, which has a total of 34 criteria and defines a fulfilment score (DIN, 2012). The evidences provided are checked for plausibility by an external reviewer, which provides a higher level of trust on the seal without requiring a formal audit.
- **ISO 16363:2012 - Audit and certification of trustworthy digital repositories:** an international standard based on the *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (CRL and OCLC, 2007; ISO, 2012), has 109 criteria and requires a formal audit as defined by ISO 16919:2014. Up to now, only one organisation is entitled to conduct audits, the Primary Trustworthy Auditing Body, primarily comprising of people who worked on the ISO 16363 and based in Europe. Even without many certifying bodies, the ISO can be used for a more complete self-assessment and self-audit, by using unofficial audit consultants<sup>14</sup> or by internal efforts.

---

<sup>13</sup><http://www.datasealofapproval.org>

<sup>14</sup>such as KEEP SOLUTIONS <http://www.keep.pt>

- **DRAMBORA:** the Digital Repository Audit Method Based on Risk Assessment is a self-audit toolkit developed by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE) to help guide repository managers along a similar route of analysis to that which an external auditor would use to examine and analyse the work of the repository. Its design is based on the experiences of the DCC audits of digital repositories conducted in 2006<sup>15</sup>. It "seeks to determine whether the repository has made every effort to avoid and contain risks that might impede its ability to receive, curate and provide access to authentic, and contextually, syntactically and semantically understandable digital information"<sup>16</sup>. (McHugh et al., 2008)

The Trusted Digital Repository initiative<sup>17</sup> agreed on a *Memorandum of Understanding* in which are defined three evaluation and certification levels. The first one is defined as *basic certification* is acquired under the DSA and represents a simple self-assessment. The second level is defined as *extended certification* is provided by the Nestor seal and represents a plausibility-checked self-assessment. The third is defined as *formal certification* and it stands for an audit by credited external experts by national or international standards such as the ISO 16363:2012. (Harmsen et al., 2013)

One of the most important evidences to support repository trustworthiness certification is the documentation of all preservation actions that were executed on the digital objects and all the reasoning behind the selection and timing of these actions. The most well accepted metadata schema for preservation metadata is PREMIS (PREMIS Editorial Committee, 2015). The PREMIS version 3.0 data model, depicted in figure 2.2 aligns well with the digital object structure presented before (figure 2.1) and extends it to include documentation on the preservation actions executed on the digital object, the program or person that executed it, and associated access and modification rights:

- **Object (or Digital Object):** is a discrete unit of information subject to digital preservation. An object can be an intellectual entity, a representation, a file, or a bitstream, where an intellectual entity may point to several representations, which may in turn point to the several files or bitstreams it contains and define their relative structure and entry point<sup>18</sup>.
- **Environment:** is the technology (software or hardware) supporting a digital object in some way (e.g. rendering or execution). Environments can be described as intellectual entities and captured and preserved in the repository as representations, files and/or bitstreams.
- **Event:** is an action that involves or impacts at least one object or agent associated with or known by the preservation repository.
- **Agent:** is a person, organisation, or software program/system associated with events in the life of an object, or with rights attached to an object. It can also be related to an environment object that acts as an agent.
- **Rights:** are assertions of one or more rights or permissions pertaining to an object and/or agent.

An example of the usage of PREMIS in a preservation action can be found in this doctoral thesis, which on its conceptual abstraction represents an intellectual entity. This intellectual entity was originally on a representation comprised by

<sup>15</sup><http://www.dcc.ac.uk/resources/repository-audit-and-assessment/drambora/drambora-faq>

<sup>16</sup><http://wiki.statsbiblioteket.dk/domswiki/NotesOnDrambora>

<sup>17</sup><http://www.trusteddigitalrepository.eu>

<sup>18</sup>Same as defined in section 2.1

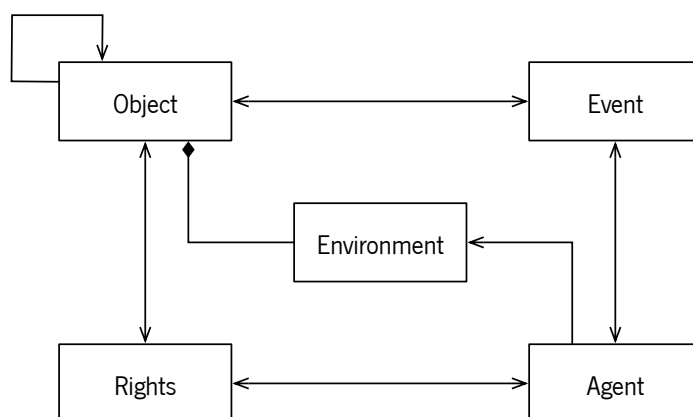


Figure 2.2: PREMIS version 3 data model (as a UML class diagram)

a set of LaTeX and images files and was converted to a single PDF file representation. The act of converting the representation is a preservation event. There can be a PREMIS object for the intellectual entity that documents its significant properties and relates it to an environment description. There will be two PREMIS objects, one for each representation, and they will both point to the intellectual entity. Each PREMIS object for the representation list all files and points to the entry point, which would be the main LaTeX file for the original representation and the PDF for the resulting representation. The PREMIS object, for each of the files, documents necessary information for its preservation like the file checksum, original name, file format, and other technical information. The PREMIS event would be the conversion (or compilation) of the LaTeX and images files into PDF, defining information such as the event date and time, outcome result and details. The PREMIS agent would be the compiler, in this case the TeXShop version 3.51 program. The PREMIS rights could detail that the PDF object is open access.

However, this preservation metadata does not include information detailing why the doctoral thesis was converted to PDF, why was the PDF format selected (and more specifically why PDF version 1.5), why was the TeXShop version 3.51 program selected to convert and if the resulting PDF maintained the significant properties of the original representation. Documenting the decision-making process and presenting the grounding reasons for the decisions, including the evidences that they followed the state-of-the-art guidelines, is necessary to ensure future users that no pernicious actions were selected and maintain the trust on the digital object authenticity. The decision-making process is named on the digital preservation literature as *preservation planning* which included monitoring (or *watch*) functions. But *preservation watch* has been emerging as a component of its own, which closely connects with preservation planning to fulfil the decision-making process. How these concepts have evolved is further explained in section [3.2](#) but before continuing with preservation watch lets first take a quick look at the preservation of the conceptual level.

## 2.4 Preservation of the conceptual level

Although the preservation of the conceptual level is ensured into a certain level by the preservation of physical and logical level, there are risks at the conceptual level that are not taken care on the lower levels. Similarly to the logical level and its technological context, for an intellectual entity to be correctly understood by the user there needs to be

a shared conceptual context. Just as with an analogue document, a user might need to understand the language in which the information is described, it might need to know some concepts, recognise some physical object, be aware of some knowledge domain or have a common social context.

For example, let's take the expression "first floor" in American and British English. In American English the "first floor" is the floor at the street level, while in British English the "first floor" is the one *right above* the street level. While this is a contemporary vocabulary difference between dialects, differences in concepts change much more aggressively through long spans of time, where the same dialect changes, social context is transformed, and even previously very recognisable physical objects become uncommon (e.g. an abacus).

A practical example would be the archive of social network content allied with the current proclivity for the use of cryptic acronyms like LOL, OMG or BFF<sup>19</sup>. When accessing this information on 100 years from now, it would be reasonable to assume that some of these acronyms will not be understood by the general public.

An approach to solve this problem is to create and maintain an ontological database that maps obsolete concepts into newer ones that can be understood by the contemporary conceptual context and allow the target community to easily access it. (Schlieder, 2010, Braud et al., 2013)

While the preservation of the conceptual level is an interesting subject, it will not be the focus of this thesis, which will centre upon the preservation of the logical level, more specifically on the preservation watch process. Therefore, and for the sake of simplicity and being succinct, on the following text of this work, when using the term "digital preservation" it would be implied the meaning of "digital preservation of the logical level".

---

<sup>19</sup>LOL: Laughing Out Loud (can also have a sarcastic connotation), OMG: Oh My God, BFF: Best Friends Forever

# Chapter 3

## Preservation watch

This chapter presents the current state-of-the-art of preservation watch, describing how the concept applies across domains and how it has evolved on the digital preservation literature. This chapter also provides a proposal on how the preservation watch can align with concepts from the risk management domain.

### 3.1 Watch concept across domains

Watch, in the sense of monitoring the surrounding environment, is a very common subject in various domains. In the domain of management science, Stafford Beer defines that any viable or autonomous system must be able to cope with the demands of the changing environment. This adaptability is what ensures its odds of survival and allows a system to be perdurable. Beer defines that information about the environment set by the outside world is a major component of input to top-level decisions. This component is responsible for looking outwards to the environment to monitor how the system needs to adapt to remain viable (Beer, 1981). Beer also defines that there is an internal representation of the real world and that the system constantly tries to match the information that comes from the monitoring activity with this model of the world. Beer further defines that we can consider the system as adaptive where the re-investment (on a business decision or plan) is fed back into the process creating a loop, and that the monitoring component must be able to track the performance of the plan so it may be continuously adapted.

From a business and economic perspective, knowledge about the external environment is also seen as a primal capability needed for survival and the strive for success. "Today, knowledge in all its forms plays a crucial role in economic processes" (OECD, 1996). This applies to nations as well as to businesses and is essential whenever long-term survival is a goal. The need to define and certify the processes that relate to knowledge led to the creation of standards that define Research, Development and Innovation (RDI) management systems, like the Portuguese standard for Research, Development and Innovation Management (IPQ, 2007). This innovation standard defines three interfaces with the outside world:

- **Technological interface** - aims to support technological watch, cooperation and prevision to interact with external technological and scientific research. The technological watch is defined as the systematic, structured and organized gathering of knowledge about economical, technological, social and commercial developments;

- **Market interface** - aims to use marketing and client feedback to gather knowledge on existing markets, requirements, drivers and preferences;
- **Organizational interface** - aims to stimulate, gather and manage internal knowledge and creativity.

## 3.2 Watch in digital preservation

In digital preservation, the focus on maintaining a system and, more importantly, the information viable and ensure its long-term survival is even more crucial. In the functional model of Open Archival Information System (OAIS) (CCSDS, 2002), the monitoring functions are part of the Preservation Planning function and can be divided into:

- **Monitor designated community** - "interact with consumers and producers to track changes to their service requirements and available product technologies";
- **Monitor technology** - "[track] emerging digital technologies, information standards and computing platforms (i.e. hardware and software) to identify technologies which could cause obsolescence in the archive's computing environment and prevent access [to content]".

In 2009, the Planets Functional Model (Sierman et al., 2009) described a model for preservation planning and watch, depicted in figure 3.1 separating monitoring from planning and coining the term *Preservation Watch*. This model does not try to substitute the OAIS, but to refine and extend it to the needs of the Planets project.

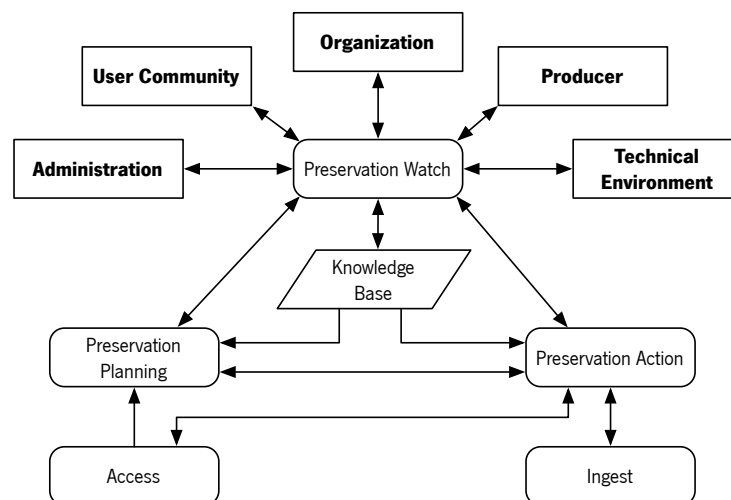


Figure 3.1: Planets Functional Model

In the Planets Functional Model, the two OAIS monitoring functions were combined into *Preservation Watch* function. Also, a *Knowledge Base* component is made relevant, and is defined as "an umbrella term for a repository of a variety of key information which will inform preservation processes conducted within the digital repository" (Sierman et al., 2009). This knowledge base loosely maps to the *Representation Network* concept described in the OAIS Model and is also analogous to the Beer's internal model of the world concept. The model defines that *Preservation Watch* is composed of the following sub-functions:



- **Monitor** - collecting preservation related information from a variety of internal and external sources;
- **Risk Analysis** - assessment of the collected information identifying critical risks that are then relayed to preservation planning;
- **Representation Information Update** - update of the knowledge base with new collected information;
- **Testbed** - controlled environment to execute experiments to better assess preservation tools and services.

The Planets Functional model further describes which entities should be monitored:

- **Administration** - monitor content collection and usage profiles;
- **Organisation** - monitor internal changes in the organisation that might affect the ability or strategy to preserve the content;
- **Knowledge base** - monitor the information in the (internal or external) Knowledge Base and assess if a threat arises;
- **User community** - monitor user community expectations and scope;
- **Producer** - monitor changes in producer technologies and scope;
- **Technical environment** - monitor developments in technology that might lead to threats or opportunities.

More recently, a new reference architecture has emerged. In the SHAMAN project (Antunes et al., 2011), Enterprise Architecture frameworks are used to define a reference architecture for digital preservation that has a broader view of the problem and is founded on System Architecture and Information Systems design principles. One of the premises is that OAIS has a restrictive view of the system, confining the perspective to the archive, and that the digital preservation problem extends to the whole life-cycle of an object. The other observation is a lack of systematic architectural principles and coherence in the frameworks adopted across the digital preservation community.

The SHAMAN project defines a capability-centred Reference Architecture, partially depicted in figure 3.2. In the capability model, *Preserve Contents*, defined as "the ability to maintain content authentic and understandable to the defined user community over time and assure its provenance", is composed by the *Preservation Planning* and *Preservation Operation* capabilities. *Preservation Operation* relates to the preservation plan deployment and execution. *Preservation Planning* is the "ability to monitor, steer and control the preservation operation of content so that goals of accessibility, authenticity, usability and understandability are met with minimal operational costs and maximal (expected) content value". Hence, *Monitoring* is one of the main capabilities of *Preservation Planning* and is defined by the ability to look towards inside, to monitor operations specified by plans and the properties of the system where they run (internal monitoring), and the ability to look towards outside, to monitor external influencers that might implicate a reevaluation of the plans.

In terms of information, tools and systems to aid the monitoring process which current literature identifies, they are limited to technical reports, file format or tool registries, and format obsolescence methodologies and systems that try to identify what preservation threats are generically associated to file formats (Ferreira et al., 2006):

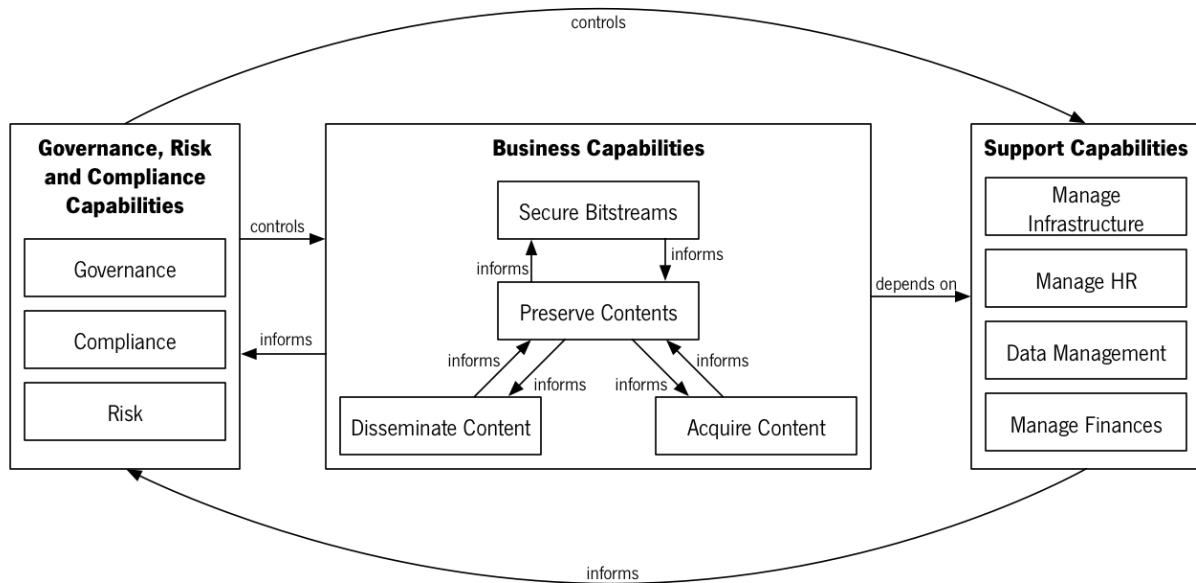


Figure 3.2: SHAMAN Detailed Capability Relationship Diagram

- **Risk Management of Digital Information: A File Format Investigation** - a report on the impact assessment on file format migration and identification of the risks that executing or postponing a migration might introduce (Lawrence et al., 2000);
- **DPC Technology Watch report series:** a periodic report intended as an advanced introduction to specific issues for those charged with establishing or running services for long term access. They identify and track developments in information technology, standards and tools which are critical to digital preservation activities<sup>1</sup>;
- **INFORM methodology** - a model to predict file format obsolescence by discovering threats to preservation and their possible impact on preservation decisions (Stanescu, 2005);
- **File format and tool registries** - online registries focused on digital preservation information about file formats, software products and other technical components relevant to preservation. Examples are PRONOM<sup>2</sup>, Global Digital Format Registry (GDFR)<sup>3</sup>, Unified Digital Format Registry (UDFR)<sup>4</sup>, P2 registry<sup>5</sup>, the Conversion Software Registry<sup>6</sup>, COPTR<sup>7</sup> and POWRR<sup>8</sup>. Unfortunately, some of these online registries are not yet or no longer functioning and are always very incomplete and outdated;

<sup>1</sup><http://www.dpconline.org/advice/technology-watch-reports>

<sup>2</sup><http://www.nationalarchives.gov.uk/PRONOM/>

<sup>3</sup><http://gdfr.info>

<sup>4</sup><http://udfr.org>

<sup>5</sup><http://p2-registry.ecs.soton.ac.uk>

<sup>6</sup><http://isda.ncsa.uiuc.edu/NARA/CSR/>

<sup>7</sup><http://coptr.digipres.org>

<sup>8</sup><http://digitalpowrr.niu.edu/tool-grid/>

- **Automatic Obsolescence Notification Service (AONS)** - a software system 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 File Format registries and receive notifications. The AONS was a good step towards automation but failed because the gathered information was not sufficient, as it assumed the file format registries would have enough preservation risk information (Pearson, 2007).

There is also a wide list of institutions that have published their preservation policies<sup>9</sup>, which give an insight on how these peer institutions cope with risks and the approach taken to mitigate them (Sierman et al., 2014). Other similar resources include repository implementation specific format policy registries, like Archivematica FPR, which "indicates the actions, tools and settings to apply to a file of a particular file format (e.g. conversion to preservation format, conversion to access format)".<sup>10</sup>

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<sup>11</sup>, FileInfo<sup>12</sup>, alternativeTo<sup>13</sup>, iUseThis<sup>14</sup> and Download.com<sup>15</sup>. 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.

### 3.3 Watch and risk management

The reason why we should worry about preservation of digital content and why some preservation action needs to be performed is closely related to the idea that content is at risk. Risk is defined as "the effect of uncertainty on objectives", in which the effect can be positive, negative or a deviation from the expected. (ISO, 2009a, ISO, 2009b)

In the digital preservation community, it is difficult to find quorum regarding the set of risks an institution must be aware, due to the different set objectives institutions have. This is particularly evident between institutions that have a different view of where the (potential) value of digital preservation lies. For some institutions the (potential) value is in the content, the objective would be to capture and maintain the content as it springs from production, and the uncertainties affect the ability to do so, losing content which unforeseeable future uses would potentially bring real value. For other institutions, the value is on the continuous service it provides for its intended users and uncertainties affect the ability to provide this service or maintain it available and with enough quality.

Although these views might be incompatible in many ways, they both share the long-term and continuous effort to maintain a capability, which means that there should be a continuous and long-term process that knows when objectives are not being achieved. This process is preservation watch.

<sup>9</sup><http://wiki.opf-labs.org/display/SP/Published+Preservation+Policies>

<sup>10</sup><https://www.archivematica.org/en/docs/fpr/>

<sup>11</sup><http://filext.com>

<sup>12</sup><http://www.fileinfo.com>

<sup>13</sup><http://alternativeto.net>

<sup>14</sup><http://iusethis.com>

<sup>15</sup><http://download.cnet.com/>

Risk management and preservation watch concepts overlap to some extent. Figure 3.3 and following term definitions try to reconcile concepts from both risk management and preservation watch domains. As recognised by some literature on the subject "the terminology used to describe the steps in the risk management process is not consistent" (Airmic et al., 2010), which forces this exposition to slightly change some of the terms to better adequate the alignment of concepts in these domains.

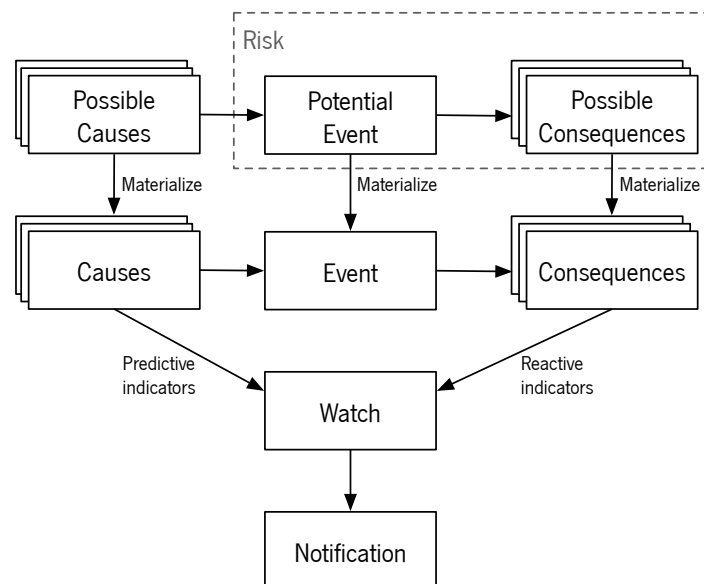


Figure 3.3: Relation between risk and watch concepts

Figure 3.3 tries to align concepts that derive from risk management as defined by (ISO, 2009a) (ISO, 2009b) and concepts of preservation watch. This alignment allows us to define the following terms:

### **Risk**

*Effect of uncertainty on objectives.* An effect is a deviation from the expected, being it positive and/or negative. The objectives can have different aspects (such as financial, health and safety, and environmental goals) and can apply at different levels (such as strategic, organization-wide, project, product and process). The formalisation of objectives align with the definition of preservation policies, which are presented in section 2.3. Risk is often characterised by reference to potential events and their (possible) consequences and is often measured by the likelihood of the potential event and the impact of its possible consequences on the objectives. (definition based on ISO Guide 73:2009)

### **Potential event**

A possible happening of particular importance that introduces opportunities for benefit (upside), threats to success (downside) or an increased degree of uncertainty. Examples: a fire, a staff strike, a new storage technology, new content of unknown file formats. (definition partly based on (Airmic et al., 2010))

For the sake of clarity, on the digital preservation domain and for the rest of this thesis, a potential event will be referred to as a **preservation threat** if it has negative consequences or a **preservation opportunity** if it has positive consequences. An increase of the uncertainty is also considered a negative consequence.

**Event**

A discrete occurrence, or a set of discrete occurrences, of a particular possible event. An occurrence happens on a defined moment in time and its consequences have a quantifiable or qualifiable impact on the objectives. Example: file X was corrupted at date Y and has become unreadable.

An event can have one or more occurrences, and can have several causes and consequences. An event can consist of something not happening and can also have no consequences. An event can sometimes be referred to as an *incident*. (*definition partly based on ISO Guide 73:2009*)

**Cause and Possible cause**

A cause is an element which alone or in combination gives rise to an event (occurrence).

A possible cause is an element which alone or in combination has the intrinsic potential to give rise to an event. (*definition partly based on ISO Guide 73:2009*)

On ISO 31000, a cause is referred to as a risk source. If the related risk or event has negative consequences the cause can be referred to as hazard (source of potential harm) or vulnerability (intrinsic properties of something resulting in susceptibility to a risk source that can lead to an event with a consequence).

**Consequence and Possible consequence**

A consequence is the outcome of an event affecting objectives.

A possible consequence is the predicted outcome of a potential event that would affect the objectives.

A consequence can be certain or uncertain and can have positive and/or negative effects on objectives. The impact of consequences in the objectives can be expressed qualitatively or quantitatively. (*definitions partly based on ISO Guide 73:2009*)

**Indicator**

Measurement of the probability that an event has already occurred (reactive indicator) or might occur in the near future (predictive indicator).

A predictive indicator bases its measurements on directly monitoring the causes of an event or indirectly monitoring other aspects of the environment that might suggest that an event is about to occur on the near future.

A reactive indicator bases its measurements on directly monitoring the consequences of an event or indirectly monitoring other aspects of the environment that might suggest that an event has occurred.

**Watch**

Watch is the continuous process that gathers indicators in order to identify the possible occurrence of events and notify the relevant parties.

Watch is also the "continual checking, supervising, critically observing or determining the status in order to identify change from the performance level required or expected". (*definitions partly based on ISO Guide 73:2009*)

**Notification**

Notification is a message that the watch component sends to relevant parties informing that there is a high probability that an event of importance has occurred or is about to occur.

When receiving the notification, the relevant parties should make a manual assessment of the event, i.e. to verify if it has indeed happened or if there are enough evidences to sustain the high probability of it happening soon that would justify taking an appropriate action, by interfacing with the external decision-making and action-taking processes. These processes are denominated in the digital preservation context as "preservation planning" and "preservation operations" and in the risk management context as "risk assessment" and "risk treatment". ISO 31000 also recognises the importance of feedback mechanisms by defining that the watch process should also monitor the result of the risk assessment and treatment processes to measure risk performance and learn from experience. This framework aligns well with the *SCAPE preservation life-cycle* architecture, to be presented on section [4.2](#)

# Chapter 4

## Context and methodology

This chapter describes the SCAPE research project, on which much of the work of this thesis was done. It also describes the research questions that will be addressed on this work and explains the approach taken to derive the answers and validate the hypothesis.

### 4.1 SCAPE: Scalable Preservation Environments

The SCAPE project was set between 2011 and 2014, and its mission was to "develop scalable services for planning and execution of institutional preservation strategies on an open source platform that orchestrates semi-automated workflows for large-scale, heterogeneous collections of complex digital objects"<sup>1</sup>. The work presented on this thesis was partly developed within this project, and lays out in a bigger and overarching architecture for scalable digital preservation. Partners included:

- Austrian Institute of Technology (AIT), Austria
- The British Library, United Kingdom
- Internet Memory Foundation, France
- Ex Libris, Israel
- Fachinformationszentrum Karlsruhe Gesellschaft für Wissenschaftlich-Technische Information (FIZ Karlsruhe), Germany
- Koninklijke Bibliotheek (National Library of the Netherlands), Netherlands
- KEEP SOLUTIONS, Portugal
- Microsoft Research

---

<sup>1</sup><http://www.scape-project.eu/>

- Österreichische Nationalbibliothek (Austrian National Library), Austria
- Open Planets Foundation (now named Open Preservation Foundation), United Kingdom
- Statsbiblioteket (Danish State and University Library), Denmark
- Science and Technologies Facilities Council (STFC), United Kingdom
- Technische Universität Berlin (Technical University of Berlin), Germany
- Technische Universität Wien (Technical University of Vienna), Austria
- The University of Manchester, United Kingdom
- Universite Pierre et Marie Curie (UPMC), France
- Brno University of Technology, Czech Republic
- Poznan Supercomputing and Networking Center, Poland
- West University of Timisoara, Romania
- Wielkopolskie Center of Pulmonology and Thoracosurgery, Poland

The SCAPE project is organised into six sub-projects comprising 22 distinct work packages:

1. **Cross-project Activities:** usual coordination and roadmap activities;
2. **Technical Platform:** work packages related to design, development and provisioning of a platform for the execution of scalable preservation services;
3. **Preservation Components:** work packages related to the development of components to deal with specific preservation tasks like characterisation, quality assurance and file format migration;
4. **Planning and Watch:** work packages related to the development of tools to support the streamline and scalability of the watch and planning processes, while regulated by defined policies;
5. **Testbeds:** work packages that aim to assess the large-scale applicability of the preservation platform and components developed in the project;
6. **Take-up:** work packages that focus on the dissemination of results, training and sustainability of developed services, in order to maximise the impact of the project on the community.

The author of the present thesis is part of KEEP SOLUTIONS and was the leader of the Preservation Watch work package, part of the Preservation and Watch sub-project, managing the design and development of a solution that would fit the Watch requirements of the overarching architecture named as the *preservation life-cycle*, described on the following section. KEEP SOLUTIONS also led the Preservation Components sub-project.

The preservation watch work package had the following specific goals:



1. To develop and implement a model of watch mechanisms, triggers, and suitable actions to be taken for each trigger. To support closed-loop automated preservation processes, in which automated monitoring of collections, actions, plans, systems, and the environment triggers diagnosis and reaction.
2. To develop an environment that simulates repository growth in terms of number of objects, file formats, versions, storage, etc., and how it affects preservation plans and triggers new preservation events.

The work described on this thesis is mainly related with the first goal, while the second goal which focused on simulation is mainly described on (Duretec, 2014).

Some of the main results of the SCAPE project are the *SCAPE preservation life-cycle* and its implementation: the *SCAPE preservation suite*. Both are described on the following sections. (Faria et al., 2013b, Becker et al., 2014, Duretec et al., 2014)

## 4.2 SCAPE preservation life-cycle

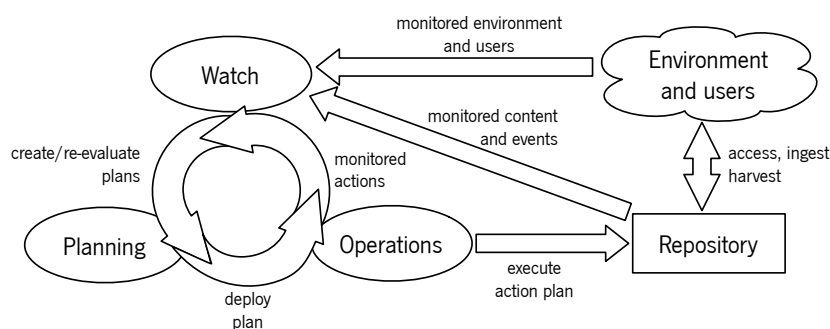


Figure 4.1: Base SCAPE preservation life-cycle with processes interaction

Figure 4.1 shows the key building blocks of the SCAPE preservation life-cycle. To preserve content, the preservation threats that hinder the continuous and authentic access to the content need to be identified and continuously monitored. To this purpose, a continuous **watch** process monitors the alignment of what the **repository** has and does with its objectives, the environment and its users. Digital preservation starts by understanding what content a repository holds and what are the specific characteristics of that content. This process is supported by the characterisation of content and allows a content owner to be aware of the storage volumes, characteristics, format distributions, and specific peculiarities such as digital rights management issues, complex content elements, or other preservation threats.

The characterisation process feeds the key characteristics of the **monitored content** into the preservation **watch** process that should cross-relate the results of this internal content characterisation with the institutional policies and external information about the technological, economic, social and political environment that the repository is set upon, allowing for the identification of preservation threats and cost-reduction opportunities. Checking the conformance of content with the owner's expectations or policies, identifying format or technological obsolescence in content or comparing content profile with other repositories can reveal preservation threats. Repository **events**, e.g. ingest and

download of content, can also be useful for tracking producer and consumer trends and can be used to uncover preservation threats.

These possible threats and opportunities should then be analysed by preservation **planning**. The planning process carefully examines the threats or opportunities, having in mind the institution's goals, policies, objectives and constraints. It evaluates and compares possible alternatives and produces an action plan that defines what operations should be implemented and the reasoning that supports this decision.

An action plan is **deployed** into the **operations** process that orchestrates the **execution** of the necessary actions on the repository content, if necessary in large-scale, and integrates the result back in the repository. These operations include characterisation, quality assurance, migration and emulation, metadata, and reporting. The operations process should provide to the watch process information about the executed actions (or **monitored actions**), such as quality assurance measurements, to be sure that the results conform to the expected. Also, all assumptions about internal and external information taken by the planning process should be continuously monitored so the action plans (to do some action or to remain idle) remain valid. Once a plan becomes "invalid" the preservation planning process should be called upon to **re-evaluate the plan**, creating a continuous life cycle that ensures content remains preserved. (Faria et al., 2013b)

However, there is a change in the concept design forced by practical concerns, namely that the large scale execution of preservation actions is only feasible when the processing (i.e. the operations) is brought close to the data, and not the other way around as inferred by the figure. This happens because with current technology it is very costly to transfer high amounts of data back and forth, when in the order of the tens of Terabytes, and many times the transfer takes more time than the processing itself. Due to this fact, which was observed in the several tests made on the SCAPE project (Schmidt, 2012; Schmidt and Rella, 2014; Schlarb et al., 2014; Palmer et al., 2014a; Duncan et al., 2014; Ferneke-Nielsen et al., 2014), the plan is now to be deployed directly on the repository (allowing also for its safekeeping) and it is the repository the one that communicates with the preservation operations process, allowing for a more direct access to the data. The result of the actions must then be monitored by the repository and this information should then be relayed to watch, together with information about the content and other events. Figure 4.2 shows the resulting review of the preservation life-cycle, which was implemented by the SCAPE preservation suite to be presented next.

### 4.3 SCAPE preservation suite

Each of the preservation processes can be done manually, but the common increasing volume and heterogeneity of documents in institutions make it necessary for tools to exist to support and automate part of these processes. In this section we present the tools and integration APIs that support the SCAPE preservation life-cycle into a complete architecture for large scale digital preservation named the SCAPE preservation suite.

There are several tools for content characterisation (FITS<sup>2</sup>, Apache Tika<sup>3</sup>, ffprobe<sup>4</sup>, etc.), some are very specific of the file format they work with, other wrap several tools together and work with a larger set of file formats and object classes. These tools provide technical information about the files and their key characteristics. However, these tools do not

---

<sup>2</sup><http://projects.iq.harvard.edu/fits>

<sup>3</sup><http://tika.apache.org/>

<sup>4</sup><https://www.ffmpeg.org/ffprobe.html>

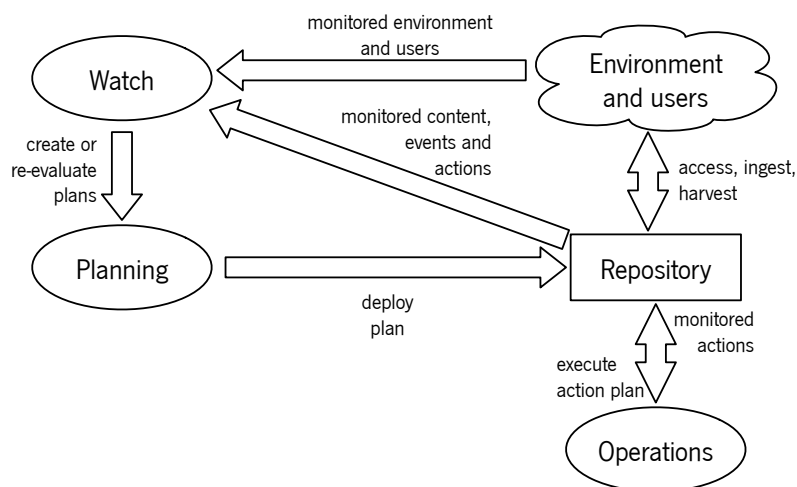


Figure 4.2: Reviewed SCAPE preservation life-cycle with processes interaction

provide aggregation and analysis of these characteristics, something considered necessary to feed back information into the watch and planning processes. To fill this gap, the C3PO tool<sup>5</sup> was developed (Petrov and Becker, 2012). C3PO collects information from characterisation tools and provides a content profile, i.e. an aggregated view of the content characteristics, necessary to support the watch process. Furthermore, the tool analyses the content and allows selection of representative datasets, which are necessary for the planning process. Also, C3PO provides an interface for browsing and drill-down of content characteristics and a programmatic API.

Scout<sup>6</sup>, the main artefact in this thesis and presented in depth on chapter 6 is a preservation component that provides an ontological knowledge base to centralize all necessary information to detect preservation threats and opportunities (Faria et al., 2012b; Becker et al., 2012). It uses plugins to allow easy integration of new sources of information, as file format registries, tools for characterisation, migration and quality assurance, policies, repository content and events, and others. The knowledge base can be easily browsed and triggers can be installed to automatically notify users of new threats and/or opportunities. Examples of such notification could be: content fails to conform to defined policies, consumers cannot read the provided format or a software tool is no longer supported.

Plato<sup>7</sup> is a well-established tool for systematic preservation planning. It allows definition of preservation objectives, criteria and restrictions necessary for decision-making and helps with the evaluation of all action alternatives, arriving to a well-determined best solution, documenting all reasoning behind the decisions, and providing traceability, one of the basis for maintaining the authenticity of digital assets (Becker et al., 2009). The result of preservation planning is an action plan that, besides documenting the process itself, defines the necessary actions to perform on content.

If, in one hand, the actions to be performed on content raise feasibility concerns due to the content volume or the action computing intensiveness, scalable platforms need to be taken into consideration. The SCAPE platform provides guidelines on how to deploy such a platform to support execution of large-scale preservation actions (Schmidt, 2012). On the other hand, if the processing scalability is not a concern, less complex platforms can provide the same action

<sup>5</sup><http://ifs.tuwien.ac.at/imp/c3po>

<sup>6</sup><http://scout.openpreservation.org/>

<sup>7</sup><http://ifs.tuwien.ac.at/dp/plato/>

plan execution features, such as the workflow engine Taverna<sup>8</sup>. Furthermore, example workflows of preservation action plans and components, e.g. characterisation, migration and quality assurance, can be found and shared in the myExperiment site<sup>9</sup>.

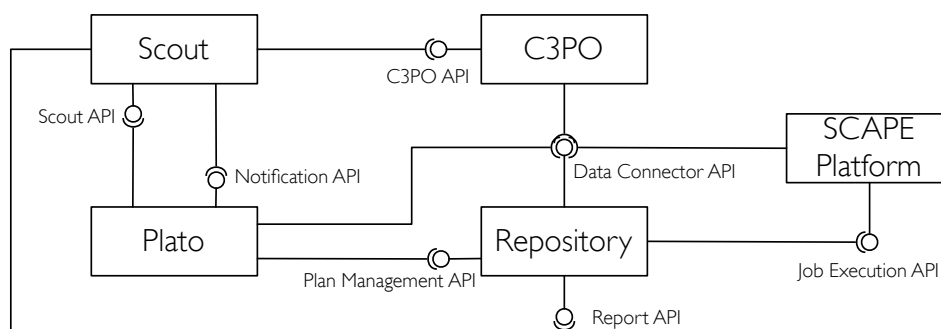


Figure 4.3: SCAPE preservation suite architecture

Figure 4.3 depicts all software components necessary for the preservation life-cycle (already described in the previous section) and focuses on the interfaces between each component. This is not a strict architecture because any of the software components can be skipped and the process it supports can be done manually or with other tools. Every programmatic interface has analogous human interface that achieves the same functionality. This is, therefore, an open and loosely coupled architecture that can be incrementally integrated into a repository implementation. A repository can integrate into this preservation life-cycle architecture by implementing three interfaces: 1) **Data Connector API**: interface to create, retrieve, search and update digital objects within a repository; 2) **Report API**: interface to retrieve information about events that take place on a repository, e.g. ingest, access, and preservation operations; 3) **Plan Management API**: interface to manage and execute preservation plans. The implementation of the Plan Management API can use the Job Execution API to actively perform preservation operations as defined by a preservation action plan. The Job Execution API provides an interface for performing and monitoring parallel data processing operations (jobs or workflows) on the platform infrastructure. (Asseg et al., 2013)

Reference implementations of the APIs are developed for the repositories RODA<sup>10</sup> and Fedora 4<sup>11</sup> with focus on creating reusable components that could help with the development of APIs for other repository implementations. Other repositories have already implemented some of the APIs, table 4.1 shows the implementation status.

Table 4.1: Repository implementations of SCAPE APIs as of November 2013

Repository	Data Connector API	Report API	Plan Management API
RODA	Yes	Yes	Yes
Fedora 4	Yes	No	Yes
Rosetta <sup>12</sup>	Yes	No	No
dArceo <sup>13</sup>	Yes	No	No

<sup>8</sup><http://www.taverna.org.uk/>

<sup>9</sup><http://www.myexperiment.org>

<sup>10</sup><http://www.roda-community.org>

<sup>11</sup><http://fedorarepository.org/>

The details of this implementation, including a walkthrough of the software components, information flow, problems and decisions, are described in section [6.3](#) named Scout implementation, and put into practice in an experiment in section [7.1](#)

## 4.4 Research questions

Digital technology is engulfing content production and we face a major problem, digital content is difficult to preserve. The digital preservation research and business domain have grown in the last decade, but there is still a major gap for effective and trustworthy digital preservation: The grounds that enact and base decision-making processes need to be documented, need to prove they follow guidelines and policies, and need to be able to cope with the large scale and heterogeneous digital content problems that institutions face. Preservation watch is the process that enacts planning (decision-making) and operations (action-taking) processes, identifying the preservation threats and relaying them to planning, so an action can be chosen which is then executed by operations. The result is again monitored by the watch process to ensure that threats were mitigated and find new possible threats that can emerge from the actions or from the environment. This continuous cycle, that can be named SCAPE preservation life-cycle, connects all processes. To achieve effective and trustworthy digital preservation all these processes need to be systemised and documented, automatized and integrated as much as possible, so the whole preservation life-cycle can cope with large scale and heterogeneous content on the long term.

Hence, the watch process has the following roles:

1. To identify threats by monitoring the repository, the users and the environment and notify the relevant parties so planning can start;
2. To assess if the threats were mitigated by executed actions and ensure if new threats were not introduced by them or by external factors, notifying the relevant parties if a plan re-evaluation is needed.

The watch process should also be automated as much as possible to allow systemisation, documentation and scalability, and should integrate with planning and operations processes.

To cope with these requirements, we need to find which preservation threats should enact planning, both to create or re-evaluate plans, starting with the most important and neglected ones as they reveal where the current gaps are. Then, we need to ensure that information from the repository, the users and the environment, i.e. the world, can be collected and formally represented on a knowledge base, i.e. an internal representation of the world, to automatically detect preservation threats on a digital repository. These requirements can be formalised on the following research questions:

**Question 1.** *What are the most important and neglected digital preservation threats?*

**Question 2.** *Can these threats be detected using a formal representation of the information about the world that is automatically monitored and collected?*

---

<sup>12</sup><http://www.exlibrisgroup.com/category/RosettaOverview>

<sup>13</sup><http://dingo.psnc.pl/darceo/>

## 4.5 Approach

The approach chosen to identify the most important and neglected digital preservation threats was a survey to the digital preservation community. This survey can identify preservation threats that are perceived as important<sup>14</sup> to the community and also ascertain their own level of practice in monitoring them. Also, it can provide the community's most well credited methods to detect these threats. The adequacy of the methods implies the community trust on these indicators, proving the final objective which is to endow credibility and support authenticity claims.

The chosen survey method is an online questionnaire using a specialised and popular software tool<sup>15</sup> and taking advantage of the SCAPE project dissemination channels. An online questionnaire was found a very adequate method to cope with the geographically distributed and tech savvy target audience: the digital preservation community. The survey is presented in chapter 5.

The information on the most important and neglected threats and the most well accepted methods to monitor them provides the base to develop a software artefact named *Scout* that implements the defined requirements of automatic digital preservation threat detection by the monitoring and collecting information about the world that is mapped to a formal representation into an internal knowledge base. It defines a set of information sources based on the state of the art and the survey and specifies how to gather information, automatically detect and monitor threats, and notify relevant parties so they can start planning. *Scout* is presented in chapter 6.

*Scout* serves as an instrument for a series of experiments based on real-world scenarios that prove *Scout* can in fact automatically detect the most important and neglected threats. This evaluation is presented in chapter 7.

---

<sup>14</sup>Please note that although threat importance could be studied in a more objective way, unbiased by the perspective of a group of people, one of the core objectives of this process is trustworthiness which is mainly based on subjective factors.

<sup>15</sup>Limesurvey is an open-source survey tool on the web that is available at <http://www.limesurvey.org>

# Chapter 5

## Survey on preservation watch concerns

This survey aims to identify the most important and neglected digital preservation threats, the preferable methods to detect them and what is the current practice in terms of digital preservation watch. The following sections show the details of the survey audience and results, appendix [A.1](#) lists the original set of questions.

### 5.1 List of questions

See below a brief description of the 30 questions employed on the survey, for a full description of the questions and the choices see appendix [A.1](#). Please note that the terms **preservation threat** and **preservation incident** are used interchangeably and that a definition of the terms is available on section [3.3](#) and further discussed on section [5.2.2](#).

Profile (2 questions):

1. What descriptions fit your organisation? [multiple options with open other]
2. What descriptions fit your role on your organization? [multiple options with open other]

Digital preservation threats [list of threats below] (3 questions):

1. What is the importance of the following threats? [1 to 5 Likert scale for each threat]
2. Which of these threats are you already monitoring? [Yes/Uncertain/No for each threat]
3. Are there any other digital preservation threats you find important? [open question]

Detecting and monitoring preservation threats (22 questions):

- For each of the threats, a list of monitoring methods is presented (list below) and for each the user is asked for their preference [1 to 7 Likert scale] and if their organisation uses the method to identify the corresponding threat [Use/Don't use/Uncertain].
- There is also for each threat an open question for other possible detection methods.

Follow up (3 questions):

1. Would you like to know more about our digital preservation monitoring tools? If so, please give us your contact.  
[open fields for name, email and institution]
2. Would you like to run our tools to know your file format distribution, along with other content characteristics, and compare it with others? [Yes/No]
3. Would you like to participate in workshops, virtual or in person, to know how to use the digital preservation tools created in the SCAPE project? [Yes/No]

List of suggested preservation threats and corresponding detect methods:

### **File corruption**

Threat of losing data due to media degradation, bit rot, (in)voluntary modification or deletion, etc.

- **Check files manually:** Verification that files are not corrupted by manually inspecting them.
- **Automatic file fixity checks:** Automatic file fixity check is a program that verifies if the file bytes have changed by processing the file and comparing with existing technical/preservation metadata.

### **Backup failure**

Backup is the copy of a file or other item of data made in case the original is lost or damaged. This is a common mitigation technique for the file corruption threat, but introduces the new threat of backup failure.

- **Manual verification:** Periodic manual check of backup success, which may include restore testing.
- **Alerted by backup program on failure:** Alerted by backup program when failure occurs.
- **Notified by backup program every time:** Notified by backup program on every execution.
- **Monitored by 3<sup>rd</sup> party program:** Third-party program monitors correct functioning of backups.

### **Hardware platform obsolescence**

Hardware no longer supported or degraded. Examples of hardware: media reader, storage, network or processing components.

- **Manual by IT staff:** Manual analysis of the hardware inventory by IT staff.
- **Manual by preservation experts:** Manual analysis of hardware inventory by preservation experts.
- **Relationship with vendors:** Have a close relationship with hardware vendors.
- **Crossing inventory with issues registry:** Automatic cross-reference of hardware inventory with a known hardware issues registry.



**Software platform obsolescence**

Software platform no longer supported or degraded. Examples of software: operative system, application server or digital repository system.

- **Manual by IT staff:** Manual analysis of software inventory by IT staff.
- **Manual by preservation experts:** Manual analysis of software inventory by preservation experts.
- **Subscribing mailing lists and others:** Subscribe relevant mailing lists or other information channels.
- **Crossing inventory with issues registry:** Automatic cross-reference of software inventory with known software issues registry.

**Producers misalignment**

A relevant percentage of the producers cannot comply with the established ingest policies.

- **Direct engagement with producers:** Feedback by direct engagement with producers.
- **Trends from web analysis:** Monitoring producer trends and problems by analysis of content from the web.
- **Monitor ingest process:** Monitor ingest process (e.g. SIP rejection statistics)

**Consumers misalignment**

A relevant percentage of the consumers cannot read the disseminated file format.

- **Direct engagement with consumers:** Feedback by direct engagement with consumers.
- **Feedback by email or other channel:** Consumer feedback by email or analogous channels.
- **Repository integrated feedback:** Repository integrated consumer feedback.
- **Manual check collections by IT staff:** Manual analysis of file formats in your collections by IT staff.
- **Manual check collections by preservation experts:** Manual analysis of file formats in your collections by preservation experts.
- **Crossing formats with rendering tools:** Automatic cross-reference of file formats in your collections with tools that can render formats.
- **Crossing formats with format registries:** Automatic cross-reference of file formats in your collections with information available on format registries.
- **Crossing formats with format issues:** Automatic cross-reference of file formats in your collections with known file format issues.
- **Trends from web analysis:** Automatic analysis of consumer trends by inspection of consumer used software (e.g. browsers and operative systems).

**Lack of context information**

There is not enough context information to understand the file content.

- **Manual check content on ingest:** Manual inspection of content on ingest.
- **User feedback:** Get information from the users by directly interfacing with them or via online channels like email or online questionnaires.
- **Check external references:** Automatic verification that external references still exist (e.g. web sites).

**Content not aligned with policies**

Content does not conform with defined institutional policies.

- **Manual check content on policy change:** Manually verify content on ingest and whenever applicable policies change.
- **Automatic check with control policies:** Define control policies in a machine readable format and have tools that automatically check the conformance.

**Staff not enough or adequate**

Organisation staff is not enough or adequately trained to maintain content.

- **Manual staff evaluation:** Manually evaluate staff performance and quality.
- **Automatic indicators of staff performance:** Have automatic indicators of staff performance (e.g. objects ingested, described, words written).
- **Consumer feedback on staff performance:** Have consumer feedback on the quality of content description and cataloguing.

**Incorrect action results**

Actions executed on files (e.g. file format migration) are not having expected results.

- **Manually verify the action results:** Manually verify all (or a sample) of the action results.
- **Automatic check with Q&A tools:** Run quality assurance tools on action results and automatically check against expected results.

**Outdated preservation plans**

Defined preservation plans (e.g. defined preservation format) became outdated.

- **Update plans periodically:** Update preservation plans periodically (e.g. yearly).
- **Use plan creation tools that notify automatically:** Use tools to create preservation plans (e.g. Plato) and be automatically notified when assumptions made become invalid.

## 5.2 Results

This section presents the results of the online questionnaire.

### 5.2.1 Sample

The survey consisted on a convenience sample made by an online questionnaire that ran from the 31st of January to the 28th of February 2014. There was an initial invitation at the start date and a reminder on the 24th of February. The invitations and reminders were sent through many channels: the SCAPE project website and newsletter, Open Planets Foundation blog, mailing list and LinkedIn, Digital Preservation Coalition news page, JISC mailing lists, and DIGLIB mailing lists. Also, requests were made to forward the invitation in the Nestor mailing list, DigCurV and DCC organisation and projects, and the DLM forum. The invitation was further disseminated on social networks, such as Twitter and Facebook.

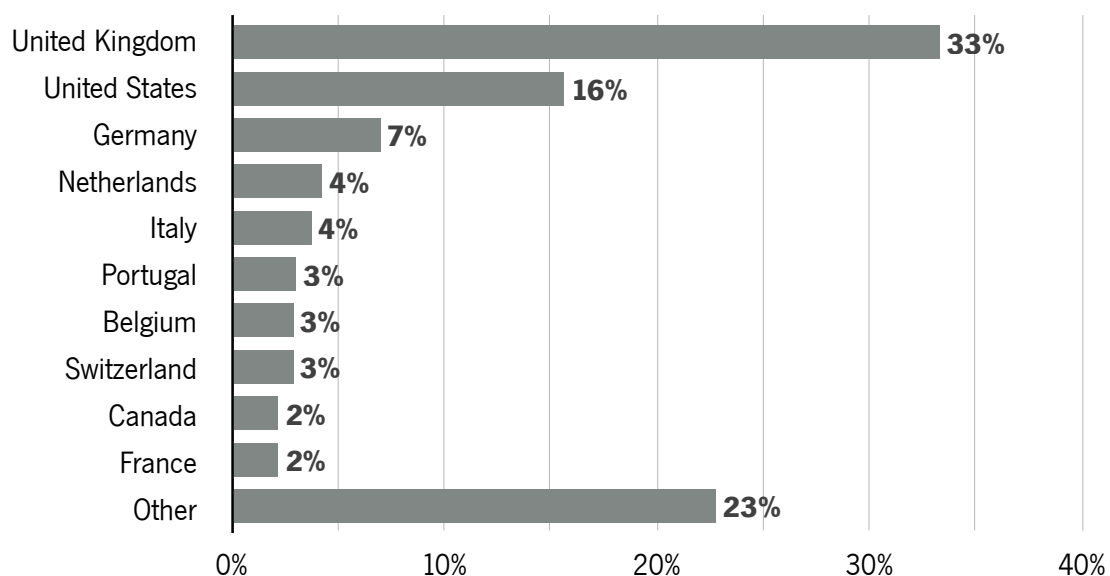


Figure 5.1: Top 10 countries that visited the survey site

There were a total of 342 responses to the survey, although not all responses answered all questions. The country of each response was not part of the questions, but anonymous web analytics show that there were a total of 588 visits<sup>1</sup> from many countries, mainly from Europe and the United States, see figure 5.1 for the top 10.

From the 259 respondents (76% of the 342 total) that answered the questions in the Profile section of the survey we can infer that responses mainly came from people in Universities and Memory institutions or content holders, followed by Government institutions, Small or medium enterprises, and Publishers or content producers (see figure 5.2). On the Profile section we can also see that the respondents mostly have the role of Digital preservation manager or Archivist, followed by Information technology, Researcher, Organisational manager and Technical support (see figure 5.3).

<sup>1</sup>How visits are calculated in Google Analytics: <https://support.google.com/analytics/answer/2731565?hl=en>

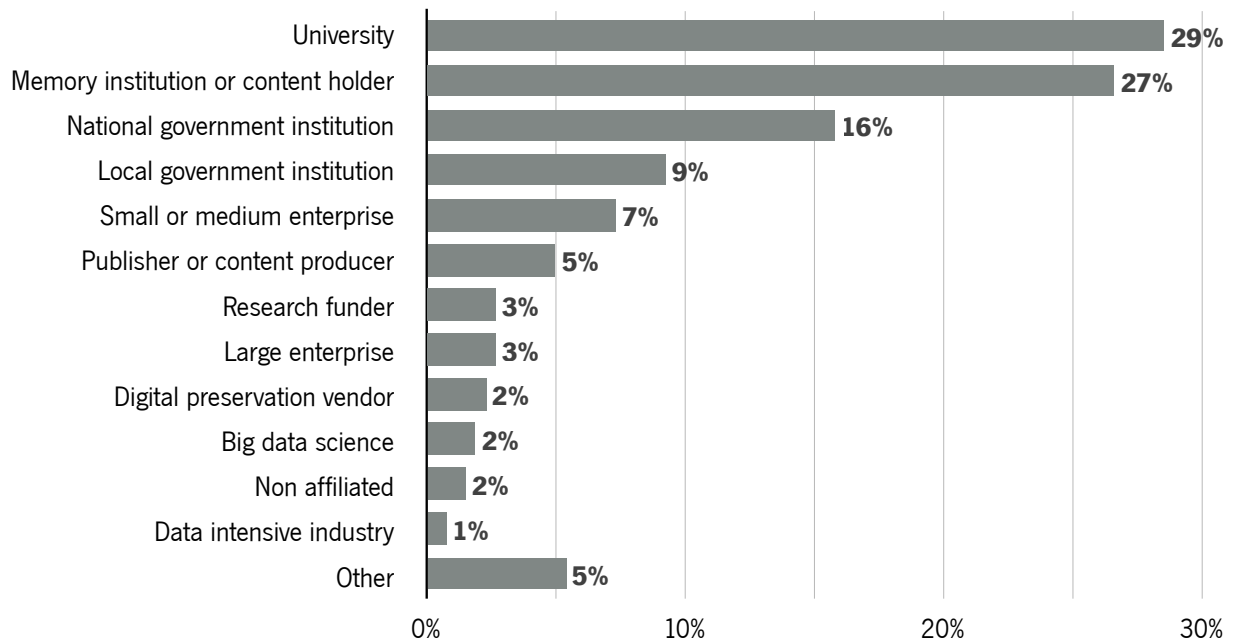


Figure 5.2: What descriptions fit your organization?

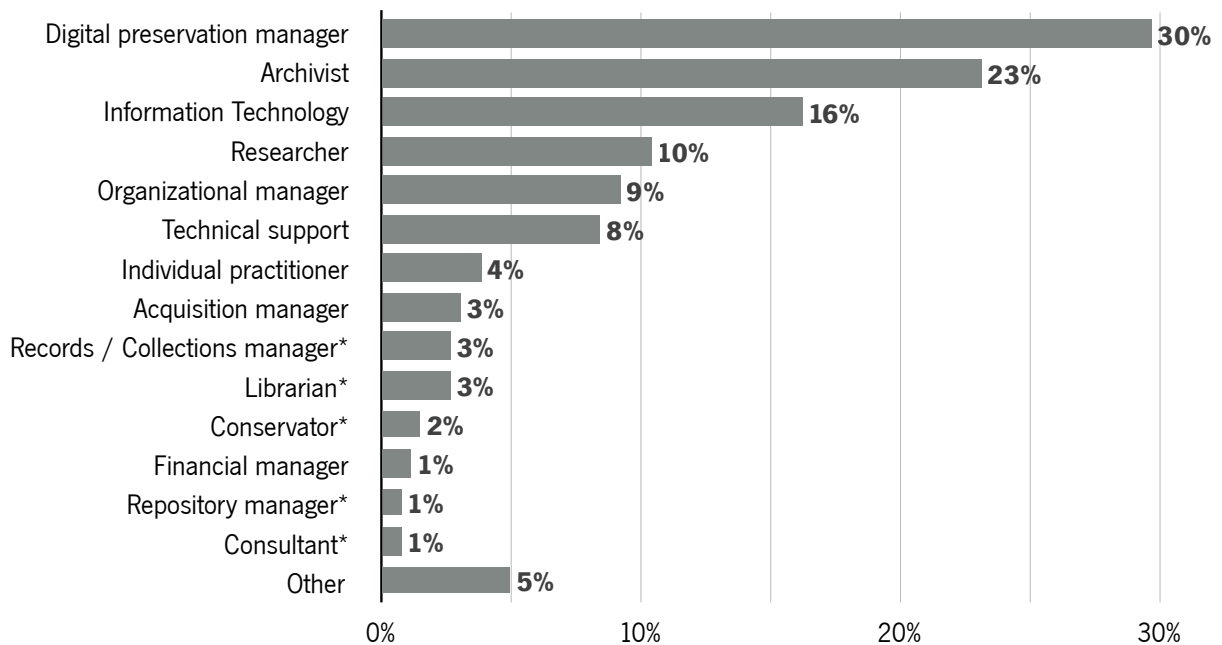


Figure 5.3: What description fit your role in the organisation? (\*' suffixed items were captured in the open option)

### 5.2.2 Preservation threats

Preservation watch mainly focuses on the detection of preservation threats (also called preservation incidents), that is the possibility that an event with negative impact on the preservation of a digital object might occur<sup>2</sup>. Knowing which preservation threats are more important for the community and which are already being monitored gives an important insight of what value can the preservation watch component bring to the community and what are the gaps this work should focus on.

To set the context for users and allow some normalisation of response and prioritisation, a suggested list of preservation threats was created by a focus group within SCAPE partners and the SCAPE user group, which was tested against a small audience at the SCAPE training event in Aarhus named "Effective, Evidence-Based Preservation Planning"<sup>3</sup>. The previous section presented the suggested list of preservation threats, with shortened names for convenience on identifying them in diagrams. The respondents were asked for their opinion on the importance of each preservation threat on a 1 to 5 scale, where:

- 1 Not at all important
- 2 Slightly important (Informational)
- 3 Important (Requires action but with low priority)
- 4 Fairly important (Requires action with average priority)
- 5 Very important (Requires urgent and immediate action)

Figure 5.4 shows the result of the question which was answered by 181 respondents. The importance of all threats have an arithmetic mean between 3.1 and 4.6, which shows that all suggested preservation threats are viewed as important and that some action is required. The black line on the limit of every bar illustrates the standard deviation, for example the minimum was 0.77 for file corruption and the maximum was 1.1 for incorrect action results, which means that more agreement was found for file corruption than for the incorrect action results, but in average there is a fair agreement on the importance of all preservation threats. Nevertheless, we can verify that file corruption and backup failure stand out in importance, which could be explained by the fact that these are the only two threats that relate to the physical preservation of content. Also, we can verify that outdated preservation plans, producer misalignment, and content not aligned with policies stand out by their relative low importance.

Figure 5.5 shows the current practice on monitoring these preservation threats. The 181 respondents that got to this point of the survey were asked which of the preservation threats they monitor, they don't monitor, or if they are uncertain. Backup failure stands out as the most monitored threat, followed by hardware and software platform obsolescence, file corruption and staff not enough or adequate. Consumer misalignment stands out as the least monitored threat, followed by lack of context information, producer misalignment, incorrect action results and outdated preservation plans.

An open question for respondents to point out other preservation threats that they perceived as important provided the following results. The responded threats below are rephrased and categorized to allow a better reading.

---

<sup>2</sup>See section 3.3 for more details on the terms

<sup>3</sup><http://wiki.opf-labs.org/display/SP/SCAPE+Training+Event+-+Effective,+Evidence-Based+Preservation+Planning>

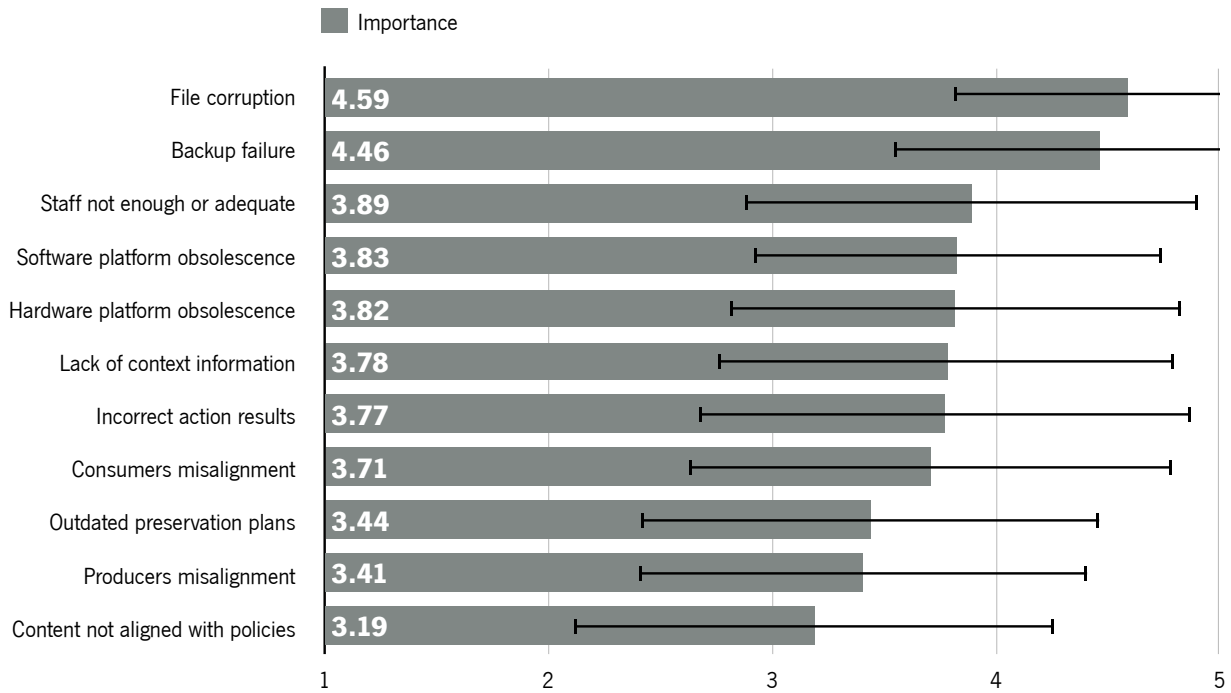


Figure 5.4: Importance of digital preservation threats

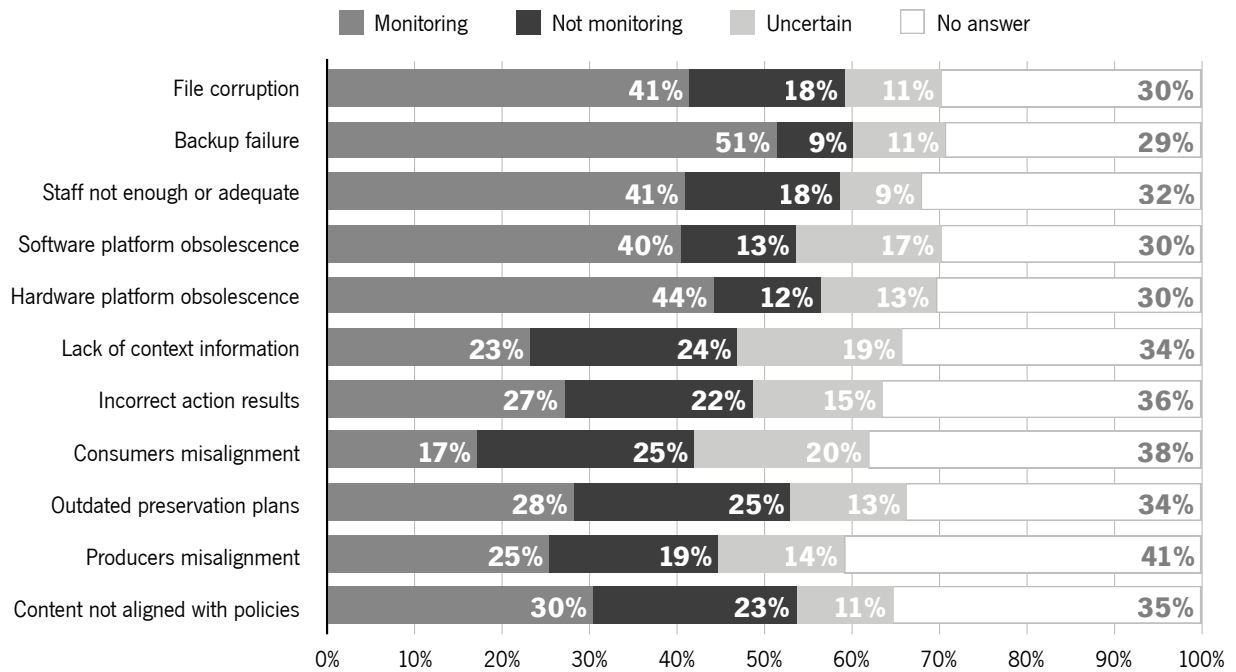


Figure 5.5: Current practice on monitoring digital preservation threats

Lack of content or metadata completeness:

- Undefined preservation metadata

Issues on content capture or production:

- Degradation of analogue media and capture hardware
- Not enough digital object representation capability or production environment information to match the content creator expectations (e.g. digital art in museums)
- Capacity to recover from analogue media and execute quality assurance

Technical capabilities or shortcomings:

- Capability to identify file formats
- Capability to validate files (e.g. check if PDF files are valid and well-formed)
- No adequate preservation action available (e.g. file format migration or upgrade has unacceptable quality and emulation is not possible/feasible/allowed)
- Long-term access to proprietary software
- Tools to validate and/or migrate are not robust

Archival storage and service:

- Security breach, malicious/accidental tampering
- Service availability
- Inadequate network services (especially for audio-visual assets)

Human or organizational environment:

- Lack of skilled personnel
- Digital preservation awareness
- Organizational or political change
- Loss of contact information for maintenance purposes (e.g. loss of tacit knowledge, know-how and ways to recover it)
- Lack of budget, management support or human resources
- Lack of clear standards for specific asset types (e.g. audio-visual or geographical datasets)

### 5.2.3 Preservation threats detection methods

A list of methods to detect or monitor each one of the previously suggested preservation threats was defined, in section 5.1, using the same focus group and trial survey as explained in the previous section. The 111 respondents who got to this part of the survey were asked for their preference on the method to detect each of the preservation threats on a 1 to 7 scale, and the practice of those same methods on a three options scale:

Preference:

- 1 Completely disagree
- 2 Disagree
- 3 Somewhat disagree
- 4 Neither agree nor disagree
- 5 Somewhat agree
- 6 Agree
- 7 Completely agree

Practice:

#### Use

The method is used by the organisation to detect the threat

#### Don't use

The method is not used by the organisation to detect the threat

#### Uncertain

Not sure if the method is used or not.

Figure 5.6 shows the preference on each of the suggested methods while figure 5.7 shows the current practice on the use of each of the methods. For each preservation threat, the respondents were also asked about other ways to detect that threat. Next, a summary of the results for each threat together with a description of the user suggested methods.

#### File corruption

The manual verification of file corruption was, understandably, one of the few methods with negative preference (i.e. below 4). The automatic file fixity checks are highly preferable and the method has a relatively high practice. Other user suggested methods were:

- User feedback
- Use of file validation (and characterization) tools like JHove, Exiftool, Droid, etc.
- Manual check of sampled files
- Comparison of similarity of visual content
- Automatic error detection on storage level
- Manual readability checks on sampled files

#### Backup failure

On detecting backup failure, the manual verification was also one of the few methods with negative preference. The method with higher preference and practice was to be alerted by the backup program on failure. There were some comments regarding this item as not helpful in inactive archives, as content does not change, or viewing backups as an outside matter, i.e. not related with the archive, managed by central IT or storage service provider. Other user suggested methods were periodic fixity checking of backups and disaster recovery testing to verify backup procedures.



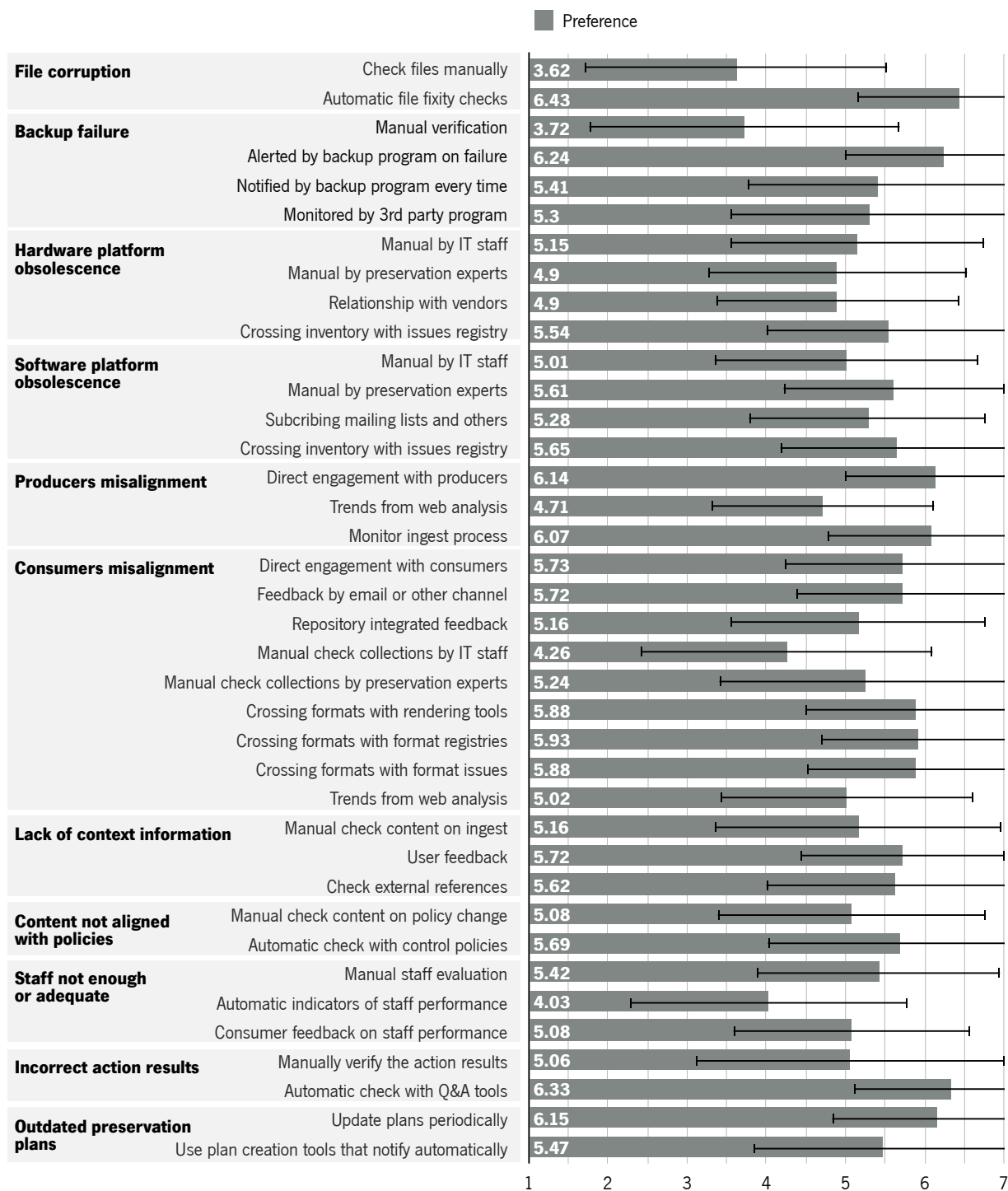


Figure 5.6: Preference on methods to detect or monitor preservation threats

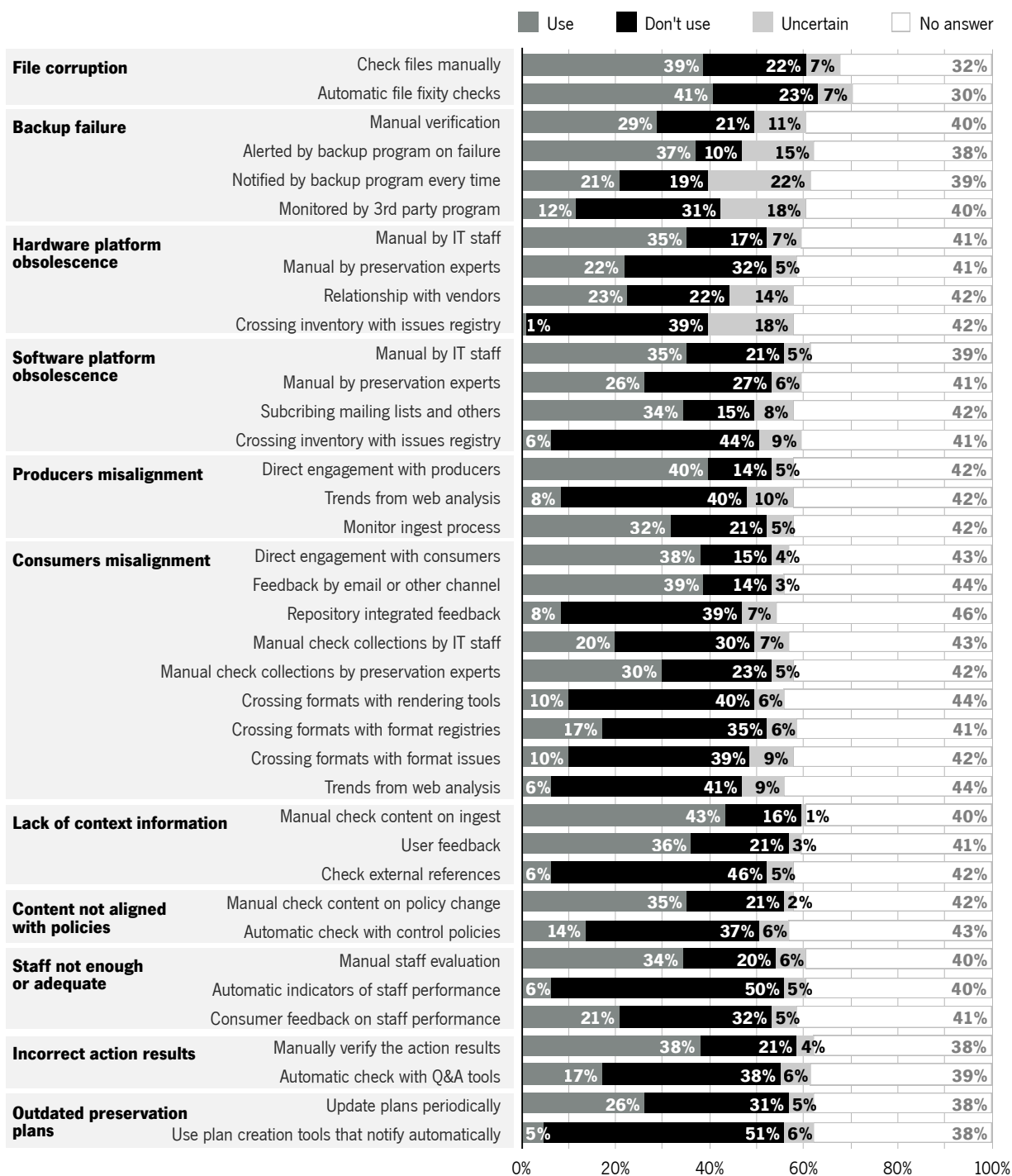


Figure 5.7: Current practice on methods to detect or monitor preservation threats

### **Hardware platform obsolescence**

All suggested methods had an average preference, but crossing inventory with issues registries stands out with a higher preference. Conversely, crossing inventory with issues registries had an almost non-existing practice (1%), whereas manual verification by IT staff stands out as the most used method. There were comments that this subject was not a subject of digital preservation but of IT management. Other user suggested methods were:

- Review of published hardware life-cycle milestones and roadmaps
- Awareness of the state of industry and common practices
- Monitoring of usage in reading room
- Maintain multiple generations of hardware

### **Software platform obsolescence**

All suggested methods have an average preference, but crossing inventory with issues registry stands out as the preferred method, followed by manual analysis by preservation experts. Conversely, these are the methods less used, with only 6% of organisations using the issues registry method. Other user suggested methods were:

- Relationship with software vendor
- Review of published software life-cycle milestones and roadmaps
- User feedback
- Creation of knowledge base of what software versions are compatible with files formats
- Monitoring of usage in reading room
- Information from the Web
- Engagement with other users of the same preservation software

### **Producer misalignment**

Direct engagement with producers and monitor of ingest process have a high preference and practice, with direct engagement being the most used and preferred method. Trends from web analysis had low preference and practice, possibly because they do not apply to all scenarios. Comments referred to the current non-existence of defined ingest policies, the general difficulty of controlling producers, and a recommendation to use open file formats. Other user suggested methods were to auto-detect compliance with machine readable transfer agreement, using feedback from users and exchange information with other institutions using the same content.

### **Consumer misalignment**

All suggested methods were well accepted by users, with the exception of manual check of collection by IT staff that showed a lower preference. The most preferred methods were cross-referencing formats with rendering tools, known format issues and format registries, have repository integrated feedback, and trends from web analysis. Most used methods were user feedback, direct or by email or other channel, and manual check by preservation experts. A user suggested method was to implement feedback channels (when such do not exist).

**Lack of context information**

All suggested methods are well accepted by users, with user feedback as the preferred method followed closely by the check of external references. Nevertheless, most organisations do a manual check of content on ingest, and very few check the external references (only 6%). Other user suggested methods were to engage with the producer, verify (or enforce) the minimum required descriptive metadata was provided, define and verify formal metadata policies and use the user feedback.

**Content not aligned with policies**

The automatic check with control policies has a greater preference than the manual check, but much fewer organisations use control policies automatic method than the manual method. There were no other recommended methods, but comments referred to the need of working with producers and intended consumers to set up the rules before content is sent, and also that checking if content is aligned with policies should not be done on ingest alone.

**Staff not enough or adequate**

Manual staff evaluation was the preferred and most used method, followed by consumer feedback on staff performance, while automatic indicators of staff performance were not found very appropriate nor very used. There was a comment referring to the need of organisational commitment to training and to include continued education and training as part of the digital preservation strategy. User suggested methods included monitoring the amount of backlog due to lack of staff availability, or content not acquired due to lack of staff expertise, or expected content not available.

**Incorrect action results**

Automatic check with quality assurance tools was by far the preferred method but conversely the least used, as compared with the manual verification alternative method. There were no other recommended methods but comments referred that testing with quality assurance tools would be the best solution but it would depend on the existence and accuracy of such tools. Also, another comment referred that quality assurance should not be completely trusted and that the original data should always be kept as a failsafe.

**Outdated preservation plans**

Update plans periodically was the preferred and most used method, but the automatic notification from planning tools was well accepted although not a very used method. Another user suggested method was to maintain preservation plans in a technical registry.

## 5.3 Analysis and discussion

This section presents an analysis of the results of the survey and a consequent discussion on their meaning and how they answer the research questions.

### 5.3.1 Sampling bias

It is important to recognise on a survey possible biases on the audience it reached to ascertain if there is a representative set of the target audience in question. Figure 5.1 shows that there is an undoubtable bias towards European countries (70%) and North America (17%). More specifically, there is a high attendance from the United Kingdom with 33% of the visits. This predominance can be explained by two facts: Firstly, the survey was done in support of the SCAPE project, which is funded by the European Union and is formed mostly by European partners. Secondly, many of the

institutions used as channels to communicate the survey: OPF, DPC, JISC and DCC, are from or have headquarters in the United Kingdom. Nestor is German and DLM and DigCurV are more international but mainly European with some liaison with the USA and Canada.

Similarly, figure 5.2 shows that universities and memory institutions or content holders make up more than 50% of all responses for the descriptions the respondent's organisation fits more adequately. This bias may also have been introduced by the channels used for the invitation which appeal preferably to the denoted organisation types, but due as well to the topic and terminology used to present the survey, which is more familiar for universities, libraries, archives and museums (LAM), in detriment of businesses and governmental institutions not related to universities or to LAM. This may also be due to the lack of digital preservation awareness within these business and institutional domains, together with a lack of proper bridging between the digital preservation terminology and the enterprise risk management terminology, which is more familiar to these target domains. Such bridging is now proposed in section 3.3 and could be further studied to reach these markets.

In another point of view, the sample may simply show the current composition of the digital preservation community, i.e. the group of people and institutions that are aware and interested on this subject, as a consequence of digital preservation being a part of their core objectives and sometimes even their main mission. Nevertheless, it must be noted that current results mainly inform us of the opinions and practice of institutions from Europe and North America (the latter with less representation) and mainly from universities and memory institutions or content holders, such as libraries, archives and museums.

### 5.3.2 Preservation threats priority

The survey focuses on two perspectives preservation threats can be measured on: the perceived importance and the actual monitoring practice. Although at a first glance one could assume a direct proportion relationship between the two, the fact is that it is not that simple. Some threats are more technically difficult to monitor than others. Furthermore, not all threats are as easy to be accepted as important by the set of roles and departments on a complex organisation that need to coordinate in order to deploy a method for continuous monitoring. Analysing preservation threats on these two perspectives together, as depicted in figure 5.8 a set of groups can be identified and the reasons for the grouping can be inferred by the typology of the preservation threats:

1. **Physical-level threats:** have higher importance and higher practice. They all relate with physical-level preservation and there is an agreement that they need urgent and immediate action to mitigate. Above 40% of organisations monitor them as it is common process on IT<sup>4</sup> across organisational domains.
  - File corruption
  - Backup failure
2. **Cross-domain logical-level threats:** have average importance and higher practice. Of average priority, they should be solved as soon as there are resources, nevertheless they are (relatively) highly monitored as they are common threats on IT or managerial processes across organisational domains.
  - Hardware no longer supported or degraded

---

<sup>4</sup>Information Technology, here to mainly refer information technology technical support

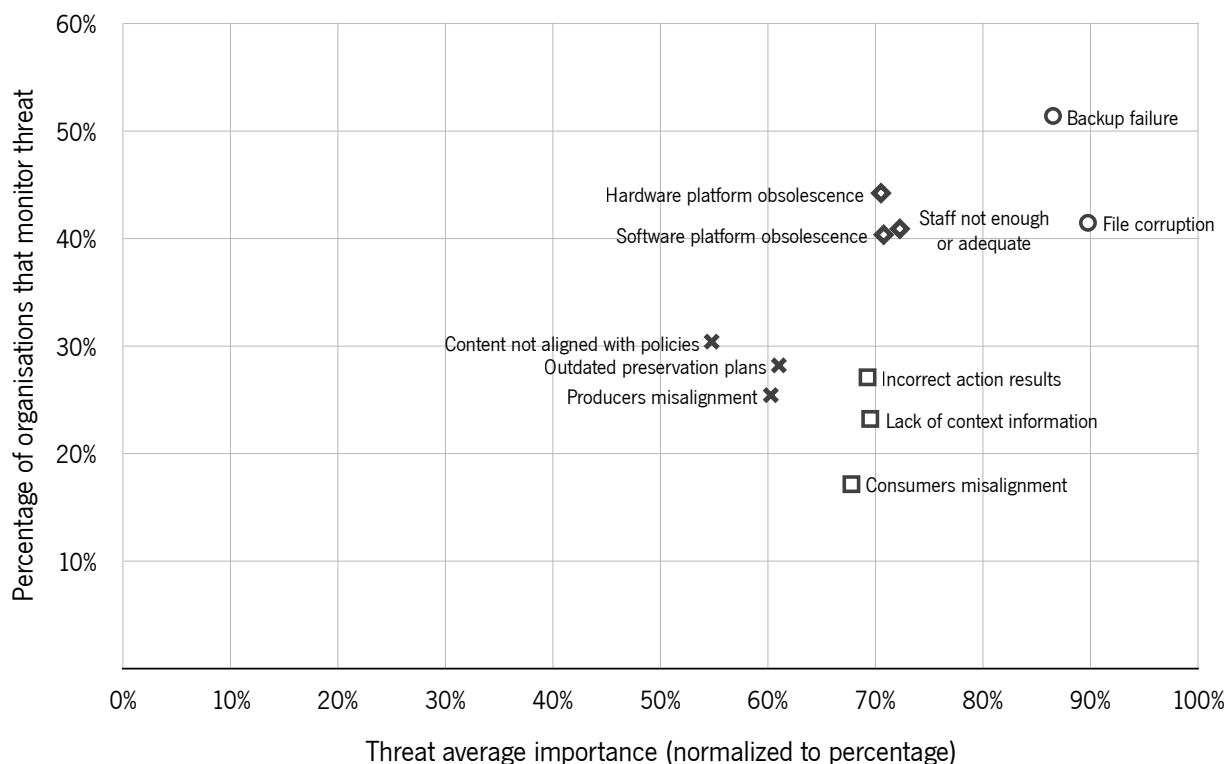


Figure 5.8: Preservation threat perceived importance vs. monitoring practice

- Software platform no longer supported or degraded
  - Organization staff is not enough or adequately trained to maintain content
3. **Digital preservation domain-specific logical-level threats:** average importance and lower practice. Considered of average priority but much less monitored than other logical-level threats as they are specific of the digital preservation domain.
- A relevant percentage of the consumers cannot read the disseminated file format (i.e. consumer misalignment)
  - There is not enough context information to understand the file content
  - Actions executed on files (e.g. file format migration) are not having expected results
4. **Best practice digital preservation procedural threats:** have lower importance and average practice. They relate to digital preservation procedures (plans and policies) and require actions but with lower priority, nevertheless a relative average amount of organisations (25% to 30%) monitor them.
- A relevant percentage of the producers cannot comply with the established ingest policies
  - Content does not conform with defined institutional policies
  - Defined preservation plans (e.g. defined preservation format) became outdated

It is understandable why physical-level threats have more priority than logical-level ones, as the logical-level depends upon the physical. Likewise, it is comprehensible that an institution would give priority to specific pressing threats than to follow and enforce procedures and best-practice guidelines such as plans and policies, which target secondary objectives as trustworthiness and credibility. Nevertheless, some of the threats could be averted with good policy enforcement and planning. Also, threats that are specific of the digital preservation domain are more difficult to monitor, as they need expertise that are more difficult to acquire, and require more effort to employ on a complex organisation where management and operation support would not fully understand the importance of the threat nor the methods to detect and monitor it.

In order to relate threat monitoring importance with its (lack of) practice by inquired organisations, so a useful prioritisation of the most important threats that are less monitored (i.e. neglected) could be drawn, the following score formula was devised. The formula normalises the importance mean to a 0-1 scale and multiplies it by the ratio of organisations that do *not* monitor the threat, relative to the set of responses that knew monitoring was or not in place (i.e. ignoring uncertain and no answer responses). This score gives equal weight to the threat monitoring importance and the (inverse of) current practice of monitoring, this is an assumption that could be further studied.

$$ThreatScore = \frac{ImportanceMean - 1}{4} \times \frac{NotMonitoring}{Monitoring + NotMonitoring} \quad (5.1)$$

Table 5.1: Preservation threat score (sorted by score)

Rank	Threat short name	Score	Difference to mean
1	Consumers misalignment	0.401	0.148
2	Lack of context information	0.352	0.099
3	Incorrect action results	0.307	0.054
4	Outdated preservation plans	0.286	0.033
5	File corruption	0.268	0.015
6	Producers misalignment	0.260	0.007
7	Content not aligned with policies	0.237	- 0.016
8	Staff not enough or adequate	0.218	- 0.035
9	Software platform obsolescence	0.175	- 0.078
10	Hardware platform obsolescence	0.152	- 0.101
11	Backup failure	0.127	- 0.126

Table 5.1 shows the score of all suggested preservation threats and highlights the ones above the mean score. The digital preservation domain-specific logical-level threats (group 3) fill up the top score with special mention for consumer misalignment. Next goes the best practice digital preservation procedural threats (group 4), with the exception of an outlier: file corruption. Although file corruption has a high practice, there is a relevant percentage of organisations that do not monitor it (18%) and its very high importance makes it still one of the threats to monitor with score above mean.

### 5.3.3 Answering the first research question

Again, the first research question is: *What are the most important and neglected digital preservation threats?*

The analysis presented on the previous sections clearly shows that the most important and neglected digital preservation threats, for the reached audience which mainly refers to the digital preservation community from Europe and North America and that belongs to universities and memory institutions or content holders, and taking into account an equal weight for the perceived importance and neglect, are the following ones by score order:

1. A relevant percentage of the consumers cannot read the disseminated file format
2. There is not enough context information to understand the file content
3. Actions executed on files (e.g. file format migration) are not having expected results
4. Defined preservation plans (e.g. defined preservation format) became outdated
5. File corruption
6. A relevant percentage of the producers cannot comply with the established ingest policies

### 5.3.4 Gaps on the preservation threats detection methods

This survey also indicates which are, in the view of the respondents, the most adequate methods to monitor the threats identified above. These results are a very important requirement input for the artefact that would respond to the second research question.

A formula to calculate a score of the most important monitoring methods to develop was devised below. As preference can have a depreciative value, it was normalised in a -1 to 1 scale on the *normalised preference (NP)*. In the same way, the lack of practice in using the method was normalised as a ratio to the *DUR* variable. Finally, the *MethodScore* tries to reveal the methods that need more attention by multiplying the preference by the lack of practice for methods that have positive preference. It also penalises methods with negative preference and high practice.

$$\text{Let } NP = \text{Normalized Preference.} \quad (5.2)$$

$$\text{Let } DUR = \text{Don't Use Ratio.} \quad (5.3)$$

$$NP = \frac{\text{PreferenceMean} - 4}{3} \quad (5.4)$$

$$DUR = \frac{\text{DontUse}}{\text{Use} + \text{DontUse}} \quad (5.5)$$

$$\text{MethodScore} = \begin{cases} NP \times DUR & \text{if } NP \geq 0 \\ NP \times (1 - DUR) & \text{if } NP < 0 \end{cases} \quad (5.6)$$



Table 5.2: Preservation threat monitoring method score

Threat short name	Detect method short name	Score	Difference to mean
File corruption	Check files manually	- 0.081	- 0.326
	Automatic file fixity checks	0.289	0.045
Backup failure	Manual verification	- 0.055	- 0.299
	Alerted by backup program on failure	0.158	- 0.087
	Notified by backup program every time	0.225	- 0.020
	Monitored by 3rd party program	0.313	0.069
Hardware platform obsolescence	Manual by IT staff	0.125	- 0.119
	Manual by preservation experts	0.179	- 0.067
	Relationship with vendors	0.147	- 0.098
	Crossing inventory with issues registry	0.503	0.257
Software platform obsolescence	Manual by IT staff	0.125	- 0.120
	Manual by preservation experts	0.273	0.028
	Subscribing mailing lists and others	0.132	- 0.113
	Crossing inventory with issues registry	0.481	0.236
Producers misalignment	Direct engagement with producers	0.181	- 0.063
	Trends from web analysis	0.198	- 0.048
	Monitor ingest process	0.273	0.029
Consumers misalignment	Direct engagement with consumers	0.166	- 0.079
	Feedback by email or other channel	0.156	- 0.089
	Repository integrated feedback	0.320	0.075
	Manual check collections by IT staff	0.052	- 0.193
	Manual check collections by preservation experts	0.182	- 0.063
	Crossing formats with rendering tools	0.502	0.257
	Crossing formats with format registries	0.434	0.188
	Crossing formats with format issues	0.500	0.254
	Trends from web analysis	0.293	0.049
Lack of context information	Manual check content on ingest	0.106	- 0.139
	User feedback	0.209	- 0.035
	Check external references	0.475	0.230
Content not aligned with policies	Manual check content on policy change	0.134	- 0.111
	Automatic check with control policies	0.414	0.168
Staff not enough or adequate	Manual staff evaluation	0.173	- 0.071
	Automatic indicators of staff performance	0.010	- 0.236
	Consumer feedback on staff performance	0.221	- 0.025
Incorrect action results	Manually verify the action results	0.125	- 0.120
	Automatic check with Q&A tools	0.534	0.290
Outdated preservation plans	Update plans periodically	0.387	0.142
	Use plan creation tools that notify automatically	0.450	0.206

Next there is a list of the top preferred monitoring methods for each of the top scored threats:

1. A relevant percentage of the consumers cannot read the disseminated file format
  - (a) Automatic cross-reference of file formats in your collections with information available on format registries
  - (b) Automatic cross-reference of file formats in your collections with tools that can render formats
  - (c) Automatic cross-reference of file formats in your collections with known file format issues
2. There is not enough context information to understand the file content
  - (a) User feedback
  - (b) Automatic verification that external references still exist (e.g. web sites)
3. Actions executed on files (e.g. file format migration) are not having expected results
  - (a) Run quality assurance tools on action results and automatically check against expected results
4. Defined preservation plans (e.g. defined preservation format) became outdated
  - (a) Update preservation plans periodically (e.g. yearly)
  - (b) Use tools to create preservation plans (e.g. Plato) and be automatically notified when assumptions taken may have become invalid
5. File corruption
  - (a) Automatic file fixity checks
6. A relevant percentage of the producers cannot comply with the established ingest policies
  - (a) Feedback by direct engagement with producers
  - (b) Monitor ingest process (e.g. SIP rejection statistics)

For the consumer misalignment threat, all methods identified above have similar preference and practice. All of them refer a similar technique, automatic cross-reference of file formats in the collections with outside registers. What differs is the registers content, where registers of file formats are much more well established than registers of tools or format issues. Examples are the PRONOM<sup>5</sup> and UDFR<sup>6</sup> file format registries, versus the COPTR tool registry<sup>7</sup>, and scarce file format issues registers. Due to this, the method of cross-referencing with file format registries has a higher practice, 17%, versus the 10% of the other options, and consequently a smaller score. Nevertheless, due to the low practice and higher preference, cross-referencing with file format registries still seems to be the best detection method and the one which will have the biggest impact on the community.

---

<sup>5</sup><https://www.nationalarchives.gov.uk/PRONOM/>

<sup>6</sup><http://udfr.cdlib.org/>

<sup>7</sup><http://coptr.digipres.org/>

For the lack of context information threat, the user feedback is by far the most used monitoring method (36%), as it can use any of the normal channels available in institutions, as email or direct contact on the reading room. Nevertheless, the method of verifying that external references still exist has a similar preferability and much less practice, leading to a bigger score which points to a gap on the monitoring techniques.

On the incorrect action results threat, the use of quality assurance tools has a very high acceptance (6.33) albeit low practice (17%) which points to a high score and can be identified as a gap.

Periodic update of preservation plans is very well accepted method to deal with possible outdate of preservation plans but its use is relatively low (26%). Nevertheless, there is a high acceptance for planning tools that would allow automatic notifications, as in a possible with an integration of Scout and Plato.

Automatic file fixity checks are the most well accepted and used method for detecting file corruption, but its practice is still surprisingly low at 41%, with 23% of responses specifically detailing they do not run automatic file fixity checks.

The producer misalignment is well monitored by both direct feedback and the monitoring of the ingest process, but as the monitoring of ingest process has a smaller practice (32%) it achieves the bigger score.

## 5.4 Final remarks

This survey answers the first research question by identifying the most important and neglected preservation threats and lays the foundation for the second by prioritising the monitoring methods a software artefact could employ to effectively identify and continuously monitor preservation threats. In summary, the top scored methods are:

1. Automatic cross-reference of file formats in your collections with information available on format registries
2. Automatic verification that external references still exist (e.g. web sites)
3. Run quality assurance tools on action results and automatically check against expected results
4. Use tools to create preservation plans (e.g. Plato) and be automatically notified when assumptions made become invalid
5. Automatic file fixity checks
6. Monitor ingest process (e.g. SIP rejection statistics)



## Chapter 6

### SCOUT - A preservation watch system

To achieve automated preservation watch, a preservation watch system must be developed to enable automated monitoring of preservation threats by collecting, fusing and analysing information from various sources. These sources enable discovery of relevant aspects of the world, i.e. the entities and their properties about which information should be gathered, so preservation threats can be identified. The relevant aspects of the world reveal internal and external influencers that cause or indicate preservation threats to the digital content. These influencers need to be continuously monitored and when a threat is identified the relevant parties must be immediately notified to decide and deploy mitigation actions while there is time to contain the possible damage.



Figure 6.1: Scout base concept

Figure 6.1 shows the base concept behind Scout: to gather and monitor information from several different sources, using it to identify threats and notify the relevant parties. Linking information *across sources* would allow to uncover preservation threats by answering critical questions such as:

- Do I have any file format in my content that is rare to appear in other repositories or web archives<sup>1</sup>?
- Does my content reference outside content that no longer exists?
- Are the actions that I have chosen to mitigate my threats running successfully and with minimum quality?
- Are the actions I intend to execute having good results for other users?
- Is my content physical integrity assured?
- Is the storage technology I intend to purchase or deploy having a good integrity assurance rate for other users?
- Are my producers having problems sending me content?
- Are other users more successful in captivating and accepting producers and their content?
- Is any of my content files characteristics non-compliant with my policies?

The example questions listed above refer to the result of the survey presented in chapter 5 and detail possible questions on the information fetched from external sources that would help identify the preservation threats deemed as important. Some of the presented questions imply sharing of information between users, which would greatly improve the value of the created knowledge base. As the preservation community becomes aware of this platform and the mutual benefits coming from synergetic data collection, the possible questions and size of the knowledge base should evolve, allowing new threats to be detected and new scopes to be introduced.

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 (Kulovits et al., 2009). 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 the consumer of the answers. Essentially, the watch system 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. (Becker et al., 2012) (Faria et al., 2012a)

---

<sup>1</sup>i.e. is not well disseminated or is a niche format

## 6.1 Requirements

The requirements for Scout link closely to the second research question: *Can these threats be detected using a formal representation of the information about the world that is automatically monitored and collected?* To collect, link and analyse information as described, a system should aim for the following high-level goals (Becker et al., 2012):

1. **Enable to pose questions about entities and properties of interest.** Software components and human users will be able to pose questions to Scout and receive answers about the measures. They can also deposit conditions to receive a notification upon significant changes.
2. **Collect information from different sources through adaptors.** Different sources will be relevant for Scout. Each source of information provides specific knowledge in different information models that will have to be mapped, normalised, merged and linked.
3. **Act as a central place for collecting relevant knowledge that could be used to preserve an object or a content set.** Information for the preservation of the object or content set shall be collected and linked so that Scout provides a uniform reference point for gathering information about a variety of aspects.
4. **Notify interested agents when an important event occurs** through configurable notification channels.
5. **Integrate with the decision-making process**, i.e. planning, to start planning when threats are found, and monitor if the base assumptions taken by planning are continuously valid, making sure that the decision taken is still the best and notify when a re-evaluation is needed.
6. **Act as an extensible platform.** Allow additional information sources to be easily added and connected.

## 6.2 Architecture

A suitable design for the proposed system is a three-tier architecture (Fowler et al., 2002), as depicted in figure 6.2. Information comes from the outside world via source adaptors on the interface tier (top layer). **Pull source adaptor plugins** fetch information from external sources of information and filter, normalise, structure, aggregate and anonymise the gathered information. Adaptors can also run in the external source, pushing the information into Scout via the **push source adaptor API**. New source adaptors can be added to the system at any time, they are managed by the **adaptor manager** and periodically run by the **scheduler**. Every source adaptor structures the gathered information to the way it is modelled in the **knowledge base** and delegates it to the **data merger** on the business logic tier (middle layer).

As different sources might refer to the same entities and properties of the world, inconsistencies and incompatibilities between different sources of information are to be expected. For example, format registries sometimes serve contradictory information about file formats. The **data merger** solves this by merging the data and treating incoherencies 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 questions that relate and interlink

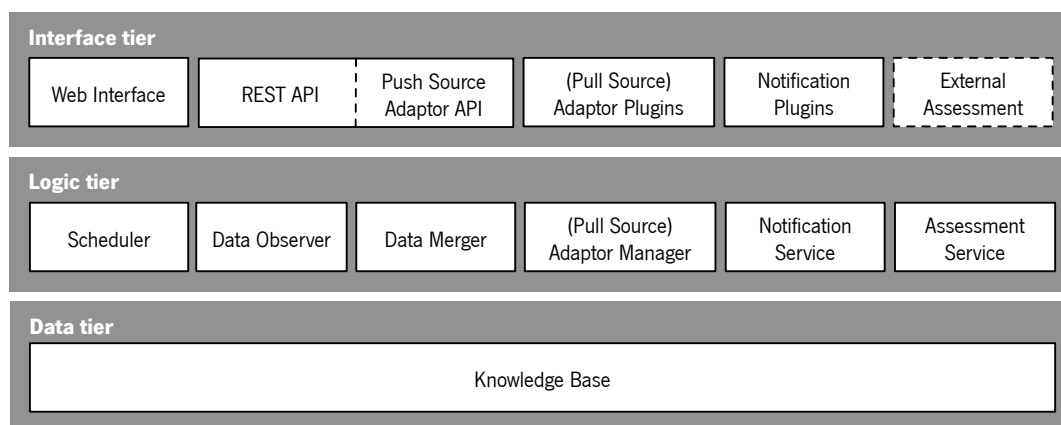


Figure 6.2: High-level architecture of the Watch component

several entities. For example, the file format distribution of a repository, given by its content profile adaptor, could 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 information cross-reference and provides a query engine that is expressive enough to cross-examine the different properties and define conditions that identify symptoms of threats or opportunities. Whenever the knowledge base is changed the **data observer** is called, so it can notify all components interested in a specific type of information.

Back to the top layer, the **Web interface** serves as the human interface for browsing the knowledge and to create triggers so they are notified when significant events occur. The **REST API** allows the same functionalities to software components. These triggers are kept in the knowledge base and the **scheduler** and the **data observer** determine when they need to be re-assessed, either by periodic checks or by monitoring the changes in the knowledge base.

The assessment is done by the **assessment service**. If the result contradicts the conditions, interested parties are alerted via the **notification service**, which uses **notification plugins** to send messages to predefined interested parties. Additionally, the system could support an **external assessment**, which would allow for more complex and deep assessment systems, like the Plato planning component, to be directly used in the assessment phase, allowing for an optimisation of the system that would minimize false alerts (i.e. false-positives). However, this feature is not a priority and will not be the focus of this artefact. (Faria et al., 2012b)

The next sections describe how information is kept on the knowledge base, how it is collected from the outside world, and how it is used to detect and monitor threats.

### 6.2.1 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 a model of interesting aspects of the world to 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 an entity might define as the entity can only define a **property value** for a **property** defined on its type. A property can also restrict how the data is formatted in the property value by defining its **data type** and further describe it with **rendering hints**, to be able to correctly present it to human users. A property is, therefore, the generic definition of an attribute that an entity of a defined 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.

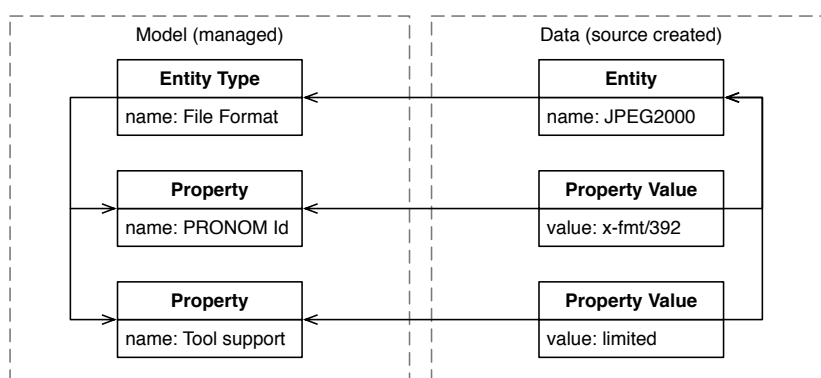


Figure 6.3: Example of managed model of the world and source created information

Figure 6.3 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<sup>2</sup>. 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 can 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 while time passes by 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**. A **source** is the documentation that describes the external source of information that provided the measurement. The **source adaptor** documents which program logic, e.g. pull source adaptor plugin name and version, which extracted the information from the source and mapped it into the internal knowledge base model. A **measurement** connects the source adaptor to the provided value and documents the date in which the value was observed. Several

<sup>2</sup>Properties can prescribe the format of the value, ignored here for sake of simplicity

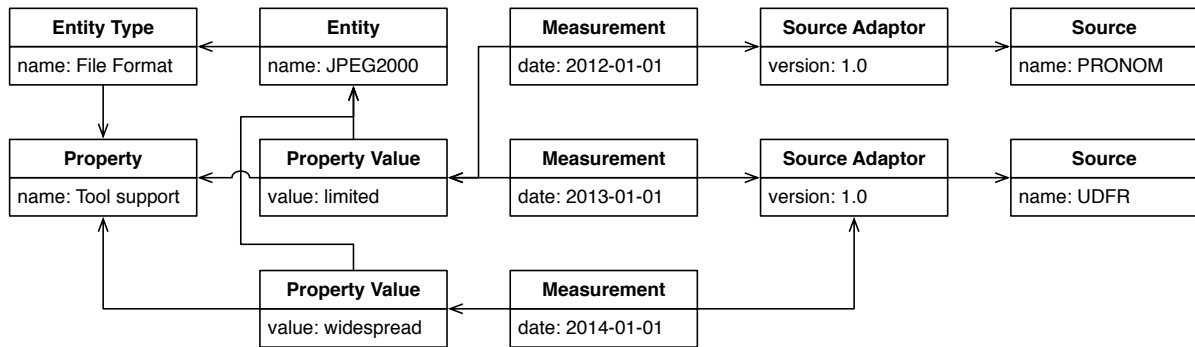


Figure 6.4: Example of knowledge base with history and provenance

measurements can connect the same pair of property value and source adaptor, whenever the same value is repeatedly observed.

An example of the knowledge base with value history and provenance is depicted in figure 6.4. 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'.<sup>3</sup>

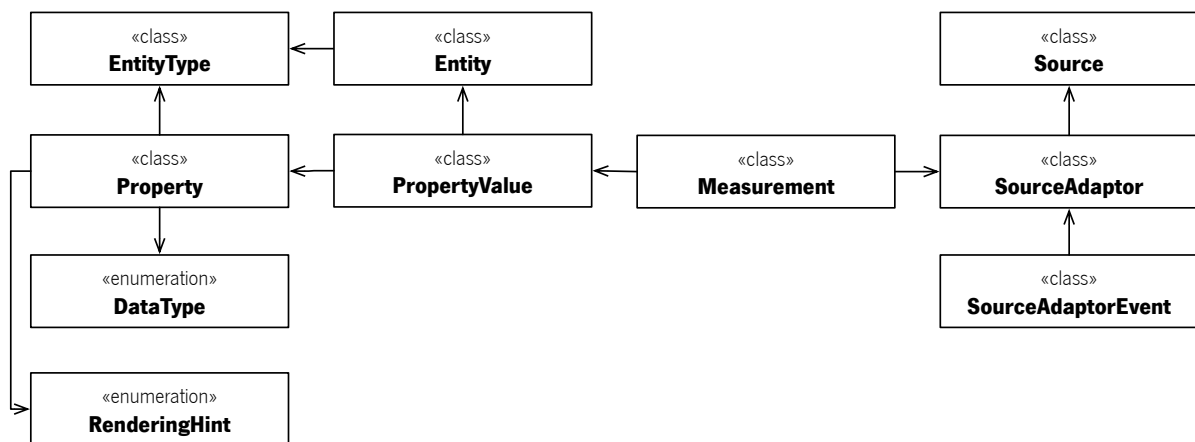


Figure 6.5: UML class diagram the knowledge base domain model

Figure 6.5 puts all concepts together in the form of an UML class diagram. A more complete version of this diagram is available on appendix A.2. The figure shows how data types and rendering hints relate to the property, and the more complete version in the appendix shows also how the values of properties align with the data types. The version on the appendix also includes the attributes of every class presented on this figure.

<sup>3</sup>Please note that the example is purely fictional and presented only for illustration purposes, neither PRONOM nor UDFR have enough information about file formats tool support to be able to infer if it is limited or widespread.

### 6.2.2 Information sources

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 categorisation (Becker et al., 2012):

#### Repository content

A content profile provides statistical data about a repository 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 threats. 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. Content profiles can be shared without intellectual property right concerns and when compared between organisations can provide a very important insight on how aligned the organisations are and if they are alone on their problems.

#### Repository events

Repositories perform continuous operations that can provide valuable information to feed into decision making. This includes validity check as well as ingest and access operations (What content is being produced and at what rate? 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 repository and preservation activities are successful. This also supports anticipation of trends in the repository operations and usage and give an important feedback on the alignment of producers and consumers towards the objectives.

#### Format registries

Information on format registries can help to enhance the knowledge from content profiles by expanding it with known properties of the existing file formats, which may be needed to compare with organisational objectives. Format registries can also help to detect the emergence of new file formats which can bring opportunities for better quality or operational cost reduction. 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 (Tarrant et al., 2011), 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. Characterisation components provide measures of properties in content, such as a format identification or the presence of encryption or compression. Quality assurance components, finally, perform quantitative or qualitative analysis of the quality of preservation actions, such as algorithmic comparisons of original and converted objects or run-time analysis of rendering quality.

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 (Becker et al., 2009). 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 (Becker et al., 2009). 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 (Kilbride, 2010). On the other hand, research experiments are commonly done to answer domain-specific preservation questions (see section 7.4 for an example). Such research results are a fundamental input for other institutions with similar problems, specially when resources are lacking.

### 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, such as using the *SCAPE digital preservation ontology*<sup>4</sup>, they will provide a critical starting point for monitoring fulfilment of these objectives on specified indicators. This information can be fed into a knowledge base to enable direct queries resolving objectives against the state of the world.

### Human knowledge

Finally, human users could also be able to insert information about every possible entity (objects, format, tool, experiment, repository status, etc.). This method would serve as a fallback when no other public source of information exists, such as information that is kept on internal institutional infrastructures, or as tacit knowledge (i.e. not formalised in any document) and that could be of public service. Scout can become a central repository where this information can lay, becoming on itself and important source of information, if the community embraces it.

---

<sup>4</sup><https://github.com/openpreserve/policies>

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 harvesting of information through a loosely coupled architecture of information adaptors. It should allow the update, addition and replacement of sources, so that the configuration of complementary information sources can evolve over time.

Also, 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 (Antunes et al., 2011). 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 content 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.

### 6.2.3 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 Scout's model of the world.

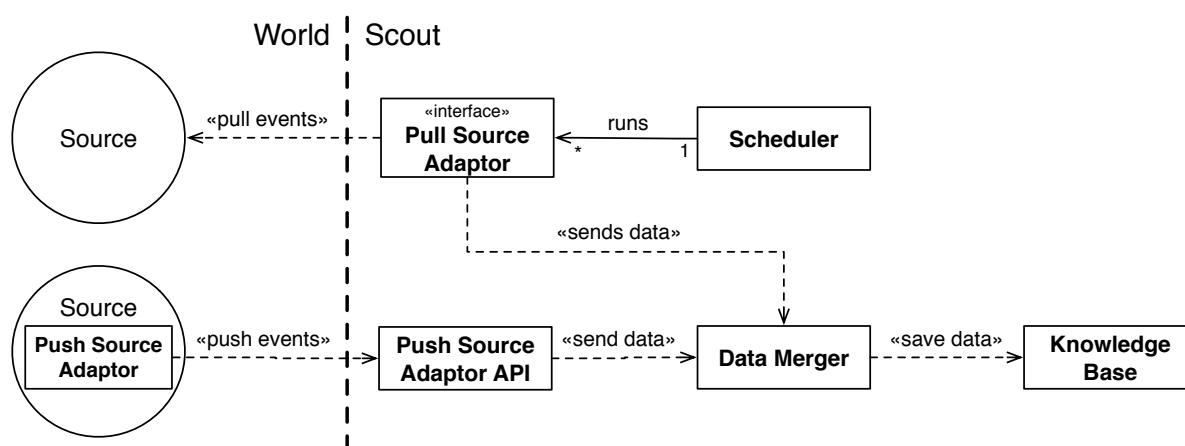


Figure 6.6: Source Adaptor model

A **source adaptor** is a software component that can be deployed in two ways: inside Scout, frequently polling the external sources for relevant information (pull source adaptor), or inside the external sources (push source adaptor), pushing information into Scout via the **push source adaptor API** (figure 6.6). The **pull source adaptor** must conform to a defined interface and will be run automatically by a scheduler. The **push source adaptor** runs freely on the external source but must conform with the API specification. The push and pulled data is sent to the data merger to be processed and integrated into the knowledge base.

The source adaptor must filter the data of interest of the external source, aggregate data, analyse and infer new knowledge when possible, map and normalise data to the model specifications and anonymise personal data. Depending

on the external source, a push or pull source adaptor might be more adequate. For example, if the external source needs to be agnostic of Scout, 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 anonymise the information before sending it to Scout.

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

## 6.2.4 Extensibility via plugins

One important design requirement for Scout is the easy extensibility of behaviour, specially for adding support for new information sources. To support adding behaviour to the system, for example new pull source adaptors, without it implying a change on the system source code, an extensibility design pattern would need to be followed. To achieve that objective, a plugin architecture was selected, to be developed as a simple service provider interface instead of using existing plugin framework due to their overhead.

Selected architecture is depicted in figure 6.7. A generic plugin interface defines the minimum set of information needed for a plugin. Every plugin is identified by its name and version, and can be further documented with a description. There are two different types of supported plugins: 1) notification plugin to add new methods of sending notification to interested parties, 2) adaptor plugin, i.e. a pull source adaptor plugin, which allows to add new ways of gathering information from the outside world. The plugin also defines a list of configuration parameters, which document how a new plugin instance can be configured. A generic plugin would also have methods to initialise and shutdown, setting up any needed resources before starting, and tearing down any used resources after ending.

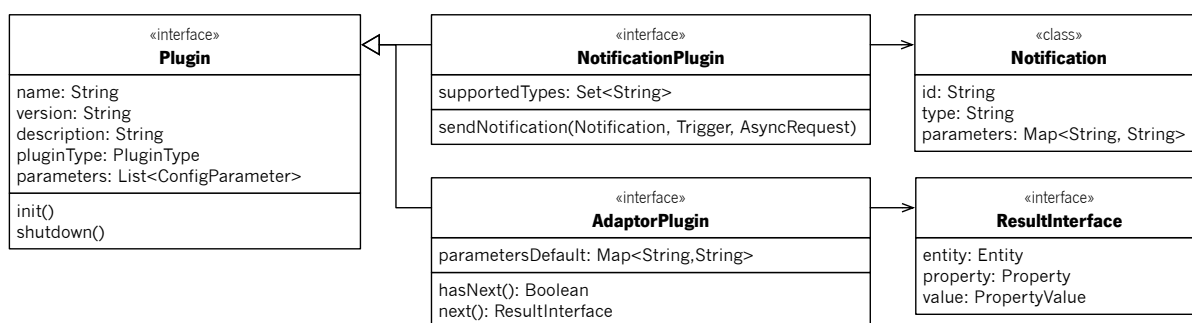


Figure 6.7: Plugin UML class diagram

The plugin manager monitors a folder on the file system<sup>5</sup> periodically and all found JAR files<sup>6</sup> are loaded into a separate class-path. A special metadata key is used to find all classes in the JAR that adhere to the plugin interface and they are loaded in the system, registered separately by their type.

<sup>5</sup>By default, the plugin folder is at /usr/local/scout/plugins

<sup>6</sup>A JAR (Java Archive) is a platform-independent file format that serves as a container for several compiled Java classes, associated metadata and resources (e.g. images) and is used to distribute software libraries (i.e. collection of implementations of behaviour) in the Java platform.

A new notification plugin must implement the NotificationPlugin interface, announcing its supported notification types and providing a method to send a notification to a user. The notification types were left purposely open so they can be easily extended. A notification will carry all needed information to inform the user of the significant event that has occurred.

A new (pull source) adaptor plugin must implement the AdaptorPlugin interface, announcing a default value of their required parameters and methods to iterate over the results it provides. An adaptor will fetch information from outside sources and provide information to the knowledge base encoded as entities, properties and property values. The adaptor does not need to care if the information already exists on the knowledge base as a data merger and the knowledge base will take care of aligning the provided data with the knowledge base data model, will create new property values when needed, and will document the measurement connecting to the correct source adaptor and, transitively, the source of information.

A plugin defines an extension to the behaviour available in Scout. Each one can have several configurable instances running in parallel. An adaptor plugin instance will be run periodically by a scheduler in a 60 minutes' period by default. The period can be configured by the plugin itself. A notification plugin will only have one instance by default, managed internally, and will be executed every time a notification from a type the plugin supports is emitted by the system.

### 6.2.5 Collecting information

The collection of information from external sources begins by starting all components described above on the tree-tier architecture. The **scout manager** is the main orchestrator of the system, putting together all components and managing their life-cycle. It begins by ordering the adaptor manager to load all active source adaptors, which previous state was saved on the knowledge base, and add each one to the scheduler, so they run periodically.

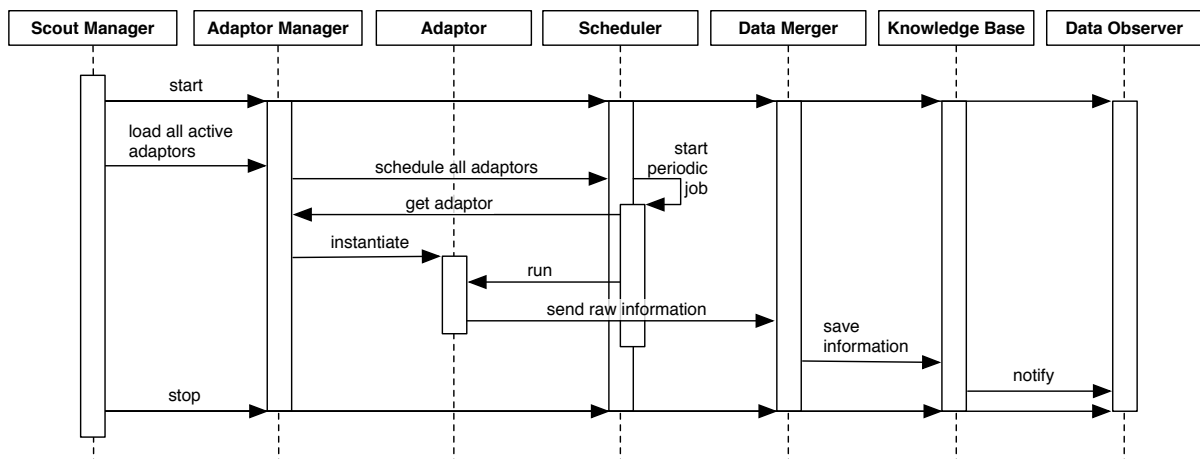


Figure 6.8: UML sequence diagram detailing how components interact to collect information from external sources

At each periodic run, which happens on system start and every 60 minutes by default, the scheduler starts a new job which requests the adaptor manager for an instance of an adaptor and uses it to scan the source and fetch new information (in the form of a result interface). This information is then sent to the data merger which enforces a

set of rules for pre-processing the information, preparing it to be saved in the knowledge base. The knowledge base integrates the new knowledge into its internal representation and notifies the data observer, which can pass on the change events to the interested components. This process is illustrated in figure 6.8 in the form of a UML sequence diagram, which details the life-cycle of each component and how messages and requests are passed through them.

### 6.2.6 Optimising value and measurement representation

Every time an adaptor runs, the current set of property values of one or more entities are extracted from the source and added to the knowledge base. This new observation is the verification at a discrete moment in time that a property has a certain value. But it is not in the scope of the adaptor to know the previous values of this property, or if this new observation derives a change on the property value or is simply a confirmation that it remains the same as before.

Being aware that a new observation reports a change or a confirmation of a property value is an important information for Scout and its only possible in the moment the new information is merged into the existing knowledge base. As the new property value observation is sent to be integrated, it is compared with the previous values of this property. If, in one hand, it reports that it is a confirmation of the existing value, then only a new measurement is created documenting that observation. In the other hand, if it reports that it is a change of the value (i.e. the new value is different from the previous one), then a new property value is created, with an incremented version number, to document this value change, and a new measurement documents the date of the observation.

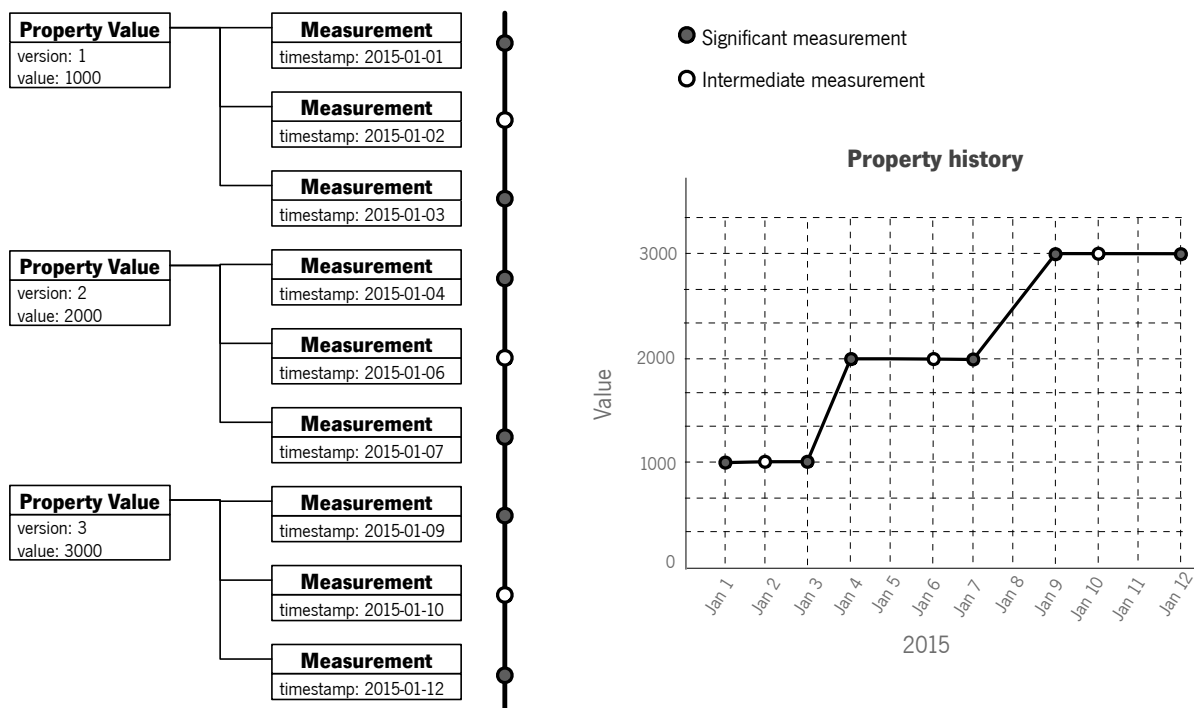


Figure 6.9: Property value versions and significant measurements

The left side of figure 6.9 depicts the creation of the property values and measurements, showing a property that is extracted from the source by an adaptor on a daily frequency, and that changes every three days, having therefore three



measurements that document the daily observations for each of the three different values it took. The same figure also demonstrates that not all measurements have the same importance. As the discrete values are connected to form a chart line to illustrate the evolution of the property throughout time, some of the measurements are superfluous and could be ignored in this representation. This by no means would signify that the measurements are of no importance, in fact they are the empirical proof that such a simplification could be done with no significant information loss. Therefore, and for the objective of optimising the chart representation performance, although all measurements are recorded in the knowledge base, their significance to the chart representation is pre-calculated, so only the significant measurements are reported to interface to create the property chart.

### 6.2.7 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 threats. These questions relate to entities and properties and may need to use cross-referencing or analyse how property values evolve throughout time (e.g. sudden growth in content volume) to identify certain threats.

Not all changes on the questions output are important: one would not want to be alerted when, for example, the content volume mildly increases some Megabytes. 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. threshold definition) that define when the result of a question, i.e. an event, becomes significant and needs attention.

Using an ontological knowledge base, conditions can be merged into questions using SPARQL<sup>7,8</sup>, a semantic query language able to retrieve and manipulate data in RDF<sup>9</sup> format. These SPARQL queries can be complex and verbose but highly reusable, as users might have similar questions with different parameters. To be able to easily reuse the same question with different parameters the **question template** is used. It defines all information needed to generate questions based on parameters. On the question template a query is defined by three parts: 1) a generic SPARQL query where possible parameters are defined as variables; 2) a list of question template parameters that define the rules to find bindings for the parameter variables; 3) a request target which is the data type of the variable that will be the query's result, which can be any one of the **request target** enumeration. See figure [6.10](#) for an UML class diagram that details these relationships.

The question template enables the creation of a predefined form in the user interface that allows the user to create a personalised question by simply setting up some parameters. The result of this selection is a list of query bindings, which simply bind a SPARQL variable to a certain value. The list of query bindings, together with the SPARQL query and request target from the question template allow the creation of a new question that aims to detect a significant event.

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). A notification must define its type and its parameters, e.g. a notification of the email type must define the recipients of the email as parameters. Types of notification are extensible so it easily adapts to the user community needs.

In essence, a question gives a result and evaluates it by a condition in SPARQL so that when significant events are found notifications are sent. This question can be set to be re-assessed frequently on a defined period using the

---

<sup>7</sup>SPARQL is a recursive acronym meaning SPARQL Protocol and RDF Query Language

<sup>8</sup><http://www.w3.org/TR/rdf-sparql-query/>

<sup>9</sup>RDF means Resource Query Framework

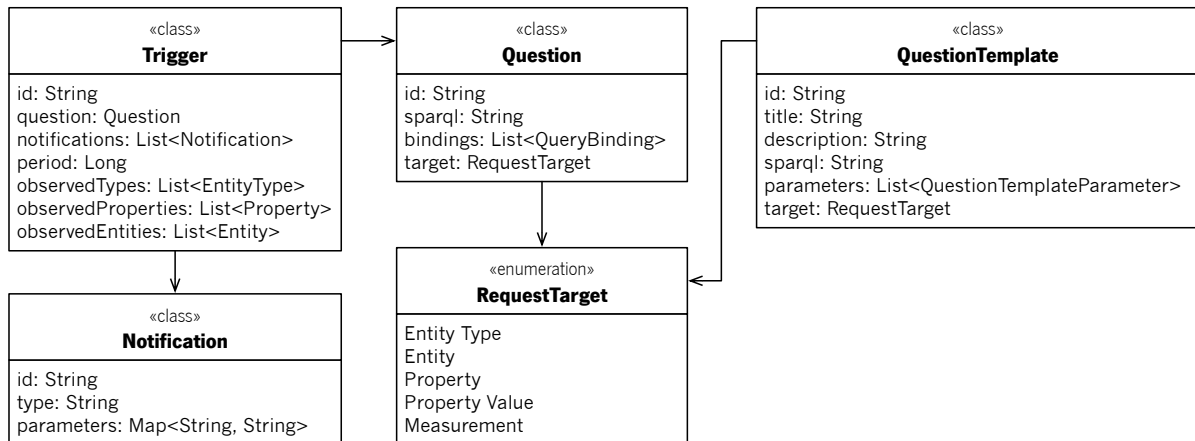


Figure 6.10: UML class diagram of trigger, question and notification domain model

scheduler, or every time new information of a certain entity type or property is updated using data observer. A **trigger** is therefore the combination of these components: the question and the list of notifications together with information on the period this question should be re-assessed and/or the list of entity types or properties that need to be observed as changes on them should enact the re-assessment.

### 6.2.8 Assessing triggers to identify threats

The main orchestrator, the **scout manager**, starts up the system and loads all previously defined triggers, which have been persisted on the knowledge base. For each one of them, it registers listeners on the **data observer** if the trigger defines entity types, entities or properties that should enact re-assessment. Also, triggers that define a period for frequent assessment are added to the **scheduler**. The scout manager can also add new triggers on runtime, via the Web Interface or the REST API. See figure [6.11](#) for a UML sequence diagram of this process.

The **scheduler** starts a new job, on the frequency defined by each trigger, to request an assessment of the trigger to the **assessment service**. Similarly, the **data observer** requests the assessment of a trigger every time it detects a change on the observed part of the **knowledge base**.

The **assessment service** uses the question inside the trigger that defines the SPARQL query, the query bindings and request target to find significant events. If any significant events are found, it uses the list of notifications of the trigger to alert all relevant parties via the **notification service**, passing on all information about the question that gave rise to the notification and all found significant events.

The role of the relevant parties that received the notification is now to focus on the preservation threat that has been raised to attention and act accordingly to mitigate the threat using the planning and operations processes. As the planning process selects the action to perform and includes quality assurance steps into the action epilogue, it may also produce and install triggers in Scout that monitor the output of the quality assurance step and verify if the action had the expected results.

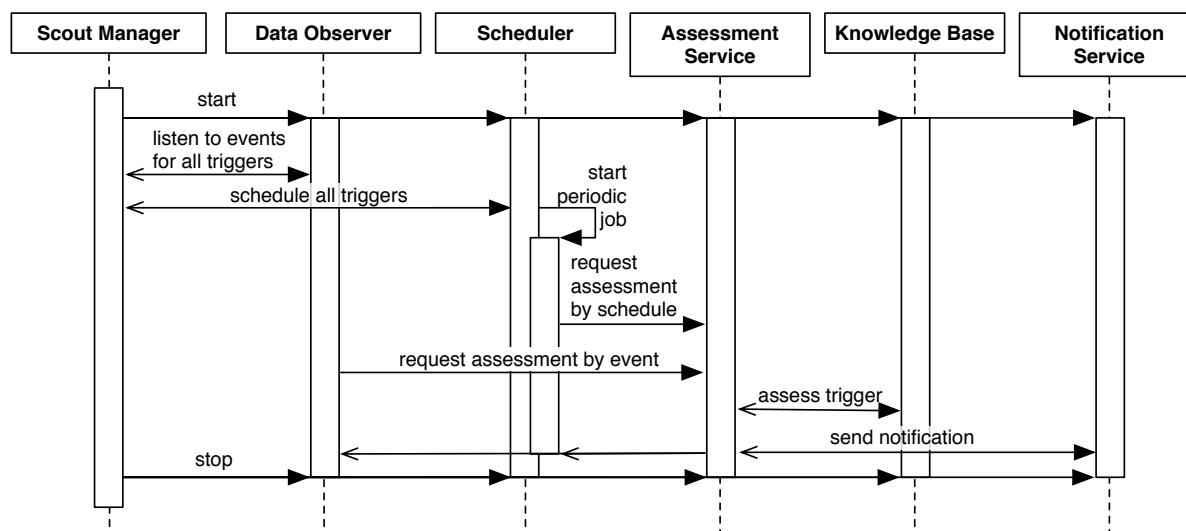


Figure 6.11: UML sequence diagram detailing how components interact to assess triggers and identify threats

## 6.3 Implementation

This section puts together state of the art tools with developments within the SCAPE project to fill the gaps found towards a preservation life-cycle that includes preservation watch by an implementation of Scout.

### 6.3.1 Monitoring content

As described in section [4.2](#) digital preservation starts by understanding what content a repository holds and what are the specific characteristics of that content. To do so, characterisation tools are executed on each file of the content. Characterisation entails the following steps ([Abrams et al., 2009](#)):

#### Identification

Process of determining the presumptive format of a digital object on the basis of suggestive extrinsic hints (for example, an HTTP Content-type header) and intrinsic signatures, both internal (a magic number) and external (a file extension).

#### Feature extraction

Process of reporting the intrinsic properties of a digital object significant to preservation planning and action. These features can function in many contexts as a surrogate for the object itself for purposes of evaluation and decision making.

#### Validation

Process of determining a digital object's level of conformance to the requirements of its presumptive format. These requirements are expressed by the normative syntactic and semantic rules of that format's authoritative specification.

### Assessment

Process of determining the level of acceptability of a digital object for a specific use on the basis of locally defined policies. Assessments can be used to select appropriate processing actions. In a repository ingest workflow, for example, the range of possible actions could include rejection, normalisation, or acceptance in original form.

There are many tools to perform file format identification, like the file command<sup>10</sup>, DROID<sup>11</sup>, FIDO<sup>12</sup> and Siegfried<sup>13</sup>, to name a few (Knijff and Wilson, 2011). Although generic file format identification tools, such as the file command, identify files with a non-controlled format name and optionally a MIME type<sup>14</sup>, this is sometimes considered not enough for an adequate identification of a file format. An adequate identification would use a unique identifier for a file format on a specific version. The commonly agreed best source for such unique identifiers are the file format registries, where the most well recognised one is PRONOM. For this reason, all digital preservation specialised file format identification tools, like DROID, FIDO and Siegfried, identify a format also by its PRONOM identifier.

After identifying the file format, it is possible to start feature extraction and validation, which processes are usually specific of each file format or family of formats (e.g. images). There are many tools and libraries specialised in specific formats, like the ExifTool<sup>15</sup> and ImageMagick<sup>16</sup> for images, MediaInfo<sup>17</sup> for video and audio, Apache POI<sup>18</sup> for Microsoft documents, and many others. Some tools wrap these smaller specialised tools and libraries together, making a more generic tool that can identify formats and also extract features and validate, when such is possible, for a bigger set of file formats. These are tools like Apache Tika<sup>19</sup>, JHOVE2<sup>20</sup> and FITS<sup>21</sup>. Although the list of supported file formats is far from being complete, these tools help immensely to characterise heterogeneous content. Of these tools, Apache Tika is the only one to be used across domains whenever extraction of metadata (i.e. features) and full-text of documents is needed, specially used to support information retrieval features. JHOVE2 and FITS are more specialised for the digital preservation domain, they identify formats using PRONOM and also feature format validation, although not for all formats. FITS is the most complete as it also wraps JHOVE2 and Apache Tika inside, making it more powerful but at a performance cost that can be prohibitive.

Experiments at the SCAPE project<sup>22,23</sup> showed the huge performance and stability penalty of using wrapping tools like FITS, instead of optimised tools like Droid and Apache Tika, on large-scale and very heterogeneous content from a web archive. Furthermore, they showed how large-scale data processing platforms like Apache Hadoop<sup>24</sup> together

---

<sup>10</sup><http://darwinsys.com/file/>

<sup>11</sup><https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/>

<sup>12</sup><http://openpreservation.org/technology/products/fido/>

<sup>13</sup><http://www.itforarchivists.com/siegfried>

<sup>14</sup><http://www.iana.org/assignments/media-types/media-types.xhtml>

<sup>15</sup><http://www.sno.phy.queensu.ca/~phil/exiftool/>

<sup>16</sup><http://www.imagemagick.org/script/identify.php>

<sup>17</sup><https://mediaarea.net/en/MediaInfo>

<sup>18</sup><https://poi.apache.org/>

<sup>19</sup><https://tika.apache.org/>

<sup>20</sup><https://bitbucket.org/jhove2/main/wiki/Home>

<sup>21</sup><http://projects.iq.harvard.edu/fits>

<sup>22</sup><http://openpreservation.org/blog/2013/01/09/year-fits/>

<sup>23</sup><http://openpreservation.org/blog/2014/05/28/weekend-nanite/>

<sup>24</sup><http://hadoop.apache.org>

with optimised integrations like Nanite<sup>25</sup> for Droid and Apache Tika, can be a significant help to horizontally scale the characterisation process.

Nevertheless, at the end of the characterisation process we would have an unbearably large amount of characterisation outputs, one for each analysed file, and each of these outputs would be on an uncontrolled structure which is dependant of the specific tool or library used to process the specific file format. FITS helps by transforming the output of each of the tools it wraps, merging into a single output in a controlled structure<sup>26</sup>, but it still falls short on the normalisation of extracted features to a common vocabulary and on the effective conflict resolution between the features reported by different characterisation tools.

To solve the problem of normalisation of features extracted by different characterisation tools a new tool was developed in SCAPE: the Vocabularyzer<sup>27</sup> — a tool that enriches characterisation information that is expressed using XML. The tool performs this task by annotating the XML elements in the original data with entities from a given vocabulary. The intention of Vocabularyzer is to link general characterisation data with the *SCAPE digital preservation ontology* (Palmer et al., 2014b).

The *SCAPE digital preservation ontology*<sup>28</sup> defines a common vocabulary<sup>29</sup> for characterisation tools, quality-assurance tools and control policies, conceptually linking together policies defined by an institution to formalise their objectives, and metrics that can measure how well the objectives are being achieved. The vocabulary includes the definition of more than 400 measures<sup>30</sup> that can be taken by characterisation tools and mapped through the Vocabularyzer. This would allow the direct evaluation of the compliance of each file to the policies, achieving the last step of characterisation: the assessment. But decisions should not be made on an individual file basis when large-scale content is concerned. A wider view of content characteristics is needed to make informed assessments and decisions based on their global impact and cost. For this objective the C3PO tool was developed.

Figure 6.12 depicts the process followed by C3PO: It aggregates characterisation outputs allowing human users real-time analysis of the content characteristics statistics. It also provides to watch and planning processes a summary of the most important content characteristic aggregates in the form of a content profile, which serves as a surrogate of the collection itself for digital preservation threat assessment. Finally, it provides planning automatic sampling features based on several algorithms that aim to produce representative sampling for automatic testing of preservation action alternatives. (Becker et al., 2009)

C3PO processes the output of characterisation tools like FITS, Apache Tika or Nanite, collecting and storing the result into an internal NoSQL database able to perform sharded cluster calculations and with map-reduce support. It has an extensible architecture able to easily add support for new characterisation tools. It runs analytical queries in real time to provide statistical information on each of the extracted features, including: histogram for feature properties with text data type values; sum, average, variance, standard deviation, maximum and minimum values for feature properties with numeric data type values; and yearly histogram for properties with date data type values (see listing 6.2 for examples). A configurable set of such properties, chosen by their relevance for digital preservation threat detection, is a content profile. (Petrov and Becker, 2012)

---

<sup>25</sup><http://nanite.openpreservation.org/>

<sup>26</sup>FITS output XML schema <http://projects.iq.harvard.edu/fits/fits-xml>

<sup>27</sup><https://github.com/openpreserve/vocabularyzer>

<sup>28</sup><https://github.com/openpreserve/policies>

<sup>29</sup><http://ifs.tuwien.ac.at/dp/vocabulary/quality#>

<sup>30</sup><http://purl.org/DP/quality/measures#>

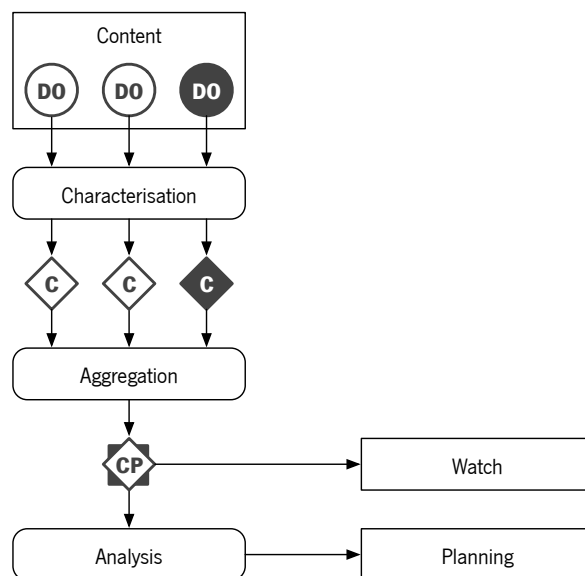
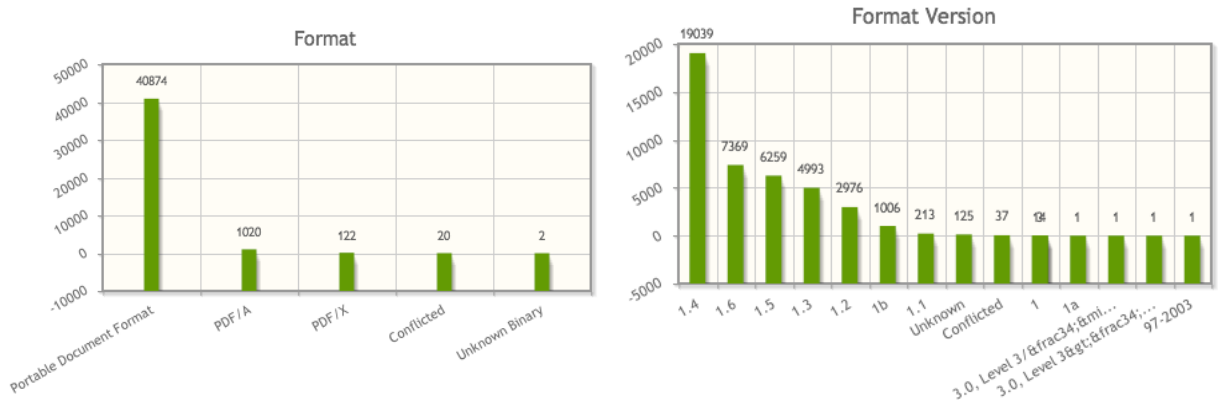


Figure 6.12: Content profiling

C3PO also provides a web interface that allows interactive analytics including presenting diagrams of the distribution of any of the processed file feature properties and also filtering and drill-down based on selected feature properties to segregate content into manageable partitions. Figure 6.13 shows several examples of property distribution diagrams from the repository and web archive domains. Figures 6.13a and 6.13b show the format name and version distribution on a repository with more than 40 thousand files. Although it is a quite controlled collection with mainly files in PDF<sup>31</sup>, there is a high heterogeneity of PDF versions. Furthermore, it may also be necessary to know more specific properties of the files for proper threat assessment and decision-making, like for example information on the creating application or validity of file towards file format specification, as depicted in figures 6.13c and 6.13d. Other collections, specially when ingest format policies are non-existent or not enforced, show a much higher heterogeneity in file formats, as indicated by the MIME type distribution of a web archive depicted in figure 6.13e. The user can create additional diagrams for any feature property present in the set in order to visualise its key aspects. 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. (Becker et al., 2014)

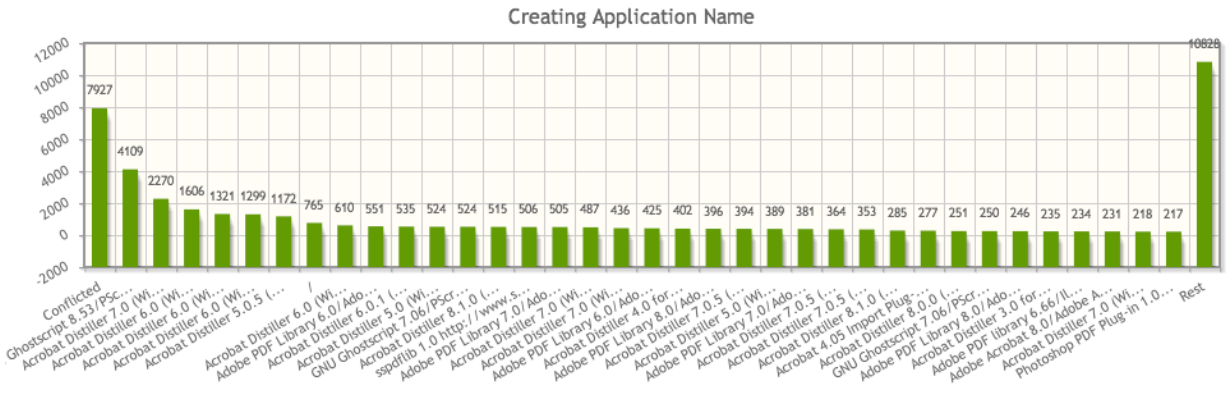
Another aspect that can be noticed on figure 6.13 is the high amount of *Conflicted* results on many of the diagrams. Conflicting results occur when different tools provide different values for the same property. The conflicting values can be caused by differences on the terminology used by each tool to encode the value (e.g. "Microsoft® Word 2013" vs. "Microsoft Word 2013"), differences on the specificity of tools with one giving more refined values than the other (e.g. "application/xml" vs. "application/xhtml+xml"), or simply by one or more tools providing wrong values. More reasons for conflict can be found at (Dappert, 2010). This lack of quorum undermines the data quality of the provided content profiles. Experiments at (Kulmukhametov and Becker, 2014) show that, on average, 68% of characterisation

<sup>31</sup>PDF means Portable Document Format

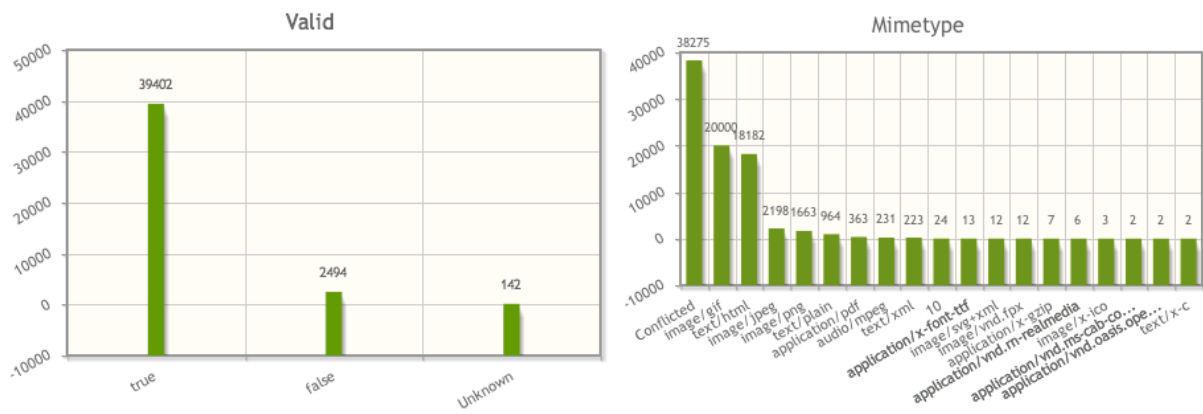


(a) Format distribution on a repository

(b) Format version distribution on a repository



(c) Creating application name distribution on a repository



(d) Format validity distribution on a repository

(e) MIME type distribution on a web archive

Figure 6.13: C3PO web interface example characteristic distribution diagrams

outputs have at least one conflict. The article also proposes a rule-based approach that was able to resolve 43% of these conflicts, greatly improving data quality. Such approach, which support was already added to C3PO, can be enhanced by adding more rules to the system, further reducing the number of conflicts.

Whilst C3PO may fit many common large-scale requirements, more extreme experiments (Faria et al., 2014) show that, on a single-server deployment<sup>32</sup>, it could process content at linear performance at around 216 files/second, but the web interface stops being responsive for collections of more than 2.5 million objects. But to allow the creation of collection profile reports on such large-scale collections an alternative approach was developed called DirectProfile. This approach doesn't use the NoSQL backend to store the data and allow the real-time statistics, but directly calculates the aggregates while processing the input, resulting in content profile XML reports which can be directly monitored by Scout. This approach executes with linear performance at a fraction of the time required by the normal mode and was able to process a half a billion object collection in less than 50 hours, around 3000 files/second and more than 10 times speedup, on a desktop computer<sup>33</sup>. A thorough description of this experiment is presented in section 7.3.

In outline, C3PO provides Scout with content profile XML reports dynamically generated by its REST API or statically generated by the DirectProfile mode. Each of the XML reports defines the state of the content on a moment in time (i.e. a snapshot), as in the example listings 6.1 and 6.2. Scout processes it and allows to inspect how the content profile evolves through time, using it as a surrogate of the collection itself for the detection of preservation threats. For example, the content profile is used to compare against the preservation control policies, which can be imported into Scout, to allow repository-wide continuous assessment of content. This is only possible due to the previous alignment of feature properties and the SCAPE digital preservation ontology, which directly maps to the control policies, and also due to all the work on conflict reduction and scalable execution.

A **pull source adaptor** for C3PO was developed for Scout to fetch information on content profiles. Firstly, the **C3PO source adaptor** fetches the list of available collections in a C3PO instance (listing 6.1). Each of the entries is a snapshot of a current collection or a snapshot of a collection in a previous moment in time. A collection in a previous moment in time would have the collection name suffixed with the date it belongs to (e.g. "websiteX\_2008-05-01"). A normal current collection would simply have the collection name (e.g. "demo"). Such a scheme was devised to allow the correct import of web archives backlog, on which each harvest of a site would potentially present a very different set of information<sup>34</sup>. A web archive might have already harvested the same website several times a year and an archive might already have several years of backlog. It is important for Scout to import the content profiles into the correct past date, showing how content has evolved throughout those years.

Each collection becomes an entity in Scout knowledge base and the content profile is processed to create properties of this entity. A collection profile contains a list of properties of different types: a histogram for text or date data types, and several statistical indicators for number data types. For example, the collection profile at listing 6.2 shows:

- A file format histogram in the `puid` property (PRONOM Unique Identifier), with 104 files of the `fmt/353` file format (TIFF: Tagged Image File Format), 23 files of the `fmt/99` file format (HTML: HyperText Markup Language, version 4), 7 files of the `fmt/18` file format (PDF: Portable Document Format, version 1.4) and many other values that were removed from the example for the sake of brevity.

---

<sup>32</sup>Test server had two CPU Intel Xeon X5670 (6 hyper-threaded cores, 12MB of cache, 2.93GHz, 6.40 GT/s Intel QPI), 288 GB for RAM, CentOS operative system

<sup>33</sup>Test desktop computer had one Intel CPU (4 cores), 8GB of RAM, Ubuntu operative system

<sup>34</sup>The output of the content profile is defined by C3PO, which is unaware that the collection profile belongs to a previous date. The name of the collection is the only configurable attribute that can be passed on to Scout to carry the date of the collection profile.



Listing 6.1: Example of a list of available content profile collections

```
<?xml version="1.0" encoding="utf-8"?>
<collections>
  <collection name="websiteX_2008-05-01"/>
  <collection name="websiteX_2009-12-01"/>
  <collection name="websiteX_2010-12-01"/>
  <collection name="demo"/>
</collections>
```

Listing 6.2: Example of a content profile XML report

```
<?xml version="1.0" encoding="UTF-8"?>
<profile collection="demo" date="Sat Sep 19 15:22:38 WEST 2015" count="206">
  <partition count="206">
    <filter id="9d62ee67-85a5-4878-9f5d-7c29857ed62e">
      <parameters>
        <parameter>
          <name>collection</name>
          <value>demo</value>
        </parameter>
      </parameters>
    </filter>
    <properties>
      <property id="puid" type="STRING" count="206">
        <!-- http://purl.org/DP/quality/measures#393 -->
        <item id="fmt/353" value="104"/> <!-- TIFF -->
        <item id="fmt/99" value="23"/> <!-- HTML 4 -->
        <item id="fmt/18" value="7"/> <!-- PDF 1.4 -->
        <!-- other values here were removed for the sake of brevity -->
      </property>
      <property id="size" type="INTEGER" count="206" sum="1.784601729E9" min="512" max="2.8598784E7"
        avg="8663115.18932039" var="6.4331934238502586E13" sd="8020719.059941109"/>
        <!-- http://purl.org/DP/quality/measures#409 -->
      <property id="created" type="DATE" count="49">
        <item id="2008" value="7"/>
        <item id="2003" value="6"/>
        <item id="2002" value="6"/>
        <!-- other values here were removed for the sake of brevity -->
      </property>
      <!-- other properties here were removed for the sake of brevity -->
    </properties>
  </partition>
</profile>
```

- Statistical indicators of the size property (file storage size in bytes), with a sum of 1.66 GB<sup>35</sup>, a minimum file size of 512 bytes, a maximum of 27.27 MB<sup>36</sup>, an average of 8.26 MB, and also information on the variance and standard deviation.
- The yearly distribution of the create property (date the file was created), with 7 files being created at 2008, 6 files at 2003, another 6 at 2002, and many other values that were removed from the example for the sake of brevity.
- Other properties were removed from the example for the sake of brevity.

The screenshot shows a web interface for an 'Entity' named 'demo'. Under the 'Properties' section, there is a table with three columns: 'Name', 'Value', and 'Action'. Each row represents a different property of the collection, with a small grid icon in the 'Action' column for each property.

Name	Value	Action
Collection size	1.66 GB	
compression_scheme distribution	5 key-value pairs	
Format distribution	15 key-value pairs	
Objects avg size	8.26 MB	
Objects min size	512 bytes	
Objects max size	27.27 MB	
Objects count	206	

Figure 6.14: Scout's content profile overview

The Scout's overview of content profile for the "demo" collection is presented in figure [6.14](#). The overview lists the last state of the properties it fetches from the content profile provided by C3PO content profile, presenting them in a human-readable way. The list of properties is configurable and can take all or a part of the properties available on the content profile. The properties presented on the figure define the collection size, objects minimum, maximum and average size, all based on the size property of the content profile; the format distribution, calculated by the same property on the content profile; and the compression scheme which is also available on the content profile although not presented on the listing example. By clicking on any of the properties a detail page for that specific property would

<sup>35</sup>GB means gigabytes

<sup>36</sup>MB means megabytes

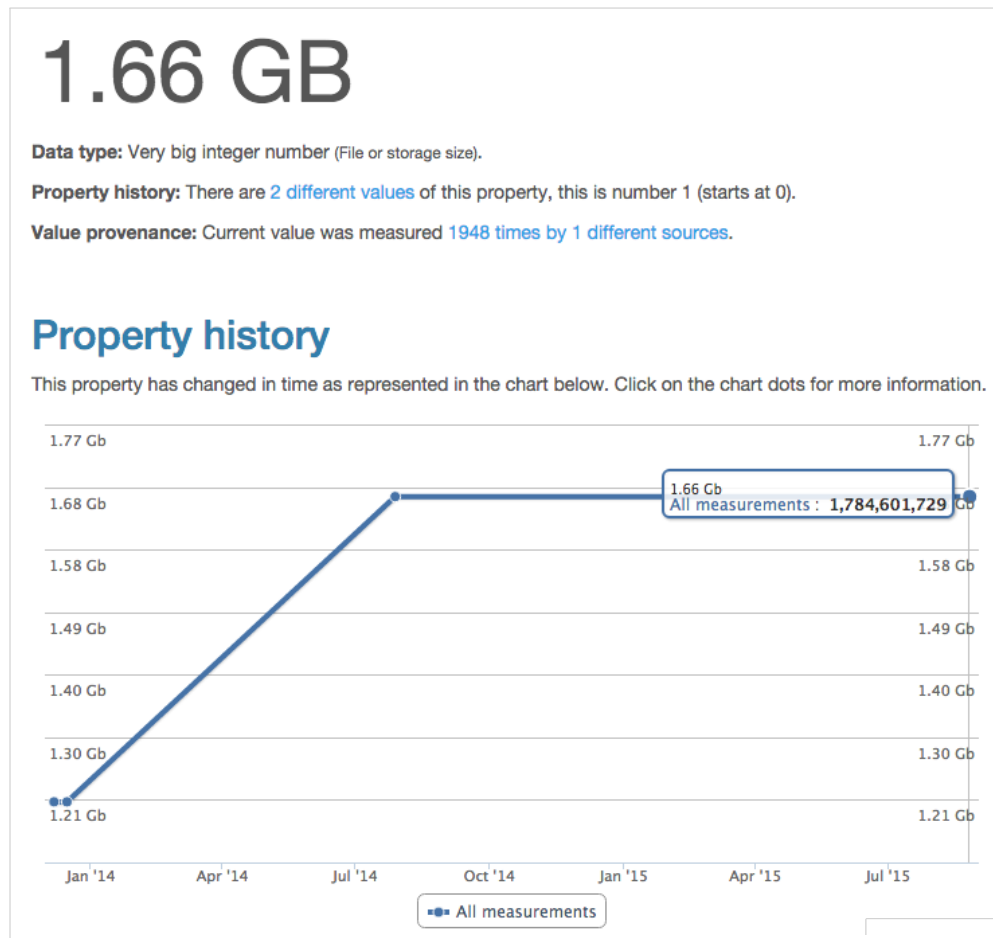


Figure 6.15: Scout's diagram of the storage size evolution through time

be presented, as in figure 6.15 for the size property, restating not just its current value (1.66 GB) but also its data type and rendering hint, which for the storage size property is "very big integer number" for the data type and "file or storage size" for the rendering hint. The figure also shows the complete property history, i.e. all previous values and the dates they were observed. A graph is presented to convey how the property changed throughout time, revealing trends. It can also be noted in figure 6.15 the value provenance, which is listed on the same page although not present in the figure. The value provenance lists the sources that have provided measures to this property, documenting the provenance of information, which is essential for trustworthiness and, consequently, for authenticity.

### 6.3.2 Monitoring the environment

One of the most important indicators for the outside environment, based on the survey results presented in section 5.4 are file format registries. As the current most well known file format registry is PRONOM, a Scout pull source adaptor plugin was developed to monitor the PRONOM SPARQL endpoint<sup>37,38</sup>. This linked data alternative access to PRONOM database is still a prototype but provides a much needed formalised version of the file format registry information encoded in a well defined vocabulary<sup>39</sup>.

The Scout PRONOM adaptor connects to the SPARQL endpoint and fetches all information to update its internal knowledge base. Each PRONOM file format is a new entity of the type "format" and all its properties are now internal properties of the created entity. For example, the "Broadcast WAVE" format will become a new entity under the entity type "format" and will have the following properties:

- **PRONOM Unique Identifier:** fmt/1
- **MIME type:** audio/x-wav
- **Version:** 0
- **File extensions:** wav
- **Format type:** <http://reference.data.gov.uk/technical-registry/formatType/Audio>
- **Byte order:** [http://reference.data.gov.uk/technical-registry/Little\\_endian](http://reference.data.gov.uk/technical-registry/Little_endian)
- **Released:** Wednesday, 1 January 1997
- **Withdrawn:** Sunday, 1 July 2001
- **Internal signature:** <http://reference.data.gov.uk/technical-registry/internalSignature/159>

Please note that the links above to the format type, byte order and internal signature, although they uniquely identify those property values as a RDF resource in the vocabulary, they do not point to a valid web resource.

The adaptor is able to fetch all information of the current 843 formats currently available from the SPARQL endpoint and integrate it with the knowledge base, where it can be used to enrich the information from the content profiles. A screenshot of Scout browsing the list of formats is available at figure 6.16.

<sup>37</sup><http://test.linkeddatapronom.nationalarchives.gov.uk/doc/file-formats>

<sup>38</sup><http://test.linkeddatapronom.nationalarchives.gov.uk/sparql/endpoint.php>

<sup>39</sup><http://test.linkeddatapronom.nationalarchives.gov.uk/vocabulary/pronom-vocabulary.htm>

## Category

Name	format
Description	Represents a file format

### Entities

← Previous 1-20 of 843 Next →















Name	Action
Broadcast WAVE[audio/x-wav; version=0]	
Broadcast WAVE[audio/x-wav; version=1]	
Graphics Interchange Format[image/gif; version=1987a]	
Graphics Interchange Format[image/gif; version=1989a]	
Audio/Video Interleaved Format[video/x-msvideo]	
Waveform Audio[audio/x-wav]	
Tagged Image File Format[version=3]	
Tagged Image File Format[version=4]	
Tagged Image File Format[version=5]	
Tagged Image File Format[version=6]	
Portable Network Graphics[image/png; version=1.0]	
Portable Network Graphics[image/png; version=1.1]	
Portable Network Graphics[image/png; version=1.2]	
Acrobat PDF 1.0 - Portable Document Format[application/pdf; version=1.0]	

Figure 6.16: Browsing the list of all formats provided by PRONOM adaptor in Scout

## Administration

Private page to manage the configurations of scout.

### Source Adaptors

All configured source adaptors that fetch information from external sources using plug-ins.

Instance	Plug-in name	Plug-in version	Source	Active	Actions
im-webarchive	c3po	0.0.9	<a href="#">IM</a>	✓	
im-browsershots	c3po	0.0.9	<a href="#">IM</a>	✓	
sb-webarchive	c3po	0.0.9	<a href="#">SB</a>	✓	
keeps-roda-demo-content	c3po	0.0.9	<a href="#">keeps-roda-demo</a>	✓	
keeps-roda-demo-events	Report API Adaptor	0.0.2	<a href="#">keeps-roda-demo</a>	✓	
pronom	Pronom Adaptor	0.0.6	<a href="#">pronom</a>	✓	

[New source adaptor](#)

### Installed plug-ins

All installed plug-ins. To install a new plugin just drop the plug-in jar into the correct folder under `/usr/local/scout/plugins`.

Name	Version	Description	Type
Pronom Adaptor	0.0.6	A Scout adaptor for the PRONOM registry.	ADAPTOR
Report API Adaptor	0.0.2	A Scout adaptor for a repository Report API.	ADAPTOR
c3po	0.0.9	A scout adaptor for the c3po content profiler source	ADAPTOR
HtmlEmail	0.0.3	Send notification via email	NOTIFICATION

Figure 6.17: Scout administration page with list of available source adaptor plugins and configured instances.

### 6.3.3 Setting up information source adaptors

Scout administration panel, shown in figure [6.17](#), lists all installed plugins and can configure new source adaptor plugin instances, enabling the user to add new information sources to the system.

Plugins, as described in section [6.2.4](#), can be installed by dropping the JAR with the plugin implementation on a predefined system folder and will automatically be loaded and appear on the administration panel. Figure [6.17](#) shows several installed plugins: PRONOM adaptor, Report API adaptor, C3PO adaptor, and HTML email notification plugin.

A new instance of an adaptor plugin can be created by clicking the "New source adaptor" button. Next pages will present a form with configuration parameters defined by the plugin, and also to select the source of information. If none of the available documented sources is adequate a new one can be created on the same panel. After configuring the adaptor plugin and connecting it to the source documentation, the instance can be created and will appear on the list available source adaptor instances.

After creation the new adaptor instance will be automatically initiated and scheduled to run periodically. Every time the adaptor instance runs a new event is registered, the list of events is presented on the adaptor detail page. If an adaptor instance run fails it will be retried up to a maximum of 5 consecutive times, after which it will be disabled.

Each run of an adaptor is documented by a source adaptor event, which includes a message describing the event, the success or failure of the run, the reason for failure if that is the case, and the date and time of the event.

### 6.3.4 Adding institutional control policies

In Scout a user can add institutional control policies, defined following SCAPE preservation watch ontology, directly on the user dashboard (see screenshot at figure [6.18](#)). The control policies can be defined in RDF and uploaded to Scout by clicking the "Upload policies" button. They are integrated as *is* into the knowledge base<sup>40</sup> and can be used in trigger questions. Control policies are defined as a set of objectives of different types:

#### **Access Objective**

Expresses a requirement for content to be provided in a representation that can be accessed by the user community. Involves information about technology available to users (e.g. web browser), formats that the objects are provided in and the capabilities of that technology.

#### **Authenticity Objective**

Requirement to maintain across preservation operations the intrinsic properties of an object (i.e. significant properties) that allow it to serve as an evidence of what it purports to record.

#### **Format Objective**

Expresses a constraint on the characteristics of a format.

#### **Action Objective**

Expresses a constraint on the characteristics of a software component.

#### **Representation Instance Objective**

Expresses a requirement on the characteristics of a particular representation.

---

<sup>40</sup>Integrated as *is* means that there are no transformations made to the input to conform to the knowledge base structure presented above.

My policies					
Objective	Measure	Description	Modality	Qualifier	Value
0	Running costs per object	Running operational costs of an action in € per object.	MUST	LT	0.24
1	elapsed time per MB	elapsed processing time per Megabyte of input data, measured in milliseconds	MUST	LT	2000
2	stability judgement	Judgement of the stability of an action	SHOULD		stable
3	ease of integration	Assessment of how easy it is to integrate an action into a particular server environment.	SHOULD		good
4	software licence source code	Indicates if and in which way the source code of the software is accessible.	MUST		openSource
5	ease of use	Assessment of how easy it is to use an action in operations	SHOULD		openSource
6	image width equal	true iff image width has been preserved.	MUST		true
7	image height equal	true iff image height has been preserved.	MUST		true
8	colour model preserved	Indicates whether the colour model has been preserved.	MUST		true
9	EXIF: all tiff data retained	Indicates whether the specific type of EXIF metadata has been retained.	MUST		true
10	number of tools	Indicator for the adoption of a format, expressed as the total number of all tools that support a format	MUST	GT	0
11	format documentation availability	Availability of the documentation for a format	SHOULD		yes-free
12	format standardization	Standardization of the outcome format.	SHOULD		international standard
13	compression type	Type of compression used in the outcome object	MUST		none
14	identification possibilities	Possibilities to identify the outcome format (cf. PRONOM)	MUST		automatic_specific
15	format complexity	Indicator of the complexity of a format as judged by experts in the domain	SHOULD		low
16	quantitative archival storage costs	Effects of a preservation action on archival storage costs, expressed as archival storage costs per object over a time period of 10 years, measured in €.	MUST	LT	0.24
17	automated QA supported	Indicates whether results of an action support automated QA with current means	MUST		true
18	comparative file size	Factor for relative output file size, calculated as: (size of output file / size of input file)	MUST	LT	10

[+ Upload policies](#)

Figure 6.18: List of control policies from Scout user dashboard



Each control policy objective is defined by:

- A **label** to easily identify the objective, e.g. "Compression scheme must be none";
- The **type** of objective, from the list above, e.g. "Format Objective";
- The related **measure** from the measures controlled vocabulary, e.g. measure #117 which has the description "Type of compression used in the outcome object";<sup>41</sup>
- The **modality**, based on the MoSCoW method, e.g. must, should, could, wont; (Clegg and Barker, 1994)
- A **qualifier**, an algebraic comparison operator that defines the rule of acceptance of a measure towards a value, e.g. equal, less than, greater than;
- A **value**, the target towards which the measure is compared via the qualifier, e.g. "none" for the type of compression.

Putting all examples together, the control policy presented above which is described as "Compression scheme must be none" enforces this rule by defining that the *measure* that relates to compression type (measure #117) *must* be *equal* to "none". It is now simple to create a trigger with a question that cross references this control policy with the related measure on the content profile, using the compression type property histogram to find out how much content has a compression type different from "none".

### 6.3.5 Detecting and monitoring threats

To detect a preservation threat, the knowledge base must have information that allows to infer the probability that a threat has happened or is about to happen. If there is no such information in the knowledge base, then it must be brought to it from external sources of information using the push or pull source adaptors. Experiments in chapter [7](#) demonstrate how relevant information can be captured from the outside world, specifically experiment in section [7.4](#) shows how any kind of information existing on the Web can be extracted into Scout.

Assuming that there is enough information in the knowledge base to detect the threat, continuously updated by source adaptors, then the next step is to make a query that would detect if the content is currently afflicted by the threat. As many threats may be common to a set of users, a question template can be created to address the common threat. The query page in Scout shows the list of these common questions, allowing the user to personalise the question using parameters. These parameters can be numeric values (e.g. for thresholds), text values, true/false options, or entity or property value selector (e.g. to select a content profile or a format). The provided form allows the easy input of the personalised parameters and the automatic creation of a question that should find the significant events. See figure [6.19](#) for a screenshot of the query page.

If none of the question templates applies to the threat, the advanced query can still be used to detect it. It enables the full use of the ontological knowledge base by allowing to create questions directly in SPARQL. It allows the selection of the request target, which defines the types of entities it will be searching on: entity types (also defined as "categories"), entities, properties, property values and measurements. It also permits the direct input of the SPARQL query, but

<sup>41</sup><http://purl.org/DP/quality/measures#117>

## Query

Select a pre-made question template or go to [advanced query](#).

**Query templates**

- [Check collection policy conformance](#)
- [Collection size limit](#)

**Check collection policy conformance**

Check if selected collection conforms to the defined policy (only compression scheme policy is checked right now)

**Collection**

sbweb

Your collection profile already inserted into scout

🔍 Search
+ Create trigger

← Previous
1-20 of 70
Next →

Category	Entity	Property	Value	Action
<a href="#">content_profile</a>	<a href="#">sbweb</a>	<a href="#">compression_scheme distribution</a>	28 key-value pairs	☰

Figure 6.19: Scout simple querying using predefined question templates

## Advanced query

Use SPARQL to make your own query

**Target**

- Category
- Property
- Entity
- Value
- Measurement

**SPARQL** [Help](#)

```
SELECT ?s WHERE { ?s rdf:type watch:PropertyValue .
?s ?p ?o }
```

🔍 Search
+ Create trigger

← Previous
1-20 of 112938
Next →

Category	Entity	Property	Value	Action
<a href="#">content_profile</a>	<a href="#">imcabinetoffice</a>	<a href="#">Collection size</a>	2.84 Gb	☰
<a href="#">content_profile</a>	<a href="#">imcabinetoffice</a>	<a href="#">compression_scheme distribution</a>	5 key-value pairs	☰
<a href="#">content_profile</a>	<a href="#">imcabinetoffice</a>	<a href="#">Format distribution</a>	22 key-value pairs	☰

Figure 6.20: Scout advanced querying using SPARQL

Figure 6.21: Create a trigger based on a question template (simple query)

limiting the binding variable to be of the defined request target type. See figure [6.20](#) for a screenshot of the advanced query page.

Clicking on the "Search" button the question can be tested, listing the current significant events. These significant events are the evidences that the preservation threat is putting content at risk, proving that preservation actions must be taken.

To continuously monitor the question, and therefore the threat, a trigger can be created with this question by clicking the button "Create trigger". This will take the user to the create trigger page, pre-filling the question, and allowing the user to define the periodicity of the question re-assessment, and the email to which a notification should be sent if new significant events are found. See figure [6.21](#) for a screenshot of the create trigger page.

As threats are detected, the relevant parties will receive an email and decide if a preservation action is indeed needed to mitigate the threat. The action is then executed and the threat re-assessed to confirm that the problem has been solved.

### 6.3.6 Planning and operations

When a preservation threat is detected and its existence verified, the planning process ensues. The planning process can be done manually or with the support of planning tools like Plato, but it would normally include the following steps: [\(Becker et al., 2009\)](#)

1. **Define objectives:** define the requirements the action must achieve and variables that must weight on the decision, e.g. open-source format, free and stable tool, good performance.

2. **Look for alternatives:** find alternatives that fit the requirements, e.g. TIFF vs. JPEG2000, ImageMagick<sup>42</sup> vs. GraphicsMagick<sup>43</sup>;
3. **Do experiments:** select a representative data set (e.g. using sampling) and experiment all alternatives, extracting measures that relate with the requirements;
4. **Evaluate results:** evaluate the experiment results towards the requirements calculating a score;
5. **Choose alternative:** select the best alternative, based on the score and other considerations, document the reasoning; e.g. selected TIFF because is more well supported, selected ImageMagick because is more widely used and faster;
6. **Document and deploy action plan:** create a document that registers the process evidences, generate an action plan that includes quality assurance, and provide all to the repository operations process.

The final result is an action plan that documents the decision making process but also contains a workflow to execute which would include the selected action alternative composed with one or more quality assurance tools that will evaluate if the action performs as expected and the outcome has enough quality to ensure that the objectives were achieved. These objectives should stem from the preservation policies and formalise the requirements set forth by the core high level objectives, like continuous access and authenticity, and would enforce the maintenance of the digital object significant properties.

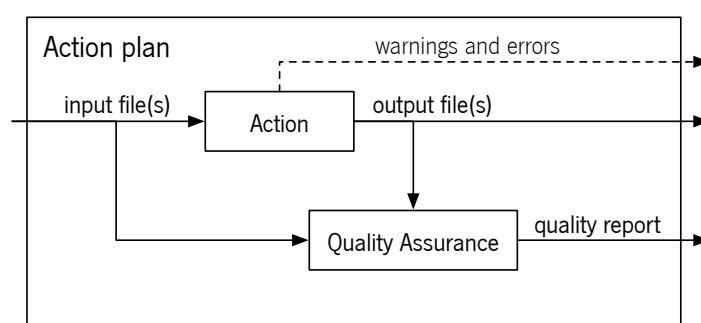


Figure 6.22: Simplified action plan workflow

Figure 6.22 shows a simplified diagram of what a preservation action plan workflow includes. Appendix A.3 shows a complete workflow diagram created with the Taverna workflow management system<sup>44</sup>, generated using Plato and SCAPE preservation components<sup>45</sup>. The action plan receives the input file(s) from the digital object and passes them on to the action tool, e.g. ImageMagick convert tool, which results in output file(s) with the result of the action plus possible warning and errors that are handed on to the output of the action plan. The input file(s) together with the

<sup>42</sup><http://www.imagemagick.org>

<sup>43</sup><http://www.graphicsmagick.org>

<sup>44</sup><http://www.taverna.org.uk/>

<sup>45</sup>The SCAPE preservation components are a set of characterisation, action and quality assurance tools developed or integrated by SCAPE to be fully documented, easily installable, easily found (using MyExperiment) and easily integrable with the SCAPE preservation suite. Ferreira et al., 2013 | Silva, 2014 | Palmer et al., 2014b | Pehlivan et al., 2014

output file(s) that resulted from the action are then sent into the quality assurance tool that will verify if the action was done with enough quality to be accepted, i.e. whereas the objectives for this action, previously defined on the plan, were achieved.

The planning process and the action plan that results from it can be applied to a single-file within each representation or to a list of files that compose part or the whole of the representation, depending on the scenario. Figure 6.23 illustrates several scenarios of how action plans can be devised and applied to a partial set of the content. These scenarios influence the quality assurance processes and their reports, which consequently impact the watch processes that monitors this information.

Scenario at figure 6.23a is the simplest one, where each digital object will have a representation composed of a single file. In this scenario the action plan can identify all files that need to be acted upon based on their file format, the action will receive a single file and produce a single file, both can be compared by the quality assurance process, and the result is easily transformed into a new representation with the action output file.

In scenario at figure 6.23b each representation can have multiple files that are independent of each other, i.e. they do not maintain any internal references nor depend on each other in order to be rendered. In this case, the action plan can still be made at the file granularity, as in the previous scenario, but when executing action plans care must be had to cope with several preservation plans acting on the same representation, and to copy or link files that are not target of any preservation action plan, in order to keep track of all files that compose the normalised representation.

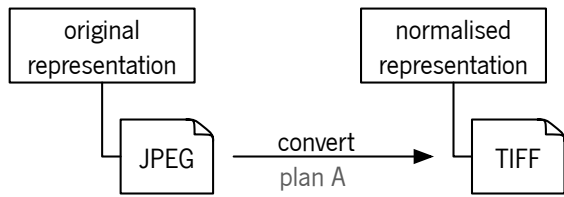
Figures 6.23c and 6.23d show two options of dealing with representations with multiple files that are dependent of each other. In this example, the HTML file references JPEG and GIF files, and needs these files in order to be fully rendered. When the objectives would force the images to be converted to another format, in this example TIFF, the file name extension may need to change, which would then imply a fix of the HTML references so they remain valid. This problem might also be overcome by renaming all files to a file format agnostic internal name, but this may not be recommended or sufficient in some cases<sup>46</sup>. Assuming this scenario would require file names to be changed and, consequently, fixing references of the HTML file, then there are two options how action plans and repository preservation operations can deal with it. One option, depicted in figure 6.23c is for the action plan to be made in the file granularity, as in the first scenario, passing to the repository preservation operations the responsibility to reconstruct the representation and ensure that internal references are correctly updated in the HTML file. Another option, in figure 6.23d would be for the action plan to take all files of the representation, including the HTML file, and make all necessary changes to produce a complete normalised representation, including updating references if necessary.

The need for an action plan to cope with a whole representation may be unavoidable, as when a representation composed of multiple files, dependent of each other, is converted to a single file. An example scenario would be a digitised book with an original digital representation made up of multiple images, one for each page, and a METS file that keeps the image order, all to be converted into a single PDF file. As depicted in figure 6.23e the action plan would need to receive all files and produce the single PDF file output, which would make up the normalised representation.

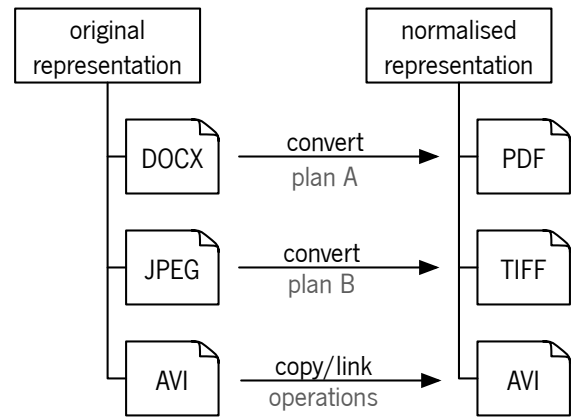
Likewise, the quality assurance steps of the action plan can be done on the file granularity and/or the representation granularity. Quality assurance that compares two files, the input and output of the action, can be done in two ways: 1) extract characteristics from the input and output file and compare them (e.g. compare image width and height); 2) apply a similarity distance function that calculates how much the two files are alike (e.g. image structure similarity -

---

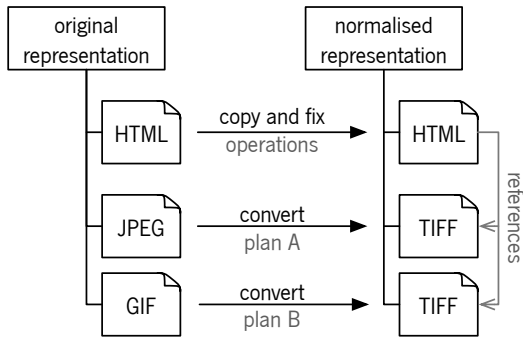
<sup>46</sup>It would not be recommended to rename files when their format can only be correctly detected with the help of the file name extension. Also, it may not be sufficient to rename files in case more information is kept than just the file name, e.g. METS may keep the checksum values of referred files.



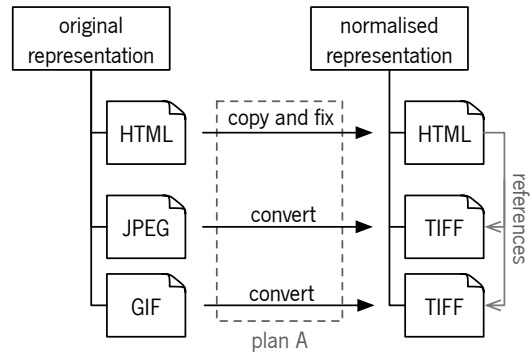
(a) Single-file representation scenario



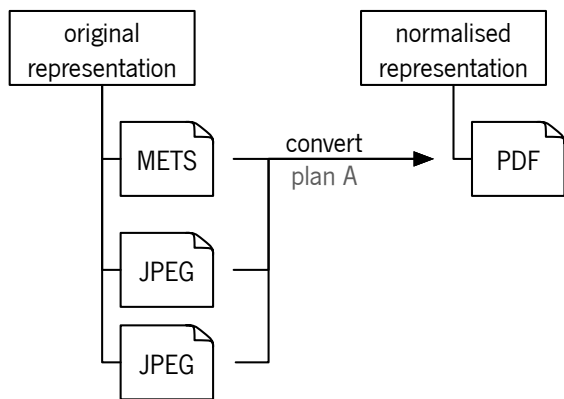
(b) Independent multiple file representation scenario



(c) Dependent multiple file representation scenario (option 1)



(d) Dependent multiple file representation scenario (option 2)



(e) Multiple file representation into a single file scenario

Figure 6.23: Representation structure scenarios for file format migration action plans

SSIM). Quality assurance to compare two representations would work in a similar fashion, but would need tools able to render the multiple-file representation. In some instances, both representation and file quality assurance could be done, as in the HTML plus images example, where each image can be compared on its own, and also the whole rendered web page can be compared, e.g. by creating a web page browser screenshot image and using the same quality assurance tools as in the individual images to ascertain if the representation as a whole still looks similar.

As a side note, it should be assured that the action and quality assurance processes do not share tools or any of their software sub-components, to avoid bias that may result on false-negatives. Moreover, other performance outputs from the action plan execution, like the execution time or used resources, can be important to monitor, specially when that was one of the reasons to select an alternative over another.

The action plan is then sent to the repository where, with the help of the operations process, it will be executed in all content that applies. This could be done manually or with the support of tools like Taverna and the SCAPE Platform (Schmidt, 2012, Schmidt and Rella, 2014), and would normally include the following steps:

1. Ensure all required resources are available: software, hardware, staff and others;
2. Identify the digital objects that the action will target;
3. Execute the workflow for each file, ensuring the correct reconstruction of the representation;
4. Document the action in preservation metadata for each digital object, connecting it to the preservation action plan that generated it, and registering all quality assurance outputs.

The resulting documentation of the preservation action would be encoded as a PREMIS event, which is referred by the PREMIS object that documents the normalised representation, and points to the PREMIS object of the original representation. Listing 6.3 shows an example of such a PREMIS event, considering a scenario alike figure 6.23d where a representation of HTML and two images is normalised as a whole representation by an action plan. The PREMIS event documents the date and type of the event, the outcome, relates it with the PREMIS agent that documents the tool used to perform the action, and points to the original representation used by the tool to produce the normalised representation. The listing also presents the action plan report which includes a reference to the plan, a quality report and a placeholder where warnings and errors can be presented.

PREMIS events don't have a direct support for relating preservation events with action plans nor to provide quality assurance reports of the executed actions, but they do have extension points, being the most appropriate one the `<eventOutcomeDetailExtension>` XML element. An XML schema was developed to encode the outputs of an action plan, extending the PREMIS event to support this information. Listing 6.3 provides an example, based on both representation and file granularity quality assurance. On the example, the representation files are laid out as defined by table 6.1

The example contains a quality report with a representation-wide quality measure, defined with label "rendered object screenshot distance using SSIM", calculated by rendering the whole representation (HTML file plus images) in a web browser and taking a screenshot, both of the original and normalised representations, comparing both screenshot images using the image distance SSIM algorithm. Also, for both image files, identified as "header" and "logo", three quality measures were taken for each and presented under a `<file>` XML element: the image distance using SSIM, comparison of image width characteristic and comparison of image height characteristic. Each measure also refers the URL for the SCAPE preservation ontology, whenever a suitable vocabulary term exists.

Listing 6.3: Example of preservation metadata produced due to the preservation action execution

```

<?xml version="1.0" encoding="UTF-8"?>
<event xmlns="info:lc/xmlns/premis-v2">
  <eventIdentifier>
    <eventIdentifierType>PID</eventIdentifierType>
    <eventIdentifierValue>event:1</eventIdentifierValue>
  </eventIdentifier>
  <eventType>migration</eventType>
  <eventDateTime>2014-09-04T14:10:51.70Z</eventDateTime>
  <eventDetail>The representation "normalised" derived from representation "original"</eventDetail>
  <eventOutcomeInformation>
    <eventOutcome>success</eventOutcome>
    <eventOutcomeDetail>
      <eventOutcomeDetailNote>detected properties</eventOutcomeDetailNote>
      <eventOutcomeDetailExtension>
        <actionPlanReport plan="planA">
          <qualityReport>
            <measure label="rendered object screenshot image distance using SSIM">0.996512</measure>
            <file id="header">
              <measure label="image distance SSIM" url="http://purl.org/DP/quality/measures#1">0.999605</measure>
              <measure label="image width equal" url="http://purl.org/DP/quality/measures#51">>true</measure>
              <measure label="image height equal" url="http://purl.org/DP/quality/measures#53">>true</measure>
            </file>
            <file id="logo">
              <measure label="image distance SSIM" url="http://purl.org/DP/quality/measures#1">0.999426</measure>
              <measure label="image width equal" url="http://purl.org/DP/quality/measures#51">>true</measure>
              <measure label="image height equal" url="http://purl.org/DP/quality/measures#53">>true</measure>
            </file>
          </qualityReport>
          <warningsAndErrors/>
        </actionPlanReport>
      </eventOutcomeDetailExtension>
    </eventOutcomeDetail>
  </eventOutcomeInformation>
  <linkingAgentIdentifier>
    <linkingAgentIdentifierType>PID</linkingAgentIdentifierType>
    <linkingAgentIdentifierValue>agent:1</linkingAgentIdentifierValue>
    <linkingAgentRole>preservation task</linkingAgentRole>
  </linkingAgentIdentifier>
  <linkingObjectIdentifier>
    <linkingObjectIdentifierType>PID</linkingObjectIdentifierType>
    <linkingObjectIdentifierValue>original</linkingObjectIdentifierValue>
    <linkingObjectRole>target</linkingObjectRole>
  </linkingObjectIdentifier>
</event>

```

Table 6.1: Structure of representation on example PREMIS event

File identifier	Original file name	Normalised file name	Action performed
root	index.html	index.html	fixed references
header	header.jpeg	header.tiff	converted file format
logo	logo.gif	logo.tiff	converted file format



### 6.3.7 Monitoring preservation actions

The produced PREMIS events are an important source of information for Scout as it is critical to monitor the quality of the performed preservation actions. To monitor this and other repository events, like ingest and access, the Report API was developed. The Report API is one of the three APIs that enable the integration with the SCAPE preservation suite (see section 4.3) and has the objective of providing metadata about repository events of interest to digital preservation concerns. Currently, the Report API allows monitoring of the following events:

- Ingest started and finished, identifying the SIP and providing information on the outcome success and details;
- Access of descriptive metadata and representations (view and download), identifying the accessed digital object and representation and providing further details on who accessed it;
- Execution of a preservation action plan upon a digital object, identifying the plan and the digital object, and providing information on the outcome success and details.

A register of each event is encoded as a PREMIS event, as in the example provided in listing 6.3. The Report API also includes, for each event, information about who or what triggered or performed the event in the form of a PREMIS agent and optionally other advanced details on the event.

The information is transmitted using the OAI-PMH protocol<sup>47</sup>, a standard that allows the transfer of metadata between two entities: the repository and the harvester. The repository holds the information and provides it to the harvester, which in our case will be Scout via the Report API adaptor, as illustrated in figure 6.24. The OAI-PMH protocol supports multiple metadata formats which can be selected on request. The Report API defines two metadata formats that can tune the level of detail: the `premis-event-v2` that provides the minimum detail associated with the PREMIS event, and the `premis-full-v2` that provides all available detail including PREMIS agents, PREMIS objects, descriptive metadata and other information associated with the event. (Asseg et al., 2013)



Figure 6.24: Repository Report API implementation and Scout adaptor

The Scout Report API adaptor fetches all repository events and calculates aggregates to allow easy checking of preservation action quality and performance. For each measure a set of statistic aggregate indicators based on the data type is calculated, having minimum, maximum, average and standard deviation for number typed measures and distribution histogram for text typed measures. For example, it can provide maximum, minimum, average and standard deviation for the image distance using SSIM, and distribution histogram of images with equal width.

Other information can also be fetched from the Report API, such as statistics of the preservation action execution time, ingest time, access performance, distribution of formats accessed, etc. See examples on appendix A.4

Triggers can be created, manually or automatically in Plato, to notify when values fall below a certain parameter. For example, a trigger can be created to notify whenever the *image distance minimum* value goes below a certain

<sup>47</sup><http://www.openarchives.org/OAI/openarchivesprotocol.html>

threshold (e.g. 0.99) and whenever the *image distance average* goes below the expected from the experiments (e.g. 0.999). It also can be set to notify whenever the width or height of an image has changed, i.e. whenever the distribution of *image width equal* measure contains one or more "false" values. Also, the original trigger used to initiate the action plan can be used to assess if the threat has been mitigated.

## 6.4 Final remarks

This chapter presents a novel design for a preservation watch system, named Scout, that improves on current state of the art and is well based on digital preservation and risk management concepts. It shows the feasibility of the design with a prototype implementation that follows all steps of the digital preservation life-cycle and proves that all requirements set forth have been achieved:

1. **Enable to pose questions about entities and properties of interest**

Scout web interface allows human users to pose questions, using templates or SPARQL, on properties on interest, and continuously monitor them using triggers that will send an email notification when significant events are found. The same functionality is also available to software agents via a REST API.

2. **Collect information from different sources through adaptors**

Scout design allows source adaptor plugins to be developed so they collect information into the knowledge base, mapping and normalising information to the knowledge base specification. Several adaptors were implemented to monitor repository content and events and file format registries.

3. **Act as a central place for collecting relevant knowledge that could be used to preserve an object or a content set**

All information is centralised into a knowledge base that allows complex querying using SPARQL. Information from different sources can be cross-referenced, mingled and filtered so significant events, that indicate preservation threats, can be extracted.

4. **Notify interested agents when an important event occurs**

Triggers can be created to monitor the knowledge base for significant events and send a notification to interested agents. New notification methods can be added via the notification plugins, so new communication channels are easily supported.

5. **Integrate with the decision-making process**

Trigger notifications are sent to relevant parties, which would enact planning (i.e. the decision-making process). Also, the planning process can create triggers to monitor the quality and performance of the defined action plans to verify if objectives were achieved and no new threats were introduced.

6. **Act as an extensible platform**

The plugin architecture can easily add new adaptors and notification channels, allowing to extend the platform functionality effortlessly. Also, the REST API allows other services to quickly integrate with Scout.

The next chapter will further validate Scout with real-world experiments in pursuit of answering the final research question.

# Chapter 7

## Evaluation

On this chapter a series of experiments are described that put the Scout artefact to test and validate the second research question hypothesis: that Scout can automatically detect the most important and neglected threats by monitoring, collecting and formally representing information about the world on its knowledge base.

This series of experiments focuses on the most important and neglected digital preservation threats found on chapter 5 and defines realistic scenarios where these threats are of importance, proceeding then to the information gathering and formal representation in Scout which enable the successful automatic detection of the identified threats.

### 7.1 Experiment 1: SCAPE preservation suite

#### 7.1.1 Scenario

The SCAPE preservation suite is a complete implementation of the preservation lifecycle, presented in section 4.3 using the tools developed in SCAPE: Scout, C3PO, Plato and the SCAPE platform supported by Taverna; and integrating them with a repository via three APIs: the Data Connector API, Report API and Plan Management API. To demonstrate and prove the viability of the SCAPE preservation suite, a deployment of the complete set of tools was developed in a virtual machine. The software, together with a fictitious scenario, allowed to create an easily repeatable experiment that proves the feasibility and usability of the approach. The SCAPE preservation suite was presented to the European Union's SCAPE review committee, at the IPRES 2013 conference<sup>1</sup> in Lisbon and at the Digital Libraries 2014 conference<sup>2</sup> in London, where it won the best demonstration award. (Kraxner et al., 2013; Duretec et al., 2014)

Consider the following scenario: an organisation has a large collection of digitised content, in image format, where the original paper analogues no longer exist. There is no information about which file formats the content is represented on. In terms of objectives, the organisation is not required to provide a service of content continuous access to users, such as in a dark archive, but it does require content to be readable in the long term, surviving technological changes with minimum preservation action interaction. The organisation therefore wishes to define as one of the control policies that no compression should be used on the content file formats and on its storage medium. Reasons for this control policy

---

<sup>1</sup><http://www.ipres-conference.org/ipres13/>

<sup>2</sup><http://www.dl2014.org>

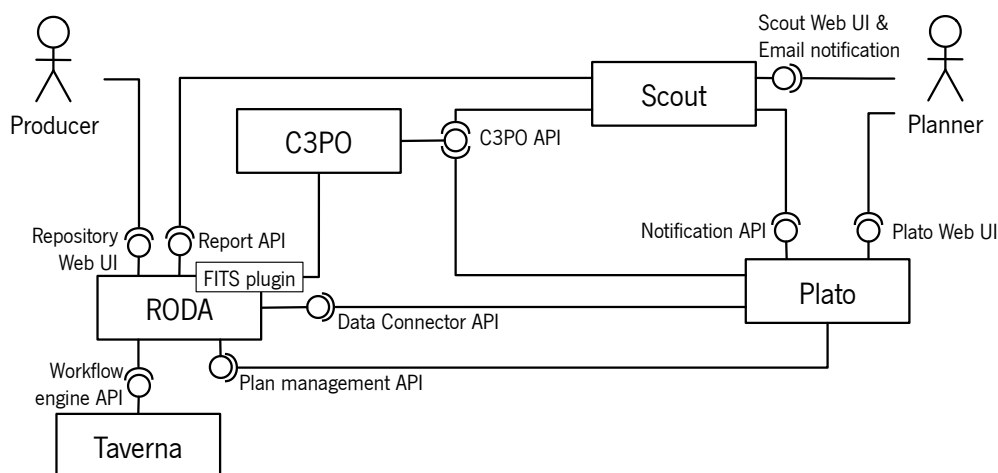


Figure 7.1: SCAPE preservation suite deployment

are twofold: compression increases entropy and obfuscates information making the effects of bit rot devastating; and compression is yet another technological layer that must be understood by future readers, one that can easily make information unintelligible if the compression algorithm is ever lost.

To enable digital preservation of content, the institution deploys the SCAPE preservation suite, prepared on a virtual machine with the arrangement depicted in figure 7.1. The **producer**, or archivist that would submit the digital content on the producers behalf, uses **RODA** to manage the content: a digital repository optimised for preservation<sup>3</sup>. It connects to the repository using its web user interface and is able to submit the content for ingest using the repository specialised methods (e.g. access to a shared storage space). As RODA repository allows integration of plugins that can be run periodically on content, a new plugin was developed to run **FITS** in all active content, i.e. in all files of all original representations or, alternatively, on the normalised representation version of content in case it exists. The output of the characterisation tool is processed by **C3PO**, by means of periodically monitoring a folder on a shared storage space, and the resulting content profile is available to **Scout** and **Plato** via the **C3PO API**. Scout also monitors all repository events via the **Report API** and contains information about the external environment via its source adaptors.

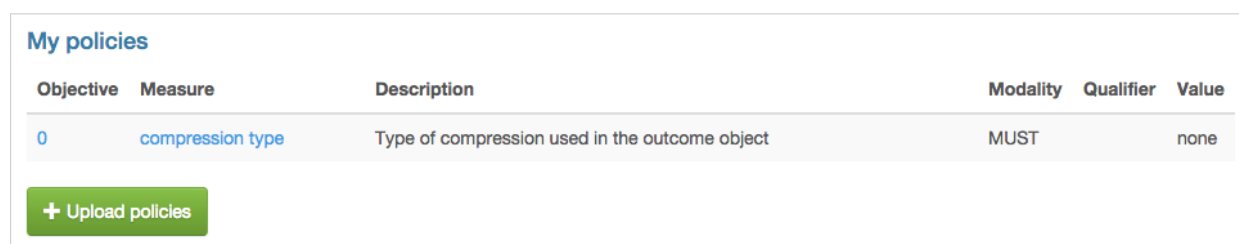
A user with the role of a **planner**, i.e. with the responsibility of managing the planning process, can use the **Scout web user interface** to configure policies and triggers that monitor the content and send notifications via **email** or via the **Notification API** when threats are detected. The planner can then use the Plato web user interface to devise a preservation action plan, following a process that uses policies, content profile and automatic sampling, detection of alternatives, automatic experimentation and comparison of results. The preservation plan can then use the **Plan management API** to deploy the plan, which will use the **Taverna** workflow engine to execute the plan on all content that applies via the **Workflow engine API**. The execution of the preservation action plan can update the repository via the **Data Connector API**, and every action is documented in the preservation metadata, which is then available to Scout via the Report API. Scout can therefore monitor the output of the preservation actions, by the captured events, and the impact they have on the content, as content profiles are updated.

<sup>3</sup><http://www.roda-community.org>

## 7.1.2 Execute experiment

The life-cycle starts by configuring Scout with the preservation policies and triggers that embody the objectives and measure their achievement. These preservation policies can change through time to fit new or updated objectives and to include the detection of new threats that the organisation would like to focus on.

The preservation policies are defined as an RDF ontology, based on *SCAPE digital preservation ontology*<sup>4</sup>. An example of a preservation policy is available on listing [7.1](#), where a fictitious organisation named "The Archive of Examples" has a "Example collection scenario" with an "example collection" content set and a "public" target user community. This preservation case has a single defined objective which mandates that compression type, identified as measure #117 in the controlled vocabulary, must be none.



Objective	Measure	Description	Modality	Qualifier	Value
0	compression type	Type of compression used in the outcome object	MUST		none

[+ Upload policies](#)

Figure 7.2: Policies in Scout

The policies can be easily uploaded to Scout using the green button on the web page partly depicted in figure [7.2](#) listing all objectives on the same page. Scout is already configured to monitor the content of the RODA repository, which is empty at this point. A trigger can now be created to notify the user, for example by email, when content does not conform to policies, using a pre-made question template (see figure [7.3](#)).

Submitting the content to RODA its ingest process unfolds, passing the several validation steps and ultimately being incorporated into the repository. As content becomes part the repository, the FITS plugin picks up the new content and outputs the deep characterisation analysis to C3PO, which aggregates the characteristics creating a content profile. At this point the user can inspect the characteristics of the content on the C3PO web user interface.

As Scout is monitoring C3PO, the content most relevant properties are incorporated into its knowledge base, together with information about the repository events, for example information about the ingest as the average ingest time. As content springs into Scout's gaze, triggers are re-evaluated detecting that some content uses compression and a notification is automatically sent to the defined user (see figure [7.4](#)).

Upon receiving the notification, the planner can verify in Scout which are the non-conformances, i.e. which properties of content are misaligned with the defined policies. Scout shows that the property compression scheme of the content profile has a number of files with compression different of none, specifically in JPEG. The planner can also use C3PO to drill-down into content and can optionally divide the problem into several parts, dividing the content into sets easier to manage and plan, or make a plan for the whole content set, depending on how heterogeneous content is.

Using Scout and C3PO, the planner can obtain a detailed view of the problem and verifies that all files are images in JPEG file format and use JPEG lossy compression. The planner then starts the preservation planning process to address the detected issue with the optimal operation available. Plato downloads sample records from the repository

<sup>4</sup><https://github.com/openpreserve/policies>

Listing 7.1: Example of control policy uploaded to Scout

```

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE rdf:RDF [
  <!ENTITY quality "http://purl.org/DP/quality#">
  <!ENTITY measures "http://purl.org/DP/quality/measures#">
  <!ENTITY preservation-case "http://purl.org/DP/preservation-case#">
  <!ENTITY control-policy "http://purl.org/DP/control-policy#">
  <!ENTITY modalities "http://purl.org/DP/control-policy/modalities#">
  <!ENTITY qualifiers "http://purl.org/DP/control-policy/qualifiers#">
  <!ENTITY owl "http://www.w3.org/2002/07/owl#">
  <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#">
]>
<rdf:RDF xmlns="http://archiveofexamples.org/policies#"
  xmlns:preservation-case="http://purl.org/DP/preservation-case#"
  xmlns:control-policy="http://purl.org/DP/control-policy#"
  xmlns:org="http://www.w3.org/ns/org#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"

  <owl:Class rdf:about="&preservation-case;ContentSet"/>
  <owl:Class rdf:about="&preservation-case;PreservationCase"/>
  <owl:NamedIndividual rdf:about="&quality;object_compression"/>

  <org:Organization rdf:about="archive_of_examples">
    <rdf:type rdf:resource="&owl;NamedIndividual"/>
    <org:identifier>The Archive of Examples</org:identifier>
  </org:Organization>
  <owl:NamedIndividual rdf:about="example_collection">
    <rdf:type rdf:resource="&preservation-case;ContentSet"/>
  </owl:NamedIndividual>
  <owl:NamedIndividual rdf:about="example_collection_scenario">
    <rdf:type rdf:resource="&preservation-case;PreservationCase"/>
    <skos:prefLabel>Example collection scenario</skos:prefLabel>
    <preservation-case:hasObjective rdf:resource="CompressionTypeMustBeNone"/>
    <preservation-case:hasUserCommunity rdf:resource="public"/>
    <preservation-case:hasContentSet rdf:resource="example_collection"/>
  </owl:NamedIndividual>
  <owl:NamedIndividual rdf:about="CompressionTypeMustBeNone">
    <rdf:type rdf:resource="&control-policy;FormatObjective"/>
    <skos:prefLabel>Compression type must be none</skos:prefLabel>
    <control-policy:value rdf:datatype="&xsd:string">none</control-policy:value>
    <control-policy:measure rdf:resource="&measures;117"/>
    <control-policy:modality rdf:resource="&modalities;MUST"/>
    <preservation-case:contentSetScope rdf:resource="example_collection"/>
  </owl:NamedIndividual>
  <foaf:Group rdf:about="public">
    <rdf:type rdf:resource="&owl;NamedIndividual"/>
  </foaf:Group>
  <owl:NamedIndividual rdf:about="&modalities;MUST"/>
</rdf:RDF>

```

## Create a new trigger

Receive a notification when your question detects a nonconformity.

---

**Name**

**Run every**

day  week  month  quarter  year

**Question**

Check collection policy conformance  
Check if selected collection conforms to the defined policy (only compression scheme policy is checked right now)

**Collection**

Your collection profile already inserted into scout

**Email**

[+ Create trigger](#)

Figure 7.3: Creation of a trigger in Scout which monitors content policy conformance

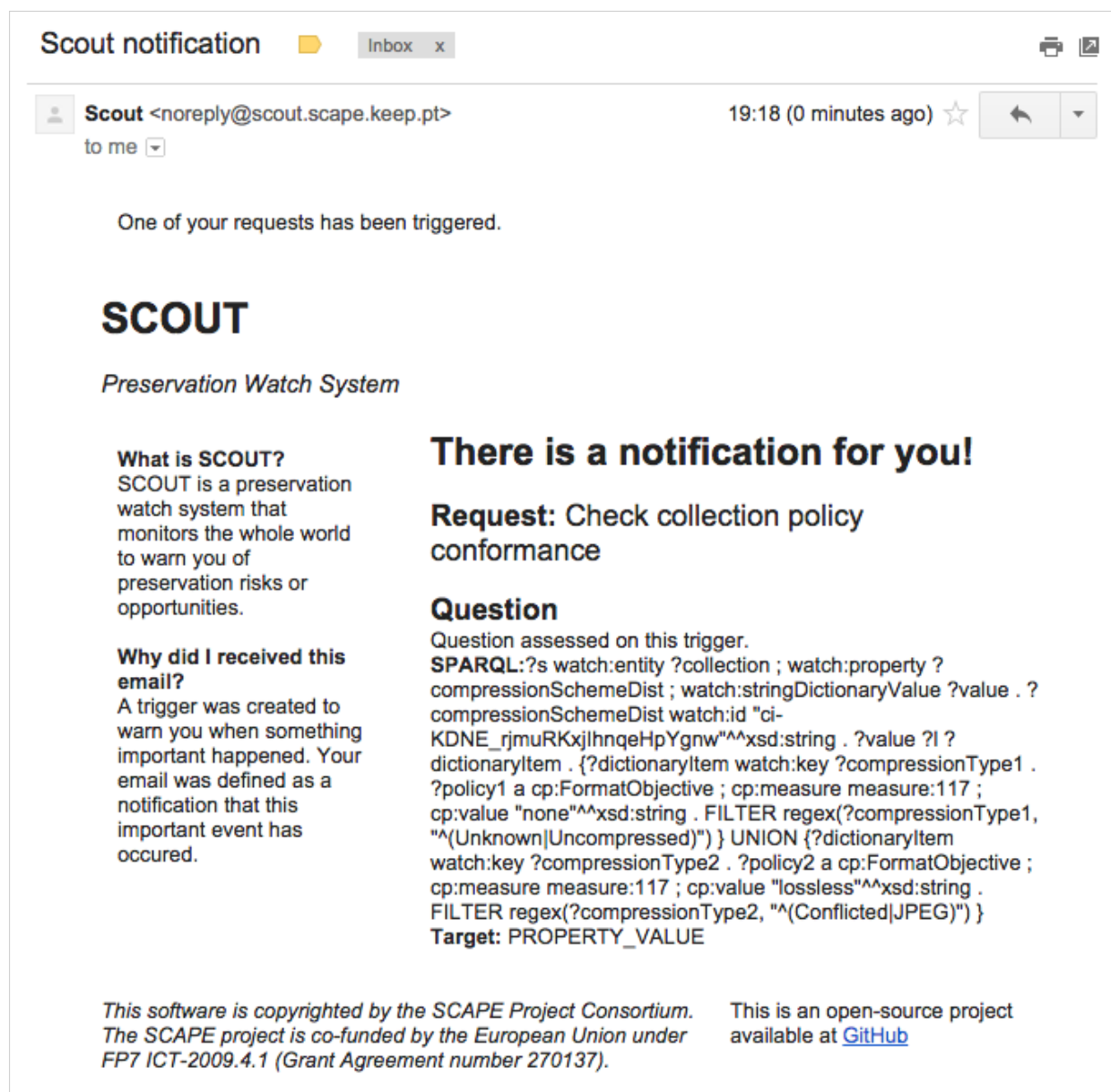


Figure 7.4: Notification sent by Scout warning planner of a content non conforming to policies threat



over the Data Connector API, understands control policies, and automatically extracts decision criteria from them. Based on the criteria and following the planning process, the planner selects alternatives and measures how well they fit the criteria and objectives, making experiments if needed. The planner selects uncompressed TIFF as the format to which images will be converted to, and pins down ImageMagick and GraphicsMagick as the software alternatives. Plato automatically selects samples using C3PO automatic sampling features, retrieves the samples using the Data Connector API, and creates an action plan workflow putting together the selected tools with quality assurance components. Then Plato requests the execution of the alternative action plan workflows using the Workflow Engine API, and finally retrieves the measurements that prove which one is the best alternative. At this point, a final decision is made by the planner on which action to employ on which part of the content, creating a document that contains a Taverna workflow which combines the selected action with quality assurance components (see figure in appendix [A.3](#)), identifies the content to which the action will be applied, and documents all steps of the decision making process, creating evidences that support the decision choice.

The newly created or revised preservation plan is then deployed to the repository via the Plan management API. As the repository has an execution environment which understands the preservation action plan workflows, the plan can be executed immediately without human intervention. The plan explicitly identifies the files to which it applies and provides a set of conditions that need to be satisfied. These conditions are simultaneously deployed to Scout, which monitors the execution of a plan and its conformance to the specified levels of quality, specifically if the action did not change the image width, height and resolution.

As content is updated by the preservation action plan, which adds a normalised representation where every image file was converted into uncompressed TIFF, the monitoring pipeline detects the changes and updates characterisation information, content profile and then Scout's knowledge base. The re-assessment of triggers now confirm that all content conforms to policies and also that file format migration did not affect significant properties as confirmed by quality assurance outputs reported by the Report API<sup>5</sup> and assessed by the triggers installed by Plato.

### 7.1.3 Results

This experiment proves that Scout, once configured, is able to automatically gather information from the repository content and events to detect preservation threats, namely that content does not conform with defined institutional policies. Furthermore, Scout confirms that the selected preservation action mitigates the threat and further monitors the execution of the action, which on and of itself introduces new threats, namely if the action executed on files is (not) having the expected results.

But this experiment is not constricted to these threats and the monitoring and detection of other relevant threats can be added to the scenario. Watch is a continuous process that requires the planner to progressively add triggers for new threats that become relevant, possibly requiring Scout to gather information from new sources so detection is possible. Below are some examples on how to monitor new threats in Scout.

#### Monitoring file corruption threat

To detect file corruption, a file fixity check RODA plugin was developed, which calculates the checksum of all files and compares it against the preservation metadata documented checksum (within the PREMIS object). Every check

---

<sup>5</sup>Examples of the quality assurance outputs reported by the Report API are available on appendix [A.4](#)

is documented as a PREMIS event and monitored by Scout via the Report API (example in listing [7.2](#)). Scout can therefore detect when a file has been corrupted and notify relevant parties.

Listing 7.2: Example of fixity check PREMIS event, result of the RODA plugin and available via the Report API

```
<event xmlns="info:lc/xmlns/premis-v2">
  <eventIdentifier>
    <eventIdentifierType>RODAObjectPID</eventIdentifierType>
    <eventIdentifierValue>roda:669314</eventIdentifierValue>
  </eventIdentifier>
  <eventType>fixity check</eventType>
  <eventDateTime>2015-09-22T13:19:48.10Z</eventDateTime>
  <eventDetail>Checksums recorded in PREMIS were compared with the files in the repository</eventDetail>
  <eventOutcomeInformation>
    <eventOutcome>success</eventOutcome>
    <eventOutcomeDetail>
      <eventOutcomeDetailNote>files checked</eventOutcomeDetailNote>
      <eventOutcomeDetailExtension>
        <p xmlns="http://www.w3.org/1999/xhtml">[FO]</p>
      </eventOutcomeDetailExtension>
    </eventOutcomeDetail>
  </eventOutcomeInformation>
  <linkingAgentIdentifier>
    <linkingAgentIdentifierType>RODAObjectPID</linkingAgentIdentifierType>
    <linkingAgentIdentifierValue>roda:189</linkingAgentIdentifierValue>
    <linkingAgentRole>preservation task</linkingAgentRole>
  </linkingAgentIdentifier>
  <linkingObjectIdentifier>
    <linkingObjectIdentifierType>RODAObjectPID</linkingObjectIdentifierType>
    <linkingObjectIdentifierValue>roda:664436</linkingObjectIdentifierValue>
    <linkingObjectRole>target</linkingObjectRole>
  </linkingObjectIdentifier>
</event>
```

## Monitoring consumer misalignment threat

Some control policies are set towards detection of threats about how well the formats align with the consumer's technological environment. Triggers that detect these threats need complementary information to be able to measure if objectives are being achieved. Listing [7.3](#) has some examples of such control policies, such as there should be at least one tool available for consumers to read the files on the provided format, the format should be an international standard and should have freely available documentation. Although PRONOM does not contain information on tools that read a file format, or which formats are standards and have free documentation, other registers exist with such information. For example, FileInfo.com contains information on programs that open a file format, grouping by operative system<sup>6</sup>, and UDFR contains some information on which tools input and output which file formats<sup>7</sup>.

This information can be used by Scout to create triggers that cross reference policies, content profiles and information on the file format registries to detect the threats of consumer misalignment.

<sup>6</sup><http://fileinfo.com/extension/tiff>

<sup>7</sup><http://udfr.org/ontowiki/view/r/u1f47>

Listing 7.3: Example of control policy for consumer misalignment

```

<owl:NamedIndividual rdf:about="NumberOfToolsMustBeGT0">
  <rdf:type rdf:resource="&control-policy;FormatObjective"/>
  <skos:prefLabel>Number of tools greater 0</skos:prefLabel>
  <control-policy:value rdf:datatype="&xsd;integer">0</control-policy:value>
  <control-policy:measure rdf:resource="&measures;141"/>
  <control-policy:modality rdf:resource="&modalities;MUST"/>
  <control-policy:qualifier rdf:resource="&qualifiers;GT"/>
  <preservation-case:contentSetScope rdf:resource="example_collection"/>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="FormatDocumentationAvailabilityShouldBeYesFree">
  <rdf:type rdf:resource="&control-policy;FormatObjective"/>
  <skos:prefLabel>Format Documentation freely available</skos:prefLabel>
  <control-policy:value rdf:datatype="&xsd:string">yes-free</control-policy:value>
  <control-policy:measure rdf:resource="&measures;147"/>
  <control-policy:modality rdf:resource="&modalities;SHOULD"/>
  <preservation-case:contentSetScope rdf:resource="example_collection"/>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="FormatShouldBeInternationalStandard">
  <rdf:type rdf:resource="&control-policy;FormatObjective"/>
  <skos:prefLabel>Format should be international standard</skos:prefLabel>
  <control-policy:value rdf:datatype="&xsd:string">international standard</control-policy:value>
  <control-policy:measure rdf:resource="&measures;161"/>
  <control-policy:modality rdf:resource="&modalities;SHOULD"/>
  <preservation-case:contentSetScope rdf:resource="example_collection"/>
</owl:NamedIndividual>

```

### Monitoring producer misalignment threat

In cases where the producers of content are not part of the organisation and the repository supports auto-deposit (i.e. the producer uploads the content to the repository), there is a threat that a relevant percentage of the producers cannot comply with the established ingest policies, or that the resources are not adequate to cope with the producers' content input.

In such cases, monitoring ingest process, as well as the content that is produced by it, can be a very relevant insight to enable producer trend analysis. Producer policies are normally enforced on the ingest process, before content is accepted to the repository's custody, and feedback about how well producers fit to this enforcement can be a valuable input. Such information is available to Scout via Report API, which allows access to the preservation events that document the ingest process. Triggers can be created to raise up an alert whenever the ingest success rate goes below a certain level, or when the ingest average processing time goes above an acceptable duration.

### Monitoring outdated preservation plans threat

Consider now that the organisation's policy changes to permit, and even enforce, lossless compression on images, for example by using LZW compression in TIFF images, to minimise required storage and drive costs down (preservation opportunity). Compression can greatly reduce the necessary storage space, and although it introduces new risks, they can be compensated with backup and by gathering and preserving the LZW compression algorithm documentation.

The previously installed trigger in Scout would automatically detect this change in the policy and alert that content is

again non-conforming to policies, notifying the planner that the plans need to be re-evaluated.

## 7.2 Experiment 2: Checking external references

### 7.2.1 Scenario

Nowadays information is ever more connected and although referencing in scholarly papers to other scientific articles is largely done by a well documented bibliography, references to other material like software, institutional websites, online tools, blogs and other web pages, is mainly done via web links, which are easily perishable. Although some link persistence can be achieved by persistent identifiers (Hilse and Kothe, 2006), the access to these depends on the institution that provides the content and although the link might perdure, the service that delivers the document still can disappear.

It is difficult to measure the impact of this loss of context information, but as in the case of this document, the loss of the links to the software may hinder the capability of a third party to reproduce the experiments and to take full advantage of the software and architecture produced in this work.

Table 7.1: List of the top twenty repositories from which the dataset was created.

Repository	Documents
ScholarsArchive at Oregon State University <sup>8</sup>	35.682
DSpace@MIT - Massachusetts Institute of Technology repository <sup>9</sup>	24.712
DSpace at Utrecht University <sup>10</sup>	21.529
DEA@DEENK - University of Debrecen Electronic Archive <sup>11</sup>	19.094
Kyoto University Research Information Repository <sup>12</sup>	17.456
Deep Blue at the University of Michigan <sup>13</sup>	14.641
UTpublications - Open Access repository of the University of Twente <sup>14</sup>	13.808
ScholarSpace at University of Hawaii at Manoa <sup>15</sup>	12.423
RUA - Institutional Repository of the University of Alicante <sup>16</sup>	12.247
Repository at Leiden University <sup>17</sup>	11.466
Veritati - Repositório Institucional da Universidade Católica Portuguesa <sup>18</sup>	9.456
Archive of European Integration <sup>19</sup>	8.915
RIA - Repositório Institucional da Universidade de Aveiro <sup>20</sup>	8.627
EconStor - Open access repository of the German National Library of Economics <sup>21</sup>	8.282
Archivo Digital UPM - Universidad Politécnica de Madrid <sup>22</sup>	7.937
DSpace en ESPOL - Escuela Superior Politecnica Del Litoral, Ecuador <sup>23</sup>	7.818
Organic Eprints <sup>24</sup>	7.769
NERC Open Research Archive - Natural Environment Research Council, UK <sup>25</sup>	7.749
Universidade do Minho Repositorium <sup>26</sup>	7.466
<b>Total</b>	<b>257.077 (59%)</b>

<sup>8</sup><http://ir.library.oregonstate.edu>

<sup>9</sup><http://dspace.mit.edu>

<sup>10</sup><http://dspace.library.uu.nl>

<sup>11</sup><http://dea.lib.unideb.hu>

<sup>12</sup><http://repository.kulib.kyoto-u.ac.jp>

In the perspective of a repository tasked with the preservation of the scientific production of a university, for example, the capacity to monitor how much of the context is being lost throughout the time is critical to manage the need for mitigation techniques, such as web archiving. This experiment brings forth the capability to do so, by using a large-scale dataset in the form of 354.850 documents in PDF format fetched from 114 open-access scientific repositories, a corpus of 437.252 documents of many formats<sup>27</sup>. A list of the top twenty of open-source repositories from which corpus was created is available on table [7.1](#).

The experiment focuses on detecting and extracting web links from the corpus collection and testing if they are still available. The date of creation will be extracted from the PDF with characterisation tools so trends can be analysed.

### 7.2.2 Execute experiment

The first step of the experiment was to extract metadata and full text from all 354.850 PDFs using Apache Tika<sup>28</sup>, which can extract the mime type (to test if indeed is a PDF), the year of creation or last modification of the PDF, and all textual content from the PDF. Only PDFs were chosen for this experiment to ensure the quality of extracted metadata and full text. Apache Tika batch mode was used to execute the program on all files in one go in a multi-threaded fault-tolerant batch process. The result of this process was 354.850 XHTML files with all metadata in <meta> tags on the head of the file and the full text on the body of the file.

At this point the first clean-up was done. First, 7.153 files were removed because they weren't an actual PDF, as identified by the extracted MIME type. These files were created when some of the accessed repositories actually needed authentication. Second, analysing the file distribution, there were not enough files from before the year 2000, and as the corpora was created in the middle of 2013, this year was not complete, so 13.398 files from these years were removed. The result of the clean-up were 322.151 PDF files that were produced between 2000 and 2012.

- 
- 13 <http://deepblue.lib.umich.edu>
  - 14 <http://doc.utwente.nl>
  - 15 <http://scholarspace.manoa.hawaii.edu>
  - 16 <http://rua.ua.es>
  - 17 <https://openaccess.leidenuniv.nl>
  - 18 <http://repositorio.ucp.pt>
  - 19 <http://aei.pitt.edu>
  - 20 <http://ria.ua.pt>
  - 21 <http://www.econstor.eu>
  - 22 <http://oa.upm.es>
  - 23 <http://www.dspace.espol.edu.ec>
  - 24 <http://orgprints.org>
  - 25 <http://nora.nerc.ac.uk>
  - 26 <http://repositorium.sdum.uminho.pt>

<sup>27</sup> This corpus was created by using the OAI-PMH protocol of these repositories to extract metadata and then scrapping the web page to find and download the documents.

<sup>28</sup> <http://tika.apache.org>

To extract the URLs (i.e. web links) from the full text, the following set of programs were executed for each file:

```

1 tail -n +3 | \
2 perl -p -e 's/^\s*\n/-/' | \
3 xurls | \
4 sort | uniq | \
5 grep -E -o "http(s)?\:\:\/\/[a-zA-Z0-9\.\-]+\(\.[a-zA-Z0-9\.\-]+\)(\.\.)*$"

```

1. Skip XHTML preamble due to URLs in namespace
2. Compensate for line breaks in URLs due to hyphenation
3. Apply URL extraction program xurls<sup>29</sup>
4. Remove duplicated URLs
5. Apply additional restrictive rules to avoid malformed URLs

The output of this program was redirected to a new file, one for each document, producing in total 1.113.286 URLs. Figure 7.5 shows the average number of web links in a file and how this number evolves from 2000 to 2012. The global average, considering files from all years, is 2.76 detected web links per file, where around 60% of files do not have any detected web links. There are many reasons for a web link in a file to not be detected, from the full text extraction from the PDF to fail to recover the text correctly, to malformed web links on the original document due to bad input or autocorrect, or to errors in the tools that extract URLs or in the script presented above.

The year trend on figure 7.5 shows that there is a linear growth of web links being used in scientific articles, spanning from an average of 0,65 web links per file in the year 2000 to an average of 6,29 in 2012, almost ten times larger. The year 2005 is an outlier with far less links than expected from the linear growth, the reason for it still needs to be investigated. The (blank) bar refers to files from which the used method was unable to extract the year of creation.

The availability of each URL was tested using curl<sup>30</sup>, an open source command line tool and library for transferring data with URL syntax. The program was set to cancel request after a 10 seconds timeout so it wouldn't block indefinitely. Such timeouts are treated further on the experiment. The program was also set only to request the HTTP headers, not downloading the content itself.

```

1 curl --max-time 10 -# --head --stderr - $URL

```

Due to time constraints, only 236,085 of URLs were tested, around 21% of the total. As files in corpora had random generated filenames, their selection was also random, so the sample was not biased. The resulting output was inserted into a MongoDB<sup>31</sup> database to enable analysis. At this point a second clean-up was done, to remove any web link checks that were not correct or conclusive, due to factors other than availability:

<sup>29</sup><https://github.com/mvdan/xurls>

<sup>30</sup><http://curl.haxx.se>

<sup>31</sup><http://www.mongodb.org>

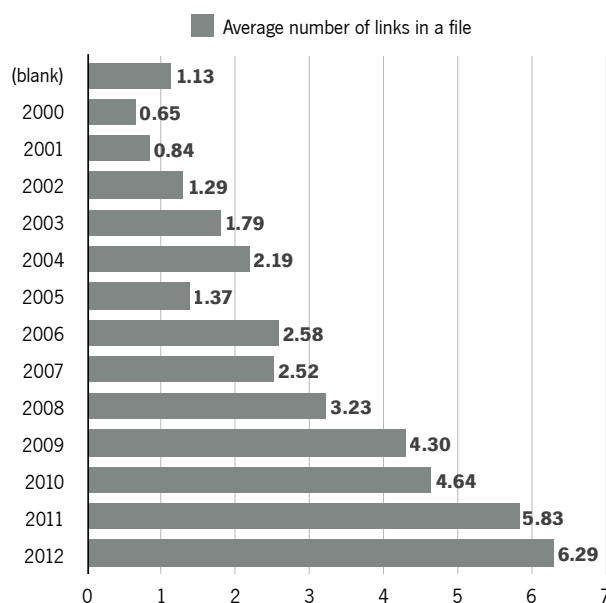


Figure 7.5: Yearly trend of the average number of web links in a file

1. Removed domain name lookup and connect timeouts as they could be due to network failures (8,773 checks);
2. Removed HTTP 405 Method not allowed responses as some servers disallow crawling robots (475 checks);
3. Removed web links with invalid top-level domain<sup>32</sup>, to ensure checks referred to valid web links (4,985 checks).

The clean-up removed 8% of the checks, ending up with 217,855 URLs checked. The result analysis is presented next.

### 7.2.3 Results

In total, 166,517 URLs (75%) were verified as available, whereas 55,335 (25%) could not be reached or the server reported that the content could not be found. Further analysing the year trend, as presented in figure 7.6, it can be noted that the global number of web links grows every year, as expected from the number of links per file trend revealed in figure 7.5. The 2012 year is an exception, mainly due to the smaller number files from that year, as it can be observed in figure 7.7a. This outlier might be due to the harvest of the corpora being done in the middle of 2013, and some documents from 2012 haven't had yet the time to find their way into the repository, be submitted, approved and published.

Although the number of dead links increases every year (with the exception of the outlier), the percentage of links that ceases to be available slowly reduces, from 31% in the year 2000, to 24% at 2012. It is expected for older links to be more vulnerable, although the yearly increase of perishability may depend on what the links point to, the care users have in selecting links less prone to disappear, the care institutions have in maintaining web resources available, and the use of persistent identifiers to maintain resources localisable.

<sup>32</sup><https://archive.icann.org/en/tlds/>

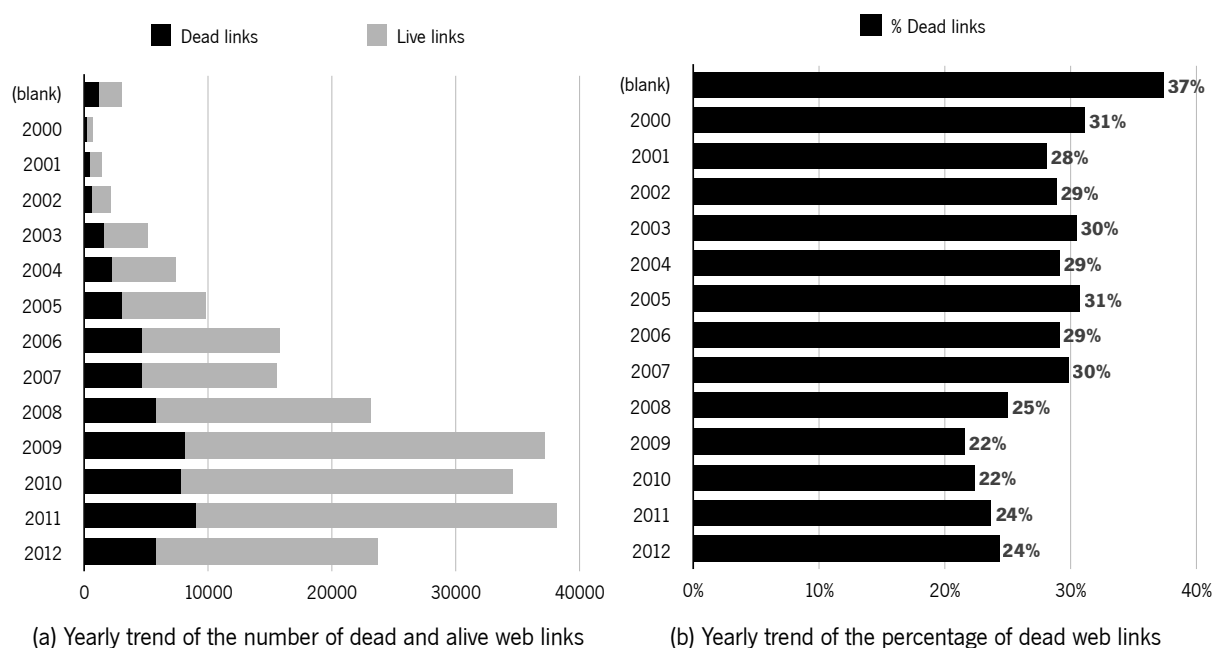


Figure 7.6: Yearly trend of web link status

Cross-referencing the web link checks with the files that contain the web links shows that around 19% of files have at least one dead link and therefore have lost context information. Graphs in figure 7.8 show that this number increases every year, in total number and percentage of files, due to the increase of web links per file trend. Graph at figure 7.7b shows that 28% of files created in 2012 already have at least one dead web link, and if the trend continues this number is expected to grow.

Losing just one web link in an scientific article or thesis might not be an huge impact, but figure 7.8 shows that the relative amount of dead web links in each file is at least 20%, with a tendency to grow as the file becomes older.

Scientific articles, thesis and all other documents available on these open-access repositories might have been submitted with web links that are already unavailable, which undermines their quality and, consequently, the quality of the repository that provides these documents. The cultural and scientific production and heritage might be starting to loose information even before the moment of capture, and continuously becomes more vulnerable as the context information is not controlled by the repositories and institutions that have the responsibility and objectives to preserve these assets. Mitigation techniques, like ingest control policy to verify the validity and availability of external web links or web archive of external references might be employed to reduce the impact of the web link perishability.



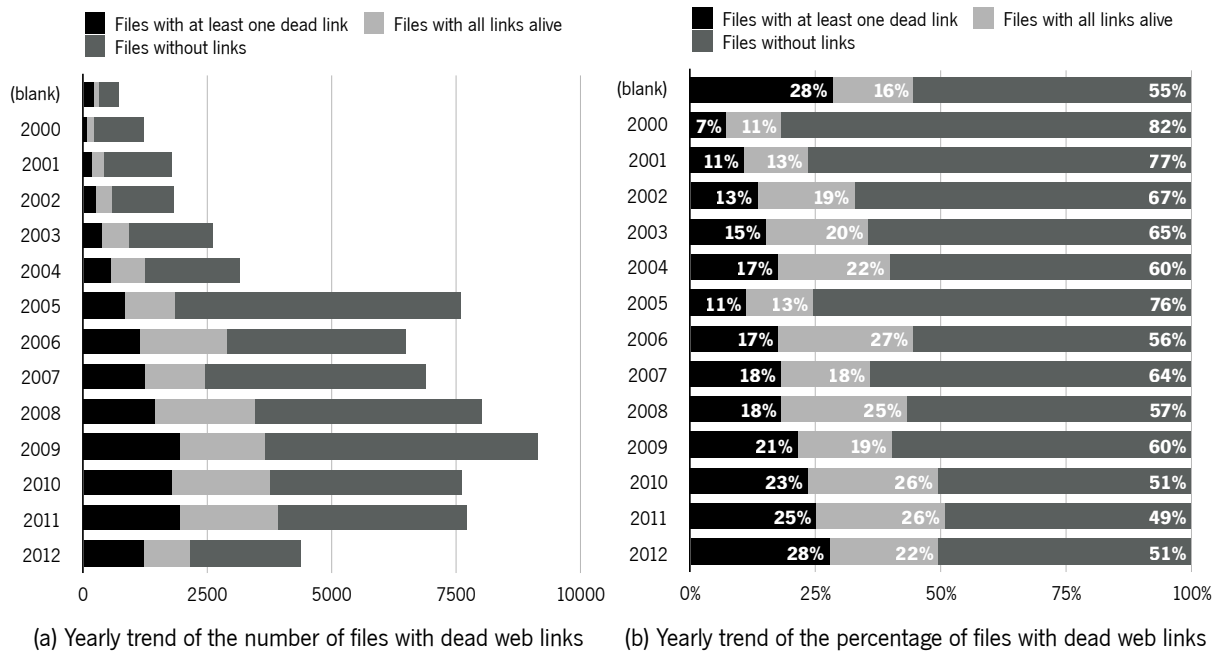


Figure 7.7: Yearly trend of files with web links

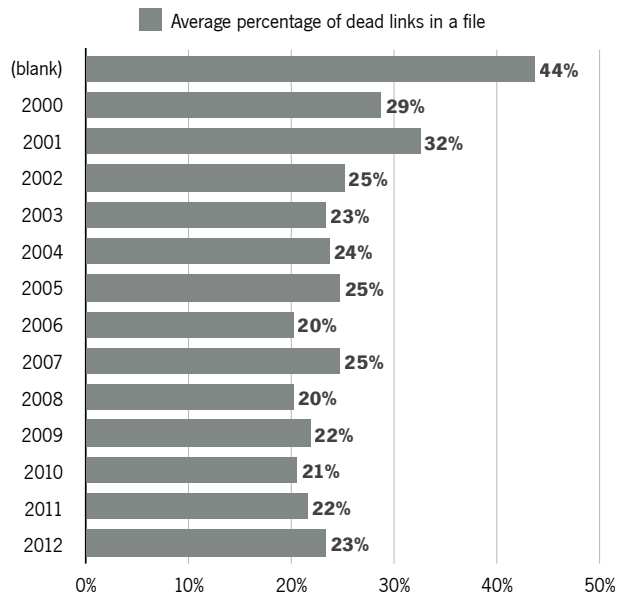


Figure 7.8: Yearly trend of the percentage of web links within each file

## 7.3 Experiment 3: Web archive deep characterisation

### 7.3.1 Scenario

Web archives preserve a very important part of the cultural heritage of the world and are, at the same time, a representative sample of the technological landscape, enabling the inference of trends in the general technological environment. Also, web archives are a good example of the large-scale issues common in many institutions.

A web archive dataset was selected to find limitations of the used tools (FITS, C3PO and Scout) for processing a large scale data in a real world situation. The dataset portrays a real world situation that occurs in large institutions with the mandate to archive big sets of the web, namely the Danish State and University Library. Extracting metadata from large scale data collections is a daunting task. Analysing it for anomalies, structure, patterns, format profiles, and other characteristics of preservation interest is no lesser task. In this experiment the File Information Tool Set (FITS) was employed on a large data set, a sample of the Danish State and University Library web archive named Netarkivet<sup>33</sup>, with 13.14 TB and half a billion objects. This web archive consists of web resources that have been harvested by crawling the Danish part of the Internet since 2005, i.e. from every publicly available URL on the Danish top level domain “.dk”. During the crawl process the web resources are collected in uncompressed ARC<sup>34</sup> containers storing approximately 100MB each that on average corresponds to 3600 web resources. A representative sample covering all the years for which data has been harvested was selected, table 7.2 lists the number of ARC files per year. The web archive had already passed the 300 TB of harvested web resources in 2013, so this sample corresponds to less than 5% of the total size of the archive at the time of the experiment. Due to Danish legislation this sample cannot be made available for the partners in the SCAPE project, but the Danish State and University Library can perform experiments on it for the SCAPE project, therefore most of this experiment was executed by this institution who was a SCAPE project partner<sup>35</sup>. The result was ingested into C3PO for profile extraction and analysis and then imported into Scout. (Faria et al., 2014)

### 7.3.2 Execute experiment

#### Deep characterization of content with FITS

At the moment this experiment was executed, in the early days of the SCAPE project, the platforms and tools for large-scale execution were not yet ready, so this experiment was implemented as a simple Bash<sup>36</sup> script based approach<sup>37</sup>. The characterisation process was executed on a group of machines described in table 7.3. The ARC files were stored on a SAN and the processing load was handled manually by giving each machine a manually defined subset of ARC files to work on.

The job was initiated in November 2011 and ran with few short interruptions until April 2013, resulting in 106 GB of

<sup>33</sup><http://netarkivet.dk>

<sup>34</sup><http://archive.org/web/researcher/ArcFileFormat.php>

<sup>35</sup>To note that this was many times the case: legal or institutional regulations impeded sharing of real world corpora due to intellectual property rights or personal information concerns, but the resulting analysis like the content profile has none of these problems and can be freely shared.

<sup>36</sup>Bash is an acronym for bourne-again shell, a command processor that can also read commands from a file, called a script.

<sup>37</sup><https://github.com/statsbiblioteket/SB-Fits-webarchive>

Table 7.2: Size of data sample

<b>Year of harvest</b>	<b>Number of ARC files</b>
2005	4,024
2006	20,497
2007	17,139
2008	30,685
2009	23,019
2010	14,090
2011	13,386
2012	17,897
<b>Sum</b>	<b>140,747</b>

Table 7.3: Hardware specification

<b>Machines</b>	5 blade servers
<b>CPU model</b>	Intel Xeon processor X5670 (12MB cache, 2.93GHz, 6.40 GT/s Intel® QPI)
<b>Processors</b>	2 processors per server, each with 6 hyper threaded cores, total 60 cores and 120 threads
<b>Memory</b>	288 GB of RAM in total
<b>Network</b>	Gigabit ethernet
<b>Storage</b>	SAN connected via the Gigabit Ethernet network
<b>Operating System</b>	CentOS

XML data distributed over 140,000 files each with an average number of 3641 records per ARC file (437,000,000 web resources total). A quantitative analysis of this data is presented in the blog post A Year of FITS<sup>38</sup>. The acquired data during the characterisation process was made available for ingest into C3PO through a simple HTTP interface. (Knijff and Wilson, 2011) presents an analysis of how FITS performs in general.

### Processing of FITS output with C3PO

The result of the previous step was a large number of FITS output files in XML. To enable the analysis of the deep characterization output, the C3PO tool was used to process the FITS output and enable real time analytics on the global result and transitive integration with Scout. But the C3PO tool had not been tested with such a large volume of data before. The data acquired on the above-presented process was ingested into C3PO to evaluate how it would perform on large-scale. At the time of this evaluation only data for the years 2005 until 2011 was available.

The three main functionalities of C3PO were tested: 1) Ingest; 2) Profiling; 3) Exploratory user interface. C3PO version 0.20 running against version 2.4 of MongoDB in Tomcat version 7 was used and all performance tests were performed on a server with the specification described on table 7.3.

The first test was to ingest the complete data set of circa 440,000,000 FITS output files. During the characterisation process described above, the FITS files were organised in TGZ files (i.e. compressed TAR<sup>39</sup> archives), one for each ARC file. The files were organised in circa 123,000 TGZ files with an average of 3,600 FITS data per TGZ. The ingest process first extracted the data from the TGZ before running the Java based C3PO ingest process. The ingest times for these extracted TGZ files are depicted in figure 7.9 where each line represents a TGZ sample which again represents an ARC file. Figure 7.9 shows that the ingest time is linear with the data size with only a few outliers. The results confirm that C3PO is able to ingest large data sets, at least up to the tens of TB in size.

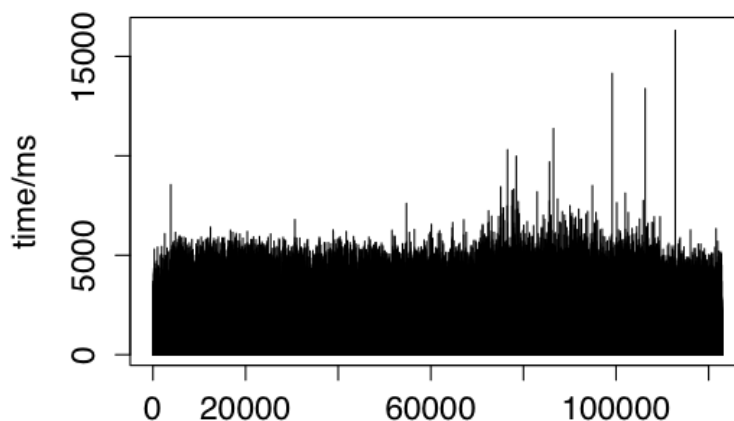


Figure 7.9: Distribution of ingest time for the complete set of TGZ files

The second test was to use C3PO command line interface for extracting profiles of the ingested data. This functionality, which can be used to create data for Scout, was tested on two samples: the 2005 data and the complete set. As

<sup>38</sup><http://www.openplanetsfoundation.org/blogs/2013-01-09-year-fits>

<sup>39</sup><https://www.gnu.org/software/tar/>

can be seen in table 7.2 there are circa 4,000 TGZ files for the year 2005. These TGZ files contain 11,905,931 FITS output files and a content profile was calculated in 15 hours and 18 minutes. The calculation for the complete set had to be interrupted because it was taking too long, but the process ran without errors for 60 hours. This shows that the C3PO ingest process could run for at least that long without crashing. If the implementation of the algorithm behaves linearly as expected, it would take about 22 days to complete. This experiment shows that the tested version of C3PO is not fit for handling complete collections.

The third test used C3PO graphical user interface for exploring extracted properties and correlations. In this test the objective is to explore how much data can actually be handled by the application and still be usable with all its features. This was done by importing a data set of a given size and then measuring the time needed to show the first screen with the overview plots in figure 7.10. Next step was to go into the first bar on the first plot by clicking it and measuring the time it took to show the second plot.

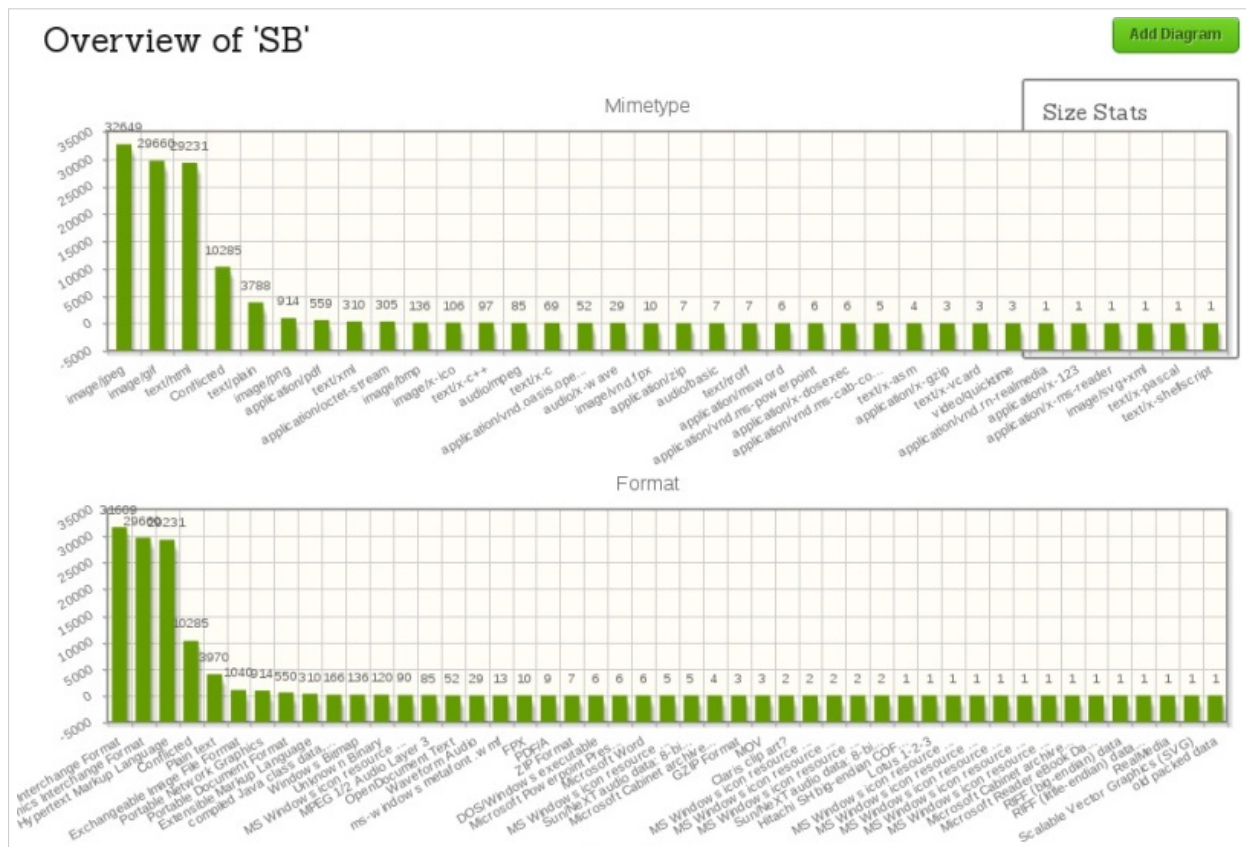


Figure 7.10: Collection overview page in C3PO

The results are shown in table 7.4 and as can be seen C3PO stopped being responsive somewhere close after 2.5 million FITS result files. Based on this performance test it is evident that C3PO at the time needed much more work on optimisation. Some of these optimisations are detailed next.

Table 7.4: Responsiveness of user interface

Run #	Number of FITS files	Database size	Processing time for 1st plot	Processing time for 2nd plot
1	13,962	0.03 GB	less than 1s	less than 1s
2	108,348	0.26 GB	18s	11s
3	363,991	1.00 GB	30s	34s
4	1,020,514	2.46 GB	2m 25s	1m 42s
5	1,639,842	3.95 GB	3m 52s	2m 50s
6	2,683,596	6.44 GB	6m 28s	4m 25s
7	11,905,931	28,63 GB	more than 3h	N/A
8	441,923,560	1183.50 GB	N/A	N/A

### Enhancing C3PO for this experiment

Producing a content profile for circa 0.5 billion FITS characterisation results is a challenging task for C3PO. The rough estimations show that it would take several weeks of processing time on a single server without sharding to ingest all the data in MongoDB and run aggregation procedure to generate a content profile. Further investigations pointed out at disk IO performance bottleneck, which occurs due to the small size of characterisation metadata file and a big amount of such files.

To solve this problem, C3PO was extended with an adaptor that generates a content profile of characterisation metadata directly while reading from storage. This improvement became possible due to the modular architecture of C3PO, where functionality of components may be easily extended through APIs. A gathering interface was implemented in a way that all needed data statistics aggregation and analysis are done on-the-fly, just after file read. There is no intermediate step of storing data in a database with following extraction of data. Also the adaptor was modified for parallel execution which scaled-up computations.

In order to evaluate the adaptor, a dataset was generated from a subset of FITS characterisation results of the large-scale dataset. A desktop PC with Intel 4 cores, 8GB RAM, 250GB Hard Drive and Ubuntu 13.04 OS was used for the tests. For experimentation, 10 first TGZ packages from each year of the collection (i.e. from the years 2005–2012) were selected. The runtime of C3PO with and without the adaptor was evaluated in order to measure improvement in computing performance. Figure 7.11 shows how much time is needed to feed a collection of FITS characterisation results into C3PO and generate a profile. To produce a profile in the normal procedure, the characterisation results must first be ingested into MongoDB and then the content profile generation procedure can be run.

As it may be noticed from the figure 7.11, the profile generation runtime on the normal mode is almost constant, because of small collection size fully stored in RAM for effective calculations. However, inner mechanisms of MongoDB take some time to initialise map-reduce jobs which take almost 60 seconds to start. The gathering runtime is dependent on a collection size and is limited by Disk IO. The developed adaptor skips the step of ingesting data into database, so there is only one process to run. Overall time to create a profile in the normal mode is equal to a sum of two process runtimes for gather and profile steps. The curve for the DirectProfile (a main function within the developed adaptor) execution does not increase as fast as gather + profile execution on the normal mode. With this adaptor, a content profile for the large-scale dataset was generated in less than 50 hours of processing time on the desktop computer previously described, processing around 3.000 files/second, more than ten times faster than the previous method.

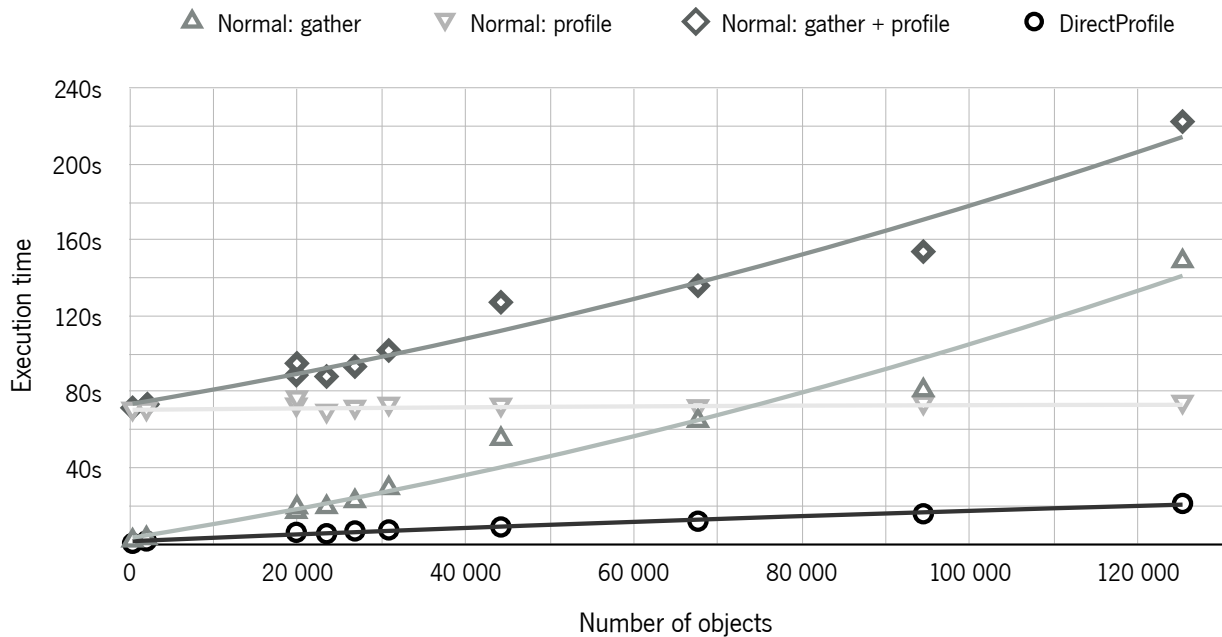


Figure 7.11: C3PO normal execution time compared with the performance of the DirectProfile mode

### 7.3.3 Results

Using the content profiles published in XML on an ad-hoc web server, Scout was able to import the data from the large-scale dataset into the system. Figure 7.12 shows Scout's content profile overview of the large-scale dataset with a total cumulative size of 13.14 TB and half a billion objects. From the statistics available on the file storage size in bytes property in the C3PO content profile the following Scout properties are calculated: collection size, objects average size and the objects minimum and maximum size; all with number data type and storage volume presentation hint. The file format histogram becomes a property named "Format distribution" with the data type "list of key-value pairs in plain text".

Figure 7.13 gives the details on the cumulative storage size of the same collection, calculated as the sum of each file size and accumulated with previous years. It shows that the current value of this property is 13.14 TB, but there is a past history for this property. The web archive backlog was processed so that past harvests would be presented on the correct timeframe. This particular collection refers to a set of web sites that were harvested every year from 2006 to 2013. The property history shows how the storage volume of the collection has grown steadily throughout the years, from 386.17 GB in 2006 to 13.14 TB in 2013. It can also be noted in figure 7.13 that every property value also refers the sources that have provided measures to this property, documenting the provenance of information, which is essential for trustworthiness and, consequently, for authenticity.

The same large-scale dataset has a format distribution with 4316 entries. For the creation of this format distribution, the reported format name was used instead PRONOM Unique Identifier, because PRONOM has a limited set of supported formats. Due to this, some incorrect or superfluous file feature information escaped into the file format name in some cases, making the format distribution much bigger than it would need to be. Table 7.5 shows the top 10 of the format distribution results, please note that there is an high percentage of conflicts (13.9%), but the previously presented

### Properties

Name	Value	Action
Collection size	13.14 TB	
compression_scheme distribution	62 key-value pairs	
Format distribution	4316 key-value pairs	
Objects avg size	25.8 kB	
Objects min size	0 bytes	
Objects max size	4.93 GB	
Objects count	546924526	

Figure 7.12: Scout's content profile overview of a large-scale collection from the Danish Web Archive

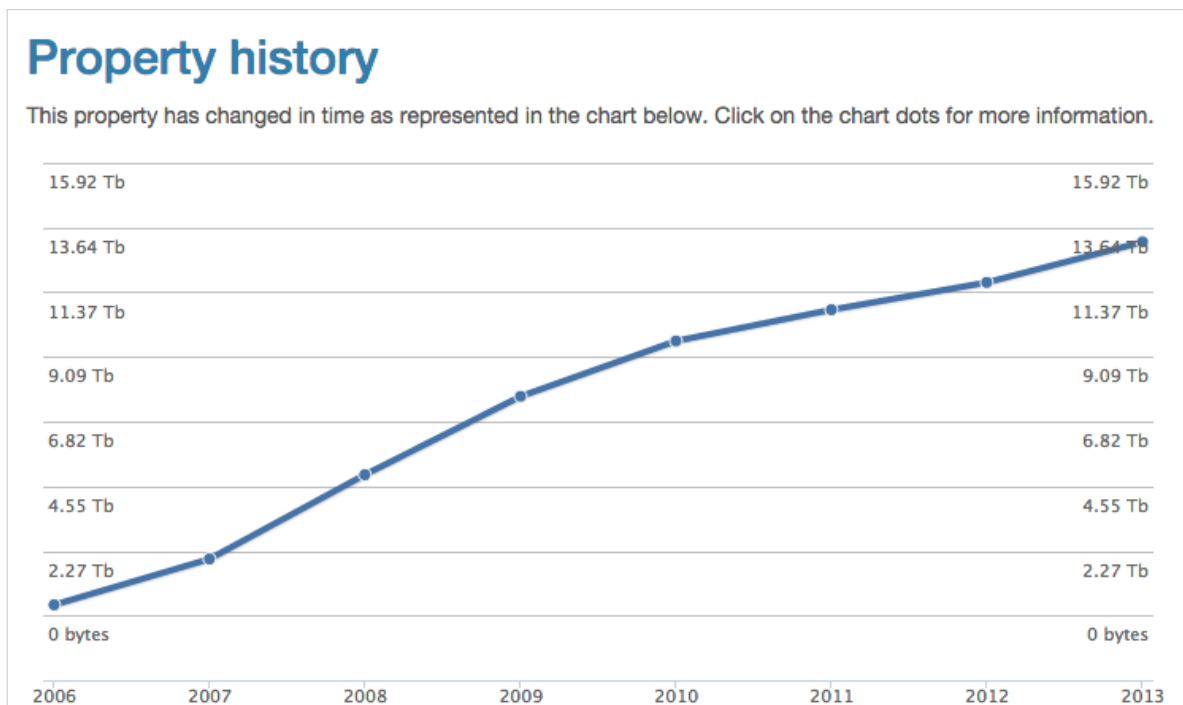


Figure 7.13: Scout's diagram of the cumulative history of the storage size property



Table 7.5: Top 10 of the most common formats of content from all harvests

Format name	Number of files	Percentage
JPEG File Interchange Format	157,939,162	28.88 %
Hypertext Markup Language (HTML)	147,345,550	26.94 %
Graphics Interchange Format (GIF)	87,649,358	16.03 %
Conflicted	75,995,185	13.90 %
Plain text	41,754,511	7.63 %
Portable Network Graphics (PNG)	17,682,954	3.23 %
Extensible Markup Language (XML)	7,484,183	1.37 %
Exchangeable Image File Format	4,078,432	0.75 %
Portable Document Format (PDF)	3,490,794	0.64 %
MS Windows icon resource (ICO)	604,771	0.11 %
<b>Total</b>	<b>544,024,900</b>	<b>99.47%</b>

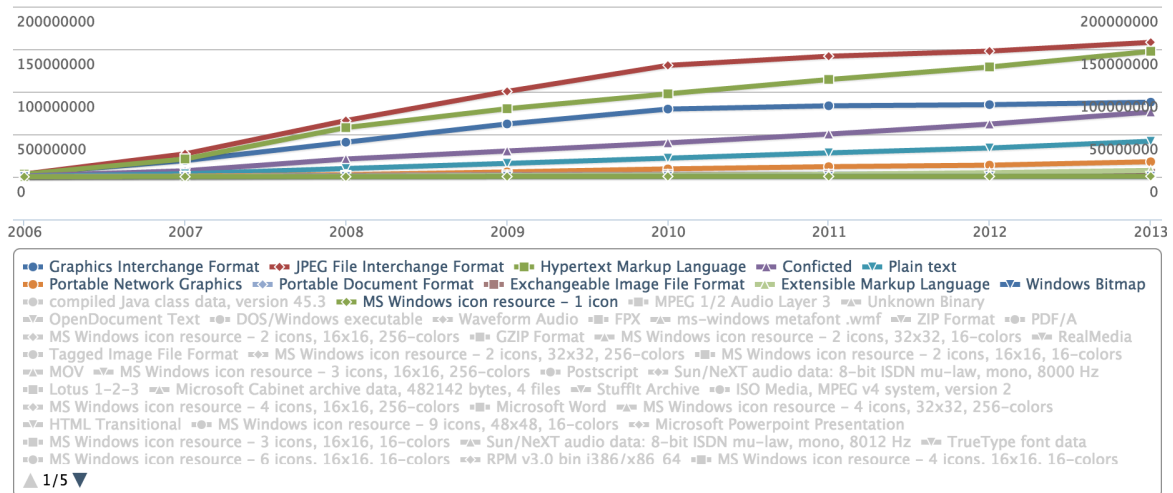


Figure 7.14: Cumulative history of the top 10 formats in format distribution

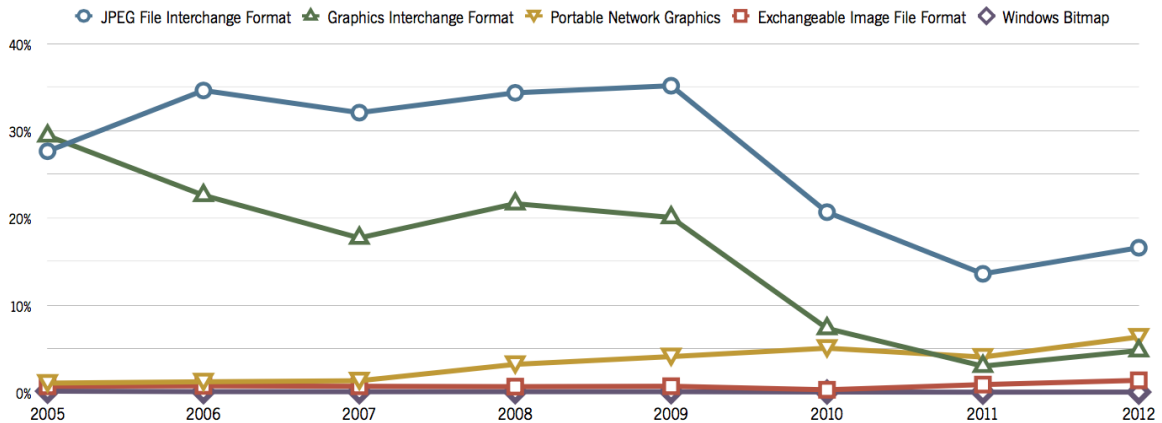


Figure 7.15: Non-cumulative history of the top 5 image formats distribution

conflict reduction approaches were not available at the time of the experiment.

The cumulative evolution of the file format distribution, filtered by the top 10 formats for readability, is available in figure [7.14](#). For long the growth of the number of JPEG images outpaced the HTML files, but this trend is inverting and it soon can be expected for the HTML files to become the most common file in the web archive whole history. The growth of HTML files number, accompanied with a growth of the number of plain text files which would probably be CSS files<sup>40</sup> point to a change of the paradigm for creating layout and formatting of web pages that would use less images and more HTML and CSS.

A comparative analysis of image format distribution is provided in figure [7.15](#) where the value of each year does not accumulate with previous ones, and the number of image files is given as a percentage of the total number of files retrieved on that year's harvest. This allows a direct comparison of the most popular image file format for each year. In the figure it can be easily observed that the Graphics Interchange Format (GIF) was the most popular image format in 2005, but it was quickly dethroned by JPEG File Interchange Format on the following year, and continued to decrease on popularity throughout the years. In 2011 the Portable Network Graphics, which had been steadily increasing since 2007, surpasses the GIF file format to become the second most widely used image file format, and first one of the lossless category, on this sample of the World Wide Web. This information is of great value to planning when selecting the preservation format to which image files should be normalised to.

<sup>40</sup>Cascade Style Sheets (CSS) is a language to describe the look and formatting information for documents written in markup language like HTML.

## 7.4 Experiment 4: e-Journal archive services completeness

### 7.4.1 Scenario

As scholars have become increasingly reliant on electronic versions of scholarly journals (e-journals), long term preservation of these resources has become of major importance and a growing need for the library community. The shift to journal content that is digital, online and held remotely has changed the responsibility that libraries have in ensuring the continuity of access to these materials. (Faria et al., 2013a)

Online libraries, like b-on<sup>41</sup>, provide content to their users by maintaining publisher subscriptions that allow access to content on the publishers own online platform. In the past, libraries have assumed preservation responsibility for the materials they collect, while publishers have supplied the materials libraries need. These well understood divisions of labour do not work in a digital environment and especially so when dealing with e-journals. These periodicals frequently change publishers or the publishers merge or cease to exist. When these events happen, libraries loose access to the journal and even to the whole publisher online platform, and consequently the library users loose access to journals, even issues that were previously accessible.

Currently, there are three leading organisations that have agreed to act as a last resort of e-journal content: Portico<sup>42</sup>, CLOCKSS<sup>43</sup> and e-Depot<sup>44</sup>. All three work very closely together and are involved in the Keepers registry which is a resource to address "who is looking after which e-journal, how, and what are the terms of access?"<sup>45</sup>.

For these archival services to effectively serve as as final safety net for e-journals preservation, coverage is of the outmost importance, but gathering all e-journal titles is a daunting task. According to Ulrichsweb<sup>46</sup> there are over 35,000 peer reviewed journal titles within the academic realm. Over 65% of them, about 23,000, are online journals. According to EBSCO<sup>47</sup> there are over 5,000 publishers who provide 25,000 electronic journals. The Web of Science<sup>48</sup> confirms that there are at least 12,000 e-journals from 3,200 publishers.

Figure 7.16 clearly shows how the required number of subscribed publishers increases exponentially as the archival services try to have all existing e-journal titles. The 100 largest publishing companies have nearly 70% of all available titles. To reach 80% of titles the archival services need to subscribe more than 500 publishers, and beyond that there is a huge long tail. EBSCO details that there are 466 publishers with two journals and around 4.000 publishers with only one journal. A similar view is given by Scopus<sup>49</sup>, the citation-index of Elsevier, that in 2009 counted almost 5,000 journal publishers in its database, 97% of them publishes 1-10 journals. This is, however, a significant part of the available journal articles, over 30%. For an e-journal archiving service it is fairly doable to sign agreements with the largest publishing companies and ingest their content into the archive, but after that the real work begins when trying to tackle with the long tail adding also that each year over 1.5 million scientific articles are published.

---

<sup>41</sup>b-on: Biblioteca do Conhecimento Online, enables permanent and unlimited access to e-journals and e-books by Portuguese higher education and research institutions. <http://www.b-on.pt>

<sup>42</sup><http://www.portico.org>

<sup>43</sup><http://www.clockss.org>

<sup>44</sup><https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources>

<sup>45</sup><http://thekeepers.org>

<sup>46</sup><http://www.ulrichsweb.com>

<sup>47</sup><http://www.ebsco.com>

<sup>48</sup><http://wokinfo.com>

<sup>49</sup><http://www.scopus.com>

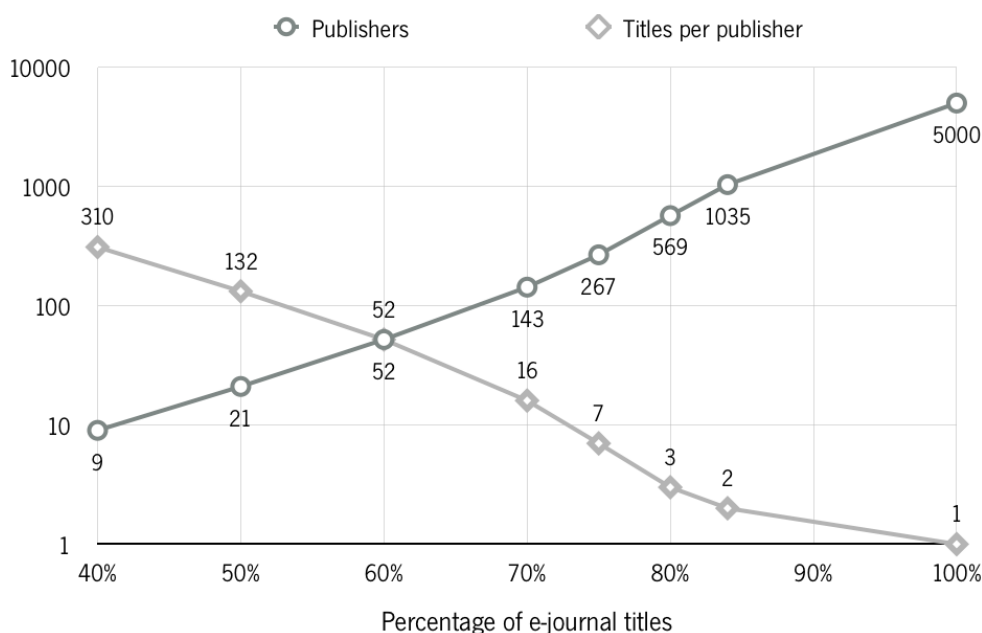


Figure 7.16: Relationship between repository completeness, number of publishers and number of e-journal titles per publisher, according to EBSCO (logarithmic scale)

Setting aside global coverage, even keeping current collections complete can be troublesome. Besides publishers getting out of business, journal title rights are often transferred between publishers. Knowing when such transfer has occurred, which publisher has taken over the title and under which conditions is crucial for archival services to keep collection completeness and also for libraries to maintain access to journals by their users. The Transfer Code of Practice from the UK Serials Group gives a set of rules for transferring journals, it "responds to the expressed needs of the scholarly journal community for consistent guidelines to help publishers ensure that journal content remains easily accessible by librarians and readers when there is a transfer between parties, and to ensure that the transfer process occurs with minimum disruption"<sup>50</sup>. But publishers do not follow these rules or do so very late. Administrative handling has no priority for a publisher and is only done months after the actual transfer. This is very problematic as archives expect titles to be received from publishers but after a transfer it suddenly ceases to receive the title anymore. This hinders the coverage and completeness of the archive and also brings a great deal of work in finding out where the title has gone and who is the new publisher.

Translating this problem to preservation policies, it can be stated that the high level aim is to create a complete collection of international scholarly e-journals for long-term access and preservation by acquiring these e-journals from publishers in order to serve the research community in case university repositories are no longer able to do this. In order to achieve this mission, various preservation procedural policies need to be developed. The list of scholarly e-journals needs to be identified and related publishers need to be contacted. Once the relationship is established via an agreement, regular monitoring needs to take place to assure that changes can be dealt with and that the goal of "completeness of the journal collection" can be continuously achieved. For this monitoring, detailed control policies will need to be established. The following list describes indications of the situations that can occur:

<sup>50</sup><http://www.uksg.org/transfer>

- Publisher  $A$  had journal  $J$ , is there a journal  $J$  provided by publisher  $A$  at time  $T_1$ ?
- Publisher  $A$  had journal  $J_1$  and the journal has been renamed to  $J_2$  (i.e. has changed title or ISSN), is there a journal  $J_2$  provided by publisher  $A$  at time  $T_1$ ?
- Publisher  $A$  transferred journal  $J$  to publisher  $B$ , is there a journal  $J$  provided by publisher  $B$  at time  $T_1$ ?
- Journal  $J$  has ceased to exist, what is its most recent issue?

In order to monitor these situations, the search results need to be filtered automatically, based on control policies. This experiment is limited to the investigation whether it is feasible to acquire relevant information from the Web and relevant registries. This relevant information relates to existing scientific journals, identified by title or ISSN, and journal-publisher relations that specify which publishers provide a certain journal. This information is manually maintained by registries within the e-Depot and also in other similar repositories and it is aggregated in the Keepers registry. But the information in the registries is only relative to the journals they collectively keep, being difficult to use it to ascertain the completeness of the journal safeguarding. Furthermore, due to the manual processes involved and the lack of cooperation from the publishers, the information in the registers is incomplete and outdated. Nevertheless, publishers provide this information on their Web sites in natural language. So there is a possibility here for expanding and improving the available information by using information extraction technologies. The following experiment will focus on using information extraction technologies to gather information that would allow us to detect the first situation described above, whereas a journal  $J$  is provided by publisher  $A$  at time  $T_1$ .

### 7.4.2 Execute experiment

Information Extraction (IE) is the task of extracting structured information from unstructured data such as natural language text. The goal of IE is to make information machine readable. Extracted information is often represented in the form of subject-predicate-object triples, where each triple is the *instance* of one semantic *relation*. The *predicate* indicates the relation to which the triple belongs, while the *subject* and *object* are the two entities between which the relation holds. IE methods use *extractors* that are either manually created or trained using machine learning methods to sieve through large volumes of text and distil for each relation a list of relation instances.

IE methods often use a pattern matching approach where for each relation a set of *patterns* is defined that when matched indicate a relation instance. The simplest example of such pattern-based information extraction might be a regular expression that finds e-mail addresses in Web pages. Patterns are often defined as *lexico-syntactic* patterns (Akbiik and Bross, 2009) that match natural language text. For example, to find which companies have acquired other companies (such as Google buying Motorola), a relation named *CompanyAcquisition* can be defined. A lexico-syntactic pattern can be constructed as "[X] acquired [Y]", where "[X]" and "[Y]" are placeholders for the subject and the object entities respectively. If this pattern matches a statement in natural language text, a triple for the corresponding relation is extracted. So, if the extractor encounters the sentence "Google acquired Motorola in 2011", the relation instance *CompanyAcquisition*(Google, Motorola)<sup>51</sup> is extracted.

As each relation is expressed in natural language text in a multitude of ways, one core challenge of Information Extraction methods is finding all patterns that belong to a specific relation. When aiming to extract relations from an

<sup>51</sup>Often, relation instances are denoted by first giving the predicate (i.e. relationship type) in camel case, and then the subject and object entities in brackets separated by a comma.

open domain corpus such as the Web, this problem becomes more challenging as there may be an unbounded number of relations, for each of which one extractor with a set of patterns must be defined. When interested only in information from a specific domain of the Web (such as digital content preservation), another challenge is to identify and gather relevant natural language text upon which Information Extraction is performed. Current IE methods are mostly limited to working on *explicit* statements in natural language text; reasoning or inferring knowledge from implicit statements is a topic of current research (Lao et al., 2011).

A range of research investigates how to reduce the workload of manually defining patterns with machine learning mechanisms that require different amounts of supervision. Approaches range from supervised (Mooney, 1999) or declarative (Krishnamurthy et al., 2009) approaches, to weakly supervised (Mintz et al., 2009) and unsupervised methods (Akbik et al., 2012). Supervised and declarative approaches generally produce high quality extractors, albeit at a cost in human effort, while unsupervised approaches are useful for information discovery.

In this experiment both declarative (Akbik et al., 2013) and unsupervised (Akbik et al., 2012; Akbik and Löser, 2012) approaches are applied to find a list of journal titles discussed in the Web, as well as a list of journal-publisher attributions in order to discover which publisher publishes which journal. The experiment consists of three steps: 1) execute a *data acquisition and pre-processing step* that first gathers relevant natural language data from the Web; 2) perform *relation discovery* on this data to mine frequent extraction patterns and gain an insight into the semantic content of the crawled corpus; 3) assign patterns to the relations that should be mined from the corpus and execute a *relation extraction step* to mine instances for each relation of interest.

### Data Acquisition and Pre-Processing

The first phase of the experiment is a data acquisition and pre-processing pipeline. Its goal is to gather large amounts of natural language text from the Web that has a high probability of containing statements that relate to this use case.

A focused crawler was implemented to address this task. It uses a list of seed keywords (such as publisher names and journal titles) and formulates a search query using a Web search engine API<sup>52</sup> for each keyword on the list. Each query returns a list of Web pages that is automatically crawled and processed with Natural Language Processing (NLP) tools. Boilerplating is applied to extract blocks of natural language text from each Web page, removing other Web page elements such as layout information or advertisements. Sentence segmentation is applied to divide blocks of text up into sentences that can be analysed individually. Finally, all sentences that do not contain at least one of the seed keywords are filtered out.

The resulting dataset consists of approximately 18 million sentences gathered from 500,000 Web sites. The total text size is 8 GB. The seed keyword list consists of 12,000 entries. A sample of seed keywords and gathered sentences is illustrated in Table 7.6. These example sentences contain information relevant to domain in scope.

### Relation Discovery

The second phase of the experiment is to discover what kind of relations are expressed in the dataset. Because manually going through a dataset of this size is infeasible, a relation discovery mechanism is applied to identify prominent extraction *patterns* in the text. The used method, explained in detail in (Akbik et al., 2012), counts and groups patterns according to distributional evidence in the corpus. This yields a list of prominent patterns in the

<sup>52</sup>The Bing API is used in this pipeline, available at <http://www.bing.com/developers/> (last checked at 2015-09-13)

Table 7.6: Sample data from the data acquisition and pre-processing pipeline.

Seed keyword	Sample sentence retrieved from the Web
Elsevier	<i>"In 1991, two years before the merger with Reed, Elsevier acquired Pergamon Press in the UK."</i>
The Asia-Europe Foundation	<i>"The Asia-Europe Foundation (ASEF) sold the Asia Europe Journal and transferred the copyright to its long-time partner Springer."</i>
Acta Chirurgica Iugoslavica	<i>"Acta Chirurgica Iugoslavica is available free of charge as an Open Access journal on the Internet."</i>
American Journal of Preventive Medicine	<i>"The American Journal of Preventive Medicine is the official journal of the American College of Preventive Medicine and the Association for Prevention Teaching and Research."</i>
Journal of Business Ethics	<i>"In 2004 the Journal of Business Ethics merged with the International Journal of Value-Based Management and Teaching Business Ethics."</i>

corpus, a sample of which is given in table 7.7. These patterns indicate that the dataset gathered by the focused crawler is indeed relevant to the target domain and suggests relationship types for which extractors can be created. Note that the used patterns are actually more complex lexico-syntactic patterns, but the syntactic elements (which denote grammatical properties of the patterns) are not indicated for the sake of readability.

Table 7.7: Top pattern in the gathered corpus.

Pattern	Rank #
[X] journal of [Y]	1
[X] published by [Y]	2
[X] journal on [Y]	3
[X] journal published by [Y]	4
[X] available as [Y] journal	5
PubMed [X] [Y]	9
[X] science proceedings of [Y]	25
[X] subscription available to [Y]	30

### Information Extraction

The third phase of the experiment has the objective of finding two relations in the crawled document collection: an extractor for journal titles (IsJournal) and an extractor for journal-publisher attributions (JournalPublisher). For each extractor, the top patterns found in the relation discovery step were manually selected to use for relation extraction. For example, the patterns "[X] journal of [Y]" and "[X] journal published by [Y]", among others, were selected of the JournalPublisher relation.

Both extractors were executed on the document collection and all found relation instances were stored in two lists: one with all journal titles found in the Web crawl, and another with all identified journal-publisher attributions.

### Information insertion into Scout

The resulting lists of journal titles and journal-publisher attributions conform to the formally specified and normalised information source restriction of Scout. This information can be inserted into Scout via a new plug-in, allowing this information to be included into the central knowledge base. Queries and notification triggers can then be created using the information on the knowledge base to alert when journals change publishers, or even to cross-relate an institution's list of subscribed publishers and journals of interest to alert when a journal of interest is no longer provided by any of the subscribed publishers.

The process of finding new journals and journal-publisher attributions used in this experience can be frequently repeated to allow automatic constant monitoring of these aspects of the world, automatically notifying interested users when the preservation risk of not acquiring a journal becomes relevant.

### 7.4.3 Results

In the experiment, a list of 2,000 journal titles and a list of 500 journal-publisher attributions were generated. The results were evaluated both automatically and manually against the e-Depot publishers. In the automatic evaluation, the results were matched against the e-Depot to find out how many of the extracted titles were already contained in the e-Depot internal registry<sup>53</sup>. Of the 2,000 journal titles only 200 were in the e-Depot, making the remaining 1,800 titles candidates for inclusion. Manually verification of a simple random sample of 200 of these titles<sup>54</sup> allowed to find out that 191 (95.5%) are titles that should be added to the registry and 9 are false-positives (4.5%).

Manually repeating this experiment with the more complete Keepers Registry found that more than 50% of all journal titles and 50% of all attributions were not in the registry, having this time 15% of false-positives and, consequently, 35% of viable candidates for inclusion in the registry. This indicates a strong potential of using Information Extraction technologies to help in keeping such registries complete and thus aiding the task of preservation monitoring.

Table 7.8 lists example instances of the JournalPublisher relation. The sample was chosen by sorting the list of all instances alphabetically by journal title and selecting the first 17 instances. The table illustrates which of these instances are already listed in Keepers Registry and which should be added to make it more complete. Some entries in the list have comments to illustrate error classes such as encoding errors or entity name boundary detection errors.

One major problem that affects extraction quality (i.e. the portion of results that are fully correct) is the lengthy nature of some journal titles or publisher names. An example of this is the "*European Journal of Nuclear Medicine and Molecular Imaging*". This causes the used method to detect wrong title boundaries in some cases; titles might be too short or too long, encompassing either only a portion of the words of the real title, or additional words that do not belong to it. The used method was adapted to cope with this, but more fine-tuning on the extractors to this specific domain will arguably increase overall extraction quality.

The information above can be directly used to answer the first scenario, whereas a journal  $J$  is provided by publisher  $A$  at time  $T_1$ , which is the time of data acquisition. The same IE pipeline can be frequently executed to get new snapshots in time, providing a continuous monitoring of this situation. Automatic monitoring of the continuity of e-journal availability can be done by cross-referencing this information about journal-publisher relations with the list of e-Depot

<sup>53</sup>the e-Depot archiving service contains an internal registry with the journal titles it archives and its related publishers, this internal registry is aggregated by the Keepers registry

<sup>54</sup>200 simple random sample has a  $\pm 6.6\%$  sampling error on a 95% confidence interval



Table 7.8: Sample of results and comparison to Keepers Registry.

Extracted relation instances		Evaluation	
Journal	Publisher	In Keepers Registry?	Comment
A Journal of Human Environment	Royal Swedish Academy of Sciences	<b>no</b> , should be added	
AAPS Journal	Springer Science + Business Media LLC	<b>yes</b>	
AAPS Journal	American Association of Pharmaceutical Scientists	<b>no</b> , should be added	
Academic Emergency Medicine	Society	<b>no</b> , error should be corrected and instance added	Error in entity detection of publisher name. It should be: " <i>Society of Academic Emergency Medicine</i> "
ACEEE International Journal of Network Security	ACEEE-Network Security Group	<b>no</b> , should be added	
Acta Applicandae Mathematicae	Springer	<b>yes</b>	
Acta Automatica Sinica	Chinese Association of Automation and Institute of Automation	<b>no</b>	"Acta Automatica Sinica" is listed only as published by "Elsevier".
ACTA AUTOMATICA SINICA	Chinese Association of Automation and Institute of Automation	<b>no</b>	All caps duplicate of previous relation
Acta Biomaterialia	Elsevier	<b>yes</b>	
Acta Geologica Slovaca	Comenius University in Bratislava	<b>yes</b>	
Acta Materialia	Elsevier	<b>yes</b>	
Acta Polytechnica Hungarica	óabuda University	<b>no</b> , error should be corrected and instance added	encoding error
Acta Radiologica	Scandinavian Society of Radiology	<b>no</b> , should be added	"Acta Radiologica" is listed with other publishers.
Aequationes Mathematicae	Birkh	<b>yes</b>	Encoding error in publisher name. Should be "Birkhäuser Verlag"
African Journal of Biomedical Research	Biomedical Communications Group	<b>no</b> , should be added	
Agricultural Economics	IAAE	<b>no</b>	"Agricultural Economics" is listed with other publishers. "IAAE" is an abbreviation for a missing publisher on the list.
Agronomy Journal	American Society of Agronomy	<b>yes</b>	

paid publisher subscriptions (throughout time) and the list of e-journals available in the e-Depot repository. Nevertheless, for the other scenarios, more information about the journals needs to be captured, like the journal renaming or ceasing. Also, machine readable information on the publisher subscription and e-journal issues available in the repository needs to be inserted into Scout in order to automatically cross-reference and discover other entailed preservation risks. These steps are some of the future work to further study the use of information extraction technologies on digital preservation processes.

Revisiting the list of patterns in table 7.7 it should be noted that only a small portion of the top patterns were used on the experiment. Incorporating additional patterns may lead to more complete extraction results. More importantly, there are many types of information in the crawled corpora that were not extracted but may also be of interest to the community. For example, the pattern PubMed [X] [Y] indicates that information on PubMed entries is contained in the corpus. Similarly, the pattern [X] journal on [Y] indicates that it is possible to extract topics for journals. Accordingly, this indicates potential for expanding the range of information extracted in future experiments.

This experiment shows that the information extraction technologies has potential not only for detection of real-time threats for digital preservation domain, but also for parsing historical knowledge to capture descriptive information and becoming an important tool for librarians and archivists to cope with the increasing scale of digital content production.

## 7.5 Final remarks

The survey presented in chapter 5 identified the most important and neglected preservation threats, answering the first research question, and also determined the most well accepted methods to identify these threats. Scout, presented in chapter 6 provides an approach to monitor and gather information from the world into its knowledge base, so it is able to automatically detect the identified threats. Scout is then proposed as the software artefact that bases an hypothesis for the second research question: Scout can detect the most important and neglected preservation threats by using a format representation of the information about the world that is automatically monitored and collected. This hypothesis is then proved by the set of experiences presented in this chapter.

The first experiment, the SCAPE preservation suite, focuses firstly on monitoring and detecting when content does not conform with the defined institutional policies, successfully detecting the threat when content that does not conform to the policies is ingested. The same experiment is then expanded to include a series of other preservation threats, including monitoring when preservation actions are not having the expected results, monitoring file corruption and monitoring when preservation plans become outdated. The experiment also shows how to further expand the scope to include producer and consumer misalignment threats, covering 5 out of the 6 threats identified as the most important and neglected.

The second experiment, checking external references, focuses on the top scored threat that is not covered on the previous experiment: there is not enough context information to understand the file content. It does so using the method of automatic verification that external references still exist, specifically web sites, which is well accepted by the community and yet not very used, bringing then the most impact to the community. This experiment is able to detect the threat on a large-scale real-world example: open-access repositories around the world. Results show that a large percentage of files have already lost much of the context information, urging for actions to mitigate the threat.

The third experiment, web archive deep characterisation, focuses on the inference of trends in the general technological environment, which give an insight to the consumers and producers technological landscape and, consequently, can

help to detect producer and consumer misalignment threats, and also provide information necessary to determine the best actions to undertake. This experiment also proves that Scout is able to cope with the scalability requirements that come from real-world scenarios. This experiment proves that Scout is able to detect trends in the technological environment, with the specific example of image format popularity.

The forth experiment, e-Journal archive services completeness, demonstrates how domain-specific problems can be monitored, and how general information available on the Web can be automatically monitored, collected and converted to a formal representation using information extraction technologies. This experiment proves that Scout can be used even when there is no formally expressed source of information by gathering information from the Web and extract from it the semantic information that will be useful for detecting preservation threats.



# Chapter 8

## Conclusion and future work

This chapter reiterates the presented work, discusses the main achieved conclusions, lists the most important contributions and defines a roadmap for future work.

### 8.1 Summary

As the world becomes dominated by digital technology, all forms of record keeping, heritage, services, goods and even art become dependent on digital information. Technological change is accelerating and today's digital content may very well disappear or become inaccessible if no preventive actions are undertaken. Digital content is produced at an impressive rate and although not all content is of enough interest to preserve, the selection and appraisal processes are onerous, resource intensive and sometimes impossible to undertake without hindsight.

But digital information is not only important for future historical uses but also for current day-to-day business operations. The success of most current human pursuits depends on the capacity to be and stay informed. From science, health, industry, art and cultural domains to personal management, social life, entertainment and transportation, our life revolves and depends on digital content and services.

Unfortunately, long-term safekeeping of digital content is more complex than its analogue counterpart. Digital content needs a technological stack in order to be consumed. This dependency also needs to be preserved in the long-term, or exchanged by a newer technological stack. The rapid evolution of digital technology becomes an enemy of older digital content as retro-compatibility is rarely a concern of software vendors, specially in emerging markets.

The problems that can afflict digital content are manifold: from hardware to software obsolescence, from loss of social-cultural context to lack of economical resources, staff, or even political leverage. Knowing which threats afflict the content is difficult and yet essential to initiate the processes that result on actions being deployed to mitigate threats. However, the sheer volume and heterogeneity of digital content make it very hard to keep track of all information needed to detect preservation threats. Furthermore, institutions commonly lack the tools to analyse digital content in such proportions, and even when such tools exist it is uncommon for an institution to have staff with the technical skills required to execute and analyse these tools and, at the same time, have the skills needed to understand and identify the preservation threats that might afflict content.

Current state of the art fails to address the main problem: lack of information. Identification of preservation threats

cannot be based on file format registries alone. Digital preservation specialised information is scarce and outdated. To effectively detect preservation threats it must be possible to monitor any source of information. Furthermore, the information monitored and collected from these sources should be formally represented, so cross-referencing information from different sources can be used to identify threats. The formal representation of information about the world of interest for digital preservation also serves to fill an important gap in current state of the art whenever authenticity is a concern: it provides evidences for the grounds that base decision-making, providing transparency on the processes that avoid pernicious actions to be taken due to self-sabotage or negligence.

This work also proposes a new set of roles for preservation watch: to be the initiator and auditor of planning (decision-making) and operations (action-taking) processes. The watch process detects preservation threats, relaying them to the planning process, so an action can be chosen and then executed by the operations process. The result is again monitored by watch to ensure that threats were mitigated and find new threats that might be introduced by the actions or from the environment. This feedback loop creates a continuous cycle, named preservation life-cycle, that continuously adapts digital content to the environment and ensures its long-term accessibility.

The watch process should therefore be able to find which preservation threats initiate planning, both to create or re-evaluate plans. But there is an unknown and presumably enormous number of possible preservation threats, so a prioritisation is in order. Then, the watch process should be able to monitor, collect and formally represent information about the world of interest for digital preservation, so these threats can be detected. These requirements can be formalised on the following research questions:

**Question 1.** *What are the most important and neglected digital preservation threats?*

**Question 2.** *Can these threats be detected using a formal representation of the information about the world that is automatically monitored and collected?*

To find the most important and neglected threats, the digital preservation community was inquired using an online questionnaire. The survey allowed to ascertain the community's perspective of the most important threats and their level of practice on monitoring these threats. The community's opinion plays here an important role, as one of the major objectives to achieve on this process is trustworthiness, which is mainly based on subjective factors. The survey also asked for the most well accepted methods to detect these threats, and their corresponding use by the community. The method preference implies trust on these indicators, proving their worth on detecting threats and serving as evidence that preservation actions are needed.

The survey results answered to the first research question and provided the base requirements for the creation of a software artefact named Scout that would help to answer the second research question. Scout implements the defined requirements of automatic digital preservation threat detection by monitoring and collecting information about the world that is mapped to a formal representation inside Scout's knowledge base. It defines a set of information sources based on the digital preservation community's preferred methods. Furthermore, it provides a mechanism for detecting preservation threats based on the gathered information, notifying relevant parties so the planning process can be started.

Scout serves as an instrument for a series of experiments that focus on the most important and neglected threats and are based on real-world scenarios, proving that Scout can in fact automatically detect the most important and neglected threats.

## 8.2 Conclusions

This section outlines the main results of this work and describes how they respond to the proposed research questions. It also digresses on discussing some of the issues found on the taken approaches.

### **What are the most important and neglected digital preservation threats?**

To answer the first research question, a survey to the digital preservation community was conducted. The survey was implemented as an online questionnaire using a convenience sample that was reached using the several communication channels available to the SCAPE project, as the project website and newsletter; the Open Planets Foundation blog, mailing list and social networks; the Digital Preservation Coalition news page; and the JISC, DigLib, nestor, DLM Forum, Digital Curation Centre and DigCurV mailing lists. As a result of the used communication channels, or simply due to the current composition of the digital preservation community, the sample was mainly composed of institutions from Europe and North America, mainly from universities and memory institutions or content holders.

The survey showed that the top score of the most important and neglected preservation threats was filled by digital preservation domain-specific logical-level threats like consumer misalignment, lack of context information and incorrect action results. Following, there were the threats on the planning and policies processes, like outdated preservation plans and producer misalignment with ingest policies. Physical-level threats, like file corruption or backup failure, were considered the most important ones and are relatively highly monitored. Still, file corruption was found to be too neglected for its importance, as 18% of organisations do not monitor it. The final prioritisation, defined by a score formula that gives an equal weight to the perceived importance and to the neglect on monitoring it, results on the following list of threats ordered by decreasing priority:

1. A relevant percentage of the consumers cannot read the disseminated file format
2. There is not enough context information to understand the file content
3. Actions executed on files (e.g. file format migration) are not having the expected results
4. Defined preservation plans (e.g. defined preservation format) became outdated
5. File corruption
6. A relevant percentage of the producers cannot comply with the established ingest policies

The list above already answers the first research question. Nevertheless, the survey also indicated which, in the view of the respondents, are the most adequate methods to monitor the threats identified above. These results are a very important input for the artefact that responds to the second research question. A similar formula was devised to calculate the preferred monitoring methods that are least used, pinpointing gaps that may be exploited to improve the impact of the developed tools in the community. The result is a list of top-scored methods for the top-score threats, ordered by decreasing priority:

1. Automatic cross-reference of file formats in your collections with information available on format registries

2. Automatic verification that external references still exist (e.g. web sites)
3. Run quality assurance tools on action results and automatically check against expected results
4. Use tools to create preservation plans (e.g. Plato) and be automatically notified when assumptions made become invalid
5. Automatic file fixity checks
6. Monitor ingest process (e.g. SIP rejection statistics)

### **Can these threats be detected using a formal representation of the information about the world that is automatically monitored and collected?**

The most important and neglected threats and the preferred methods to detect them, together with the requisites for effective and trustworthy preservation watch defined above, allow to define a set of requirements for Scout, the software artefact that implements a novel approach for preservation watch. This artefact is used as the base to answer the second research question.

Scout defines an ontological model for an internal knowledge base which allows the representation of any information about the world while keeping some form of control on how this information is structured, using a vocabulary of entity types and properties. This model also enables the update of information, keeping the whole history of changes and recording the date of these changes. This allows trends to be inferred. Furthermore, the knowledge base allows for the traceability of the information by keeping the provenance of values, documenting the source from which the property value measurements were acquired.

This knowledge base is populated with information from external sources, like repository content and events, format registries, experiments or organisational objectives. The information is brought into the system via source adaptors, pieces of software logic that can be easily created and added to Scout to monitor and gather information from external sources. The source adaptors are set to run periodically and fetch information from outside, mapping it to a formal representation compatible with the knowledge base. A REST API also allows information to be pushed into the knowledge base. The knowledge base can be browsed and queried on to find significant events that indicate preservation threats. Triggers can be configured to monitor defined queries and notify relevant parties whenever significant events are suddenly found, providing an effortless way to continuously monitor threats that afflict digital content.

The prototype implementation of Scout shows the feasibility of the design that follows all steps of the digital preservation life-cycle and proves that all requirements set forth have been achieved:

1. **Enable to pose questions about entities and properties of interest**

Scout web interface allows human users to pose questions on properties on interest and continuously monitor them using triggers that will send an email notification when significant events are found.

2. **Collect information from different sources through adaptors**

Scout design allows source adaptor plugins to be developed so they collect information into the knowledge base, mapping and normalising information to the knowledge base specification.



3. **Act as a central place for collecting relevant knowledge that could be used to preserve an object or a content set**

All information is centralised into a knowledge base that allows complex querying. Information from different sources can be cross-referenced, mingled and filtered so significant events, that indicate preservation threats, can be extracted.

4. **Notify interested agents when an important event occurs**

Triggers can be created to monitor the knowledge base for significant events and send a notification to interested agents.

5. **Integrate with the decision-making process**

Trigger notification are sent to relevant parties, which enacts planning. Also, the planning process can create triggers to monitor the quality and performance of the defined action plans to verify if objectives were achieved and no new threats were introduced.

6. **Act as an extensible platform**

The plugin architecture allows to easily add new adaptors and notification channels, extending the platform functionality effortlessly. Also, the REST API allows other services to quickly integrate with Scout.

Scout is at the base of the hypothesis for the second research question: Scout can detect the most important and neglected preservation threats by using a formal representation of the information about the world that is automatically monitored and collected. This hypothesis was then proved by a set of experiences that focus on the most important and neglected digital preservation threats and define realistic scenarios where these threats would be of importance.

The first experiment, the SCAPE preservation suite, an implementation of the SCAPE preservation life-cycle, presents an easily repeatable experiment that has won the best demonstration award on the Digital Libraries 2014 conference, in London. It describes a deployment of the SCAPE preservation suite into a virtual machine, stitching together Scout with all other SCAPE developed software components that compose an architecture that supports all digital preservation processes, that run one after the other in a continuous cycle. This experiment presents a fictitious but common scenario that goes through all steps of the preservation lifecycle and demonstrate how almost all of the top scored preservation threats are or could be detected.

The second experiment, checking external references, focuses on the only threat not detected on the first experiment, the threat of losing context information, by developing a pipeline that is able to automatically verify if all external references to web sites still exist. The experiment was executed on a real-world and large corpus gathered from several open-access repositories. Results show that a large percentage of files have already lost much of the context information, 19% of files have at least one dead link and the average of the percentage of dead links in a file is at least 20%, with a tendency to grow as the file becomes older.

The third experiment, web archive deep characterisation, delves into the web archive scenario. Web archives preserve a very important part of the cultural heritage and are also a representative sample of the technological landscape. Analysis of web archive content infers trends on the producers and consumers technological environment, which on its turn finds preservation threats. This experiment also provides a testbed to benchmark Scout and connected tools with a large-scale corpus of around 13 TB and half a billion objects from the Danish State and University Library web archive. The results showed that Scout and connected tools, after some optimisations, were able to cope with the scalability issues that afflict many institutions digital collections. Also, the experiment allowed to analyse the trends

of image format popularity on the (Danish) Web from 2005 to 2012. The analysis shows JPEG dethroning GIF as the most used image format on the Web in 2006, the rapid decrease of GIF's popularity throughout the years, and the ascension of PNG image format in 2011, becoming the most widely used lossless image format on the (Danish) Web. This information is of vital importance to infer how well ingest policies align with content producers, and how well current digital content aligns with consumers. Also, this information is important for planning to base their decisions, e.g. to which lossless image format should it convert content, and is an important evidence that the selection of preservation actions is made upon well grounded facts. This improves transparency and credibility of the planning process.

The final experiment, e-journal archive services completeness, focuses on a problem specific to libraries, where the increasing reliance on electronic version of scholarly journals has introduced a major threat. The journals are now digital, online and held remotely, and may disappear at any moment when publishers cease to exist or merge, or when journals change publishers. As a result, libraries, and consequently their users, loose access to journals, even issues that were previously available and purchased. To mitigate this threat, libraries acquire the services of e-journal archives, like Portico, CLOCKSS or the e-Depot, that act as a last resort of e-journal content. But the effectiveness of e-journal archives depends on their completeness. This introduces a new threat, that e-journal archives are not complete. To monitor this threat, natural language statements were crawled from the internet and processed using information extraction technologies. The method allowed to find journal and journal-publisher relationships stated on the Web (e.g. on publisher websites) and to compare these with the e-journal archive registers like the Keepers registry. The results showed that 35% of the found journal and journal-publisher relationships were not on the registry, detecting a major threat on the safe-keeping of e-journals even when adhering to e-journal archival efforts.

The presented set of experiments demonstrate how Scout is able to automatically detect all of the most important and neglected threats, as found on the answer of the first research question. Moreover, the experiments show how Scout and related tools are able to cope with scalability issues present in many institutions in terms of volume and heterogeneity. The experiments also confirm that the approaches used in Scout can still be applied in scenarios where no well structured and complete source of information is available, using information extraction technology to draw out of the plethora of information available on the Web the knowledge of interest to digital preservation that can detect and measure the impact and probability of preservation threats. The experiments therefore validate the hypothesis that Scout is able to automatically detect preservation threats, using a formal representation of the information about the world that is automatically monitored and collected, answering to the second research question.

### **8.3 Contributions**

This thesis agglomerates several contributions considered of importance to a different set of research and commercial domains. Below are listed some of these contributions, grouped by the domain they mostly apply to.

#### **For research, culture and memory institutions**

- The insight into the digital preservation community opinion and practice on preservation watch, including the biggest worries and methods they mostly use, allows for these institutions to find quorum and consensus on which threats should be focused on and which methods should be used or developed;

- A new approach to preservation watch that focuses on information gathering from any source, its representation on the internal knowledge base, and the use of this information to detect threats, may be of use for institutions to effectively monitor preservation threats on large-scale and heterogeneous content sets;
- The proposal on how to deal with complex scenarios of format migration, how to compose these preservation actions with quality assurance processes, and how to document executed preservation actions and their quality assurance output, can help institutions to develop their own workflows for executing digital preservation actions and properly document them;
- The SCAPE preservation life-cycle and its implementation provide a demonstration on how to have support for all preservation processes on a continuous cycle, which may help institutions to endow their own repository implementations with watch, planning and operations processes;
- The experiments that, on their own, give important insight on current threats that afflict important parts of the scientific and cultural heritage.

#### **For business and public administration institutions**

- The review on the concepts of watch and their alignment with risk management concepts. This thesis also bridges the two domains, which may spark research on how to bring digital preservation to the business and public administration domains;
- A new approach for preservation watch provides a base for research on how the same methodology could be applied to monitor business operations and the risks that afflict them. In theory, the same technique could be applied to monitor not just preservation risks but also how well business objectives are being achieved, opening this approach to the domains of risk management, quality management, and information security management.

#### **For digital preservation practitioners**

- The developed open-source tools, like Scout and all tools belonging to the SCAPE preservation suite, can be of use for preservation practitioners to support their own preservation processes;
- The architecture provided, deployments, detailed scenarios, issues, workarounds and caveats described can be an important guide for practitioners to help on their own deployment of preservation enabled repositories;
- The experiments in the context of the web archive and e-journal archival services prove the value of the information that can be extracted from using methods, such as the inference of image format trends and the capability to extract e-journal related information, like publisher or subject categories, from natural language statements crawled on the Web. Practitioners can use similar methods to extract new value, study the impact or probability of new threats, and extract other types of knowledge from the information, opening the doors for new subjects of research;

### **For digital preservation vendors**

- The results of the digital preservation community survey, including the threats the community is mostly worried about and less capable of monitoring, is an important market information for preservation vendors that can develop solutions to monitor and/or mitigate the threats based on their priority;
- The presented architecture allows vendors of digital repositories to easily add new watch, planning and operations capabilities by implementing the repository APIs that enable them to integrate with the whole or part of the SCAPE preservation suite;
- The presented proposal on how to deal with the different granularities and workflows a preservation action and consequent quality assurance processes allows vendors of preservation enabled repositories to develop support for complex digital preservation actions;
- The presented proposal on how to encode quality assurance outputs in preservation metadata can be also of use for vendors to enable complete documentation of their quality assurance outputs;
- The presented approach for preservation watch can also inspire vendors to develop novel solutions that would use similar techniques in other markets, e.g. the business or the domestic market, or other domains, e.g. risk, quality or information security management domains;

## **8.4 Future work**

Many of the conquered challenges on this thesis uncovered new ones. Knowing that much work still ahead to make Scout ready for the community, here are some of the possible developments on the route to market:

- Create more question templates, based on the community needs, which would possibly include adding more sources of information;
- Support manual input of information, allowing human knowledge to also be a source of information;
- Allow manual correction, update, or improvement of information from manual sources;
- Develop a user feedback form to be integrated with repositories or content portals to allow consumers to provide feedback on their ability to read certain file formats, the adequacy of the formats to their needs, the correctness of descriptive metadata, the lack of context, etc.;
- Add more documentation on current structure of knowledge base and how to make advanced queries;
- Develop easier methods to provide content profiles to the system;
- Advertise and create a community around the project, with the backing of respected institutions of the domain (conversations with Open Preservation Foundation, the British Library, and the National Library of the Netherlands are ongoing);

- Further integration of developed technologies and approaches into a turn-key digital preservation solution (development of RODA 2.0 is ongoing, which will include features inspired by this research);
- Explore the possibility of having a federation of local Scout services to improve the quality and volume of the knowledge base.
- Explore the dynamics of a central instance of Scout versus a federation of local instances of Scout, and research mechanisms leverage public and private pockets of information, to improve data quality and volume.



## Bibliography

- Abrams, S., Morrissey, S., and Cramer, T. (2009). "What? So what": The next-generation JHOVE2 architecture for format-aware characterization. *The International Journal of Digital Curation*, 4(3):123–136.
- Airmic, Alarm, and IRM (2010). A structured approach to Enterprise Risk Management (ERM) and the requirements of ISO 31000. Technical report, The Association of Insurance and Risk Managers (Airmic), The Public Risk Management Association (Alarm) and The Institute of Risk Management (IRM).
- Akbik, A. and Bross, J. (2009). Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns. In *Workshop on Semantic Search in Conjunction with the 18th Int. World Wide Web Conference*.
- Akbik, A., Konomi, O., and Melnikov, M. (2013). Propminer: A workflow for interactive information extraction and exploration using dependency trees. In *ACL System Demonstrations*. Association for Computational Linguistics.
- Akbik, A. and Löser, A. (2012). Kraken: N-ary facts in open information extraction. In *AKBC-WEKEX*, pages 52–56. Association for Computational Linguistics.
- Akbik, A., Visengeriyeva, L., Herger, P., Hemsén, H., and Löser, A. (2012). Unsupervised discovery of relations and discriminative extraction patterns. In *Proceedings of the 24th International Conference on Computational Linguistics*.
- Antunes, G., Barateiro, J., Becker, C., Borbinha, J., Proença, D., and Vieira, R. (2011). Shaman reference architecture (version 3.0). Technical report, SHAMAN Project.
- Asseg, F., Bacall, F., Barton, S., Castro, R., Hahn, M., Plangg, M., Schenck, M., Schmidt, R., and Withers, D. (2013). Architecture design: first version. Deliverable D4.1, SCAPE Project Deliverable (D4.1).
- Baker, M., Keeton, K., and Martin, S. (2005). Why traditional storage systems don't help us save stuff forever. In *The First IEEE Workshop on Hot Topics in System Dependability*.
- Becker, C., Duretec, K., Petrov, P., Faria, L., Ferreira, M., and Ramalho, J. C. (2012). Preservation watch: What to monitor and how. In *International Conference on Digital Preservation (iPRES 2012)*.
- Becker, C., Faria, L., and Duretec, K. (2014). Scalable decision support for digital preservation. *OCLC Systems & Services: International Digital Library Perspectives*, 30(4):249–284.

- 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*, 10(4):133–157.
- Beer, S. (1981). *Brain of the Firm*, volume 1st ed. John Wiley & Sons.
- Braud, M., Edelstein, O., Rauch, J., Rabinovici-Cohen, S., Nagin, K., Marberg, J., Voets, D., Sanya, I., Badawy, M., Shehab, E., Randers, F., Droppert, J., and Klecha, M. (2013). Ensure: Long term digital preservation of health care, clinical trial and financial data. In *International Conference on Digital Preservation (iPRES 2013)*.
- CCSDS (2002). *Reference Model for an Open Archival Information System (OAIS)*. Consultative Committee for Space Data Systems.
- Clegg, D. and Barker, R. (1994). *Case Method Fast-Track: A RAD Approach*. Addison-Wesley.
- CRL and OCLC (2007). Trustworthy repositories audit & certification: Criteria and checklist. Technical report, Center for Research Libraries (CRL) and Online Computer Library Center (OCLC).
- Dappert, A. (2010). Deal with conflict, capture the relationship: the case of digital object properties. In *International Conference on Digital Preservation (iPRES 2010)*.
- Dappert, A. and Farquhar, A. (2009). Significance is in the eye of the stakeholder. In *Research and Advanced Technology for Digital Libraries*, volume 5714 of *Lecture Notes in Computer Science*, pages 297–308. Springer Berlin Heidelberg.
- DIN (2012). Information and documentation - criteria for trustworthy digital archives. DIN 31644, Deutsches Institut für Normung (DIN).
- Duncan, A., Jones, C., Wilson, A., Palmer, W., and Proell, S. (2014). Research data sets executable workflows for large-scale execution. Technical report, SCAPE Project Deliverable (D17.2).
- Duretec, K. (2014). Final version of the simulation environment. Technical Report D12.3, SCAPE project.
- Duretec, K., Kulmukhametov, A., Kraxner, M., Plangg, M., Becker, C., and Faria, L. (2014). The scape preservation lifecycle. In *Digital Libraries conference (DL2014)*.
- Faria, L., Akbik, A., Sierman, B., Ras, M., Ferreira, M., and Ramalho, J. C. (2013a). Automatic preservation watch using information extraction on the web : a case study on semantic extraction of natural language for digital preservation. In *International Conference on Digital Preservation (iPRES 2013)*.
- Faria, L., Becker, C., Duretec, K., Ferreira, M., and Ramalho, J. C. (2013b). Supporting the preservation lifecycle in repositories. In *Open Repositories*, Charlottetown, Prince Edward Island, Canada.
- Faria, L., Duretec, K., Kulmukhametov, A., Moldrup-Dalum, P., Mejkoune, L., Pop, R., Barton, S., and Akbik, A. (2014). Scape: final version of the preservation watch component. Technical Report D12.2, SCAPE project.
- Faria, L., Duretec, K., Petrov, P., and Becker, C. (2012a). Identification of triggers and preservation watch component architecture, subcomponents and data model. Deliverable D12.1, SCAPE project.



- Faria, L., Petrov, P., Duretec, K., Becker, C., Ferreira, M., and Ramalho, J. C. (2012b). Design an architecture of a novel preservation watch system. In *International Conference on Asia-Pacific Digital Libraries (ICADL)*. Springer.
- Ferneke-Nielsen, R. B., Jurik, B. A., Andersen, B., Palmer, W., Pop, D., Schlarb, S., Duncan, A., Vujic, I., Klíma, O., Kutner, O., Parkola, T., Asseg, F., Barton, S., and Medjkoune, L. (2014). Scape final evaluation and methodology report. Technical report, SCAPE Project Deliverable (D18.2).
- Ferreira, M., Baptista, A. A., and Ramalho, J. C. (2006). A foundation for automatic digital preservation. *Ariadne*.
- Ferreira, M., Silva, H., Castro, R., Møldrup-Dalum, P., Pehlivan, Z., Wilson, C., and Schlarb, S. (2013). Gap analysis on action services tools and scape platform and testbeds requirements. Technical report, SCAPE Project Deliverable (D10.2).
- Fowler, M., Rice, D., Foemmel, M., Hieatt, E., Mee, R., and Stafford, R. (2002). *Patterns of Enterprise Application Architecture*. Addison Wesley.
- Harmsen, H., Keitel, C., Schmidt, C., Schoger, A., Schrimpf, S., Stürzlinger, M., and Wolf, S. (2013). Explanatory notes on the nestor seal for trustworthy digital archives. Technical Report 17, nestor.
- Hedstrom, M. and Lee, C. A. (2002). Significant properties of digital objects: definitions, applications, implications. In *DLM-Forum*, volume 200, pages 218–223.
- Hilse, H.-W. and Kothe, J. (2006). Implementing persistent identifiers: Overview of concepts, guidelines and recommendations. Technical report, Consortium of European Research Libraries.
- IPQ (2007). NP 4457:2007 - Portuguese standard for Research, Development and Innovation Management - Management System Requisites.
- ISO (2009a). Risk management – principles and guidelines. ISO 31000, International Organization for Standardization (ISO).
- ISO (2009b). Risk management - vocabulary. ISO Guide 73, International Organization for Standardization (ISO).
- ISO (2012). Space data and information transfer systems – audit and certification of trustworthy digital repositories. ISO 16363, International Organization for Standardization (ISO).
- Joint, N. (2008). Is digitisation the new circulation? *Library Review*, 57(2):87–95.
- Kilbride, W. (2010). Preservation planning on a spin cycle. *DPC What's New*, 28.
- Knijff, J. and Wilson, C. (2011). Evaluation of characterisation tools, part 1: Identification. Technical report, SCAPE project.
- Kraxner, M., Plangg, M., Duretec, K., Becker, C., and Faria, L. (2013). The scape planning and watch suite : supporting the preservation lifecycle in repositories. In *International Conference on Digital Preservation (iPRES 2013)*. Biblioteca Nacional de Portugal.

- Krishnamurthy, R., Li, Y., Raghavan, S., Reiss, F., Vaithyanathan, S., and Zhu, H. (2009). Systemt: a system for declarative information extraction. *ACM SIGMOD Record*, 37(4):7–13.
- Kulmukhametov, A. and Becker, C. (2014). Content profiling for preservation: Improving scale, depth and quality. In *The Emergence of Digital Libraries - Research and Practices*. International Conference on Asia-Pacific Digital Libraries, Springer International Publishing.
- Kulovits, H., Rauber, A., Brantl, M., Schoger, A., Beinert, T., and Kugler, A. (2009). From TIFF to JPEG2000? Preservation planning at the Bavarian State Library using a collection of digitized 16th century printings. *D-Lib*, 15(11/12).
- Lao, N., Mitchell, T., and Cohen, W. W. (2011). Random walk inference and learning in a large scale knowledge base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 529–539, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Lawrence, G. W., Kehoe, W. R., Rieger, O. Y., Walters, W. H., and Kenney, A. R. (2000). Risk management of digital information: A file format investigation. Technical report, Cornell University Library.
- McHugh, A., Innocenti, P., and Ross, S. (2008). Assessing risks to digital cultural heritage with DRAMBORA. In *International Documentation Committee of the International Council of Museums (CIDOC)*, Athens, Greece.
- Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.
- Mooney, R. (1999). Relational learning of pattern-match rules for information extraction. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 328–334.
- OECD (1996). *The OECD Jobs Strategy: Vol.1: Technology, Productivity and Job Creation*. Organization for Economic Co-operation and Development.
- Palmer, W., Jurik, B. A., Ferneke-Nielsen, R. B., Kutner, O., Schlarb, S., Neudecker, C., and Hahn, M. (2014a). Large scale digital repositories executable workflows for large-scale execution. Technical report, SCAPE Project Deliverable (D16.2).
- Palmer, W., Schlarb, S., Ferneke-Nielsen, R. B., Moldrup-Dalum, P., Kulmukhametov, A., and Akbik, A. (2014b). Characterisation technology, release 3 + release report. Technical Report D9.3, SCAPE project.
- Pearson, D. (2007). AONS II: continuing the trend towards preservation software 'Nirvana'. In *Proc. of IPRES 2007*.
- Pehlivan, Z., Jurik, B. A., Graf, R., Palmer, W., van der Knijff, J., Kulmukhametov, A., Medjkoune, L., and Barton, S. (2014). Quality assurance workflow, release 3 + release report. Technical report, SCAPE Project Deliverable (D11.3).

- Petrov, P. and Becker, C. (2012). Large-scale content profiling for preservation analysis. In *International Conference on Digital Preservation (iPRES 2012)*.
- PREMIS Editorial Committee (2015). Data Dictionary for Preservation Metadata: PREMIS version 3.0. Technical report, Library of Congress.
- Rosenthal, D. (2010). Format obsolescence: assessing the threat and the defenses. *Library Hi Tech*, 28(2):195–210.
- Schlarb, S., Medjkoune, L., and Palmer, W. (2014). Web content executable workflows for large-scale execution. Technical report, SCAPE Project Deliverable (D15.2).
- Schlieder, C. (2010). Digital heritage: Semantic challenges of long-term preservation. *Semantic Web*, 1(1-2):143–147.
- Schmidt, R. (2012). An architectural overview of the SCAPE preservation platform. In *International Conference on Digital Preservation (iPRES 2012)*.
- Schmidt, R. and Rella, M. (2014). Final platform release. Technical report, SCAPE Project Deliverable (D4.2).
- Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656.
- Shroeder, B. and Gibson, G. A. (2007). Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you? In *FAST '07: 5th USENIX Conference on File and Storage Technologies*.
- Sierman, B., Hofman, H., and Thaller, M. (2009). Report on the Planets Functional Model. Technical report, Planets Project Deliverable PP/7-D3+D4.
- Sierman, B., Jones, C., Bechhoffer, S., and Elstrom, G. (2013). Preservation policy levels in scape. In *International Conference on Digital Preservation (iPRES 2013)*.
- Sierman, B., Jones, C., and Elstrøm, G. (2014). Catalogue of preservation policy elements. Technical report, SCAPE Project Deliverable (D13.2).
- Silva, H. (2014). Final version of action components. Technical report, SCAPE Project Deliverable (D10.3).
- Stanescu, A. (2005). Assessing the durability of formats in a digital preservation environment: The INFORM methodology. *OCLC Systems & Services*, 21(1):61–81.
- Tarrant, D., Hitchcock, S., and Carr, L. (2011). Where the semantic web and web 2.0 meet format risk management: P2 registry. *The International Journal of Digital Curation*, 1(6):165–182.
- Thibodeau, K. (2002). Overview of technological approaches to digital preservation and challenges in coming years. In *The State of Digital Preservation: An International Perspective*. Council on Library and Information Resources.
- Wilson, A. (2008). Significant properties of digital objects. In *JISC Significant Properties Workshop*. British Library.
- Yim, J.-S., Russel, D., and Zhang, J. (2014). Market share analysis: Enterprise distributed system backup/recovery software market, worldwide, 2013. Technical report, Gartner.



# **Appendix A**

## **Appendix**

### **A.1 Survey questionnaire**

The next pages present the survey online questionnaire as exported by the LimeSurvey system.

# Digital preservation: what to monitor and how?

This survey aims to identify the most relevant digital preservation incidents, the most accepted and trusted ways to detect them and what is the current practice in terms of digital preservation monitoring.

This information will allow us to focus our developing efforts on tools that will automatically collect data on these incidents in order to provide you prompt notifications of digital preservation risks and opportunities.

This research is part of the SCAPE project, for more information visit [www.scape-project.eu](http://www.scape-project.eu) (<http://www.scape-project.eu/>).

## Why participate?

Your feedback will allow us to develop new tools to scout the Web for relevant data that will then feed into our digital preservation monitoring system (<http://openplanets.github.io/scout/>).

If you participate we will be able to better help you solve your own digital preservation monitoring issues.

## Who should participate?

From big content holders to the average person, anyone that might be interested in maintaining their digital content for dozens of years.

## Privacy and cookies

This site uses cookies to allow customisation and to collect anonymous web metrics. No other methods are used to track users and your responses will stay anonymous unless further use is explicitly permitted.

There are 30 questions in this survey

## Profile

Basic information about yourself and your organization that will help us categorize your responses.

What descriptions fit your organisation?

1

Please choose **all** that apply:

- Research funder
- National government institution
- Local government institution
- Memory institution or content holder
- University
- Big data science
- Publisher or content producer
- Digital preservation vendor
- Large enterprise
- Small or medium enterprise
- Data intensive industry
- Non affiliated
- Other:

What descriptions fit your role on your organization?

2

Please choose **all** that apply:

- Archivist
- Digital preservation manager
- Information Technology
- Technical support
- Financial manager
- Organizational manager
- Acquisition manager
- Researcher
- Salesperson
- Individual practitioner
- Other:

## Digital preservation incidents

An incident is a discrete occurrence of an event with positive or negative impact on the preservation of a digital object.

**Heads up!** Some questions are more technical than others. **If you don't feel you are the right person to answer some question, just skip to the next one!** Also, forward this survey to whom you know might know the answer to the questions you skip.

What is the importance of the following incidents (see scale below)?

3

Please choose the appropriate response for each item:

	1	2	3	4	5
<b>File corruption</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Backup failure</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Hardware no longer supported or degraded</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Software platform no longer supported or degraded</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>A relevant percentage of the producers cannot comply with the established ingest policies</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>A relevant percentage of the consumers cannot read the disseminated file format</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>There is not enough context information to understand the file content</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Content does not conform with defined institutional policies</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Organization staff is not enough or adequately trained to maintain content</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Actions executed on files (e.g. file format migration) are not having expected results</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Defined preservation plans (e.g. defined preservation format) became outdated</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

1 - Not at all important

2 - Slightly important (Informational)

3 - Important (Requires action but with low priority)

4 - Fairly important (Requires action with average priority)

5 - Very important (Requires urgent and immediate action)

Which of these incidents are you already monitoring?

4

Please choose the appropriate response for each item:

	Yes	Uncertain	No
<b>File corruption</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Backup failure</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Hardware no longer supported or degraded</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Software platform no longer supported or degraded</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>A relevant percentage of the producers cannot comply with the established ingest policies</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>A relevant percentage of the consumers cannot read the</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>



**disseminated file format**

**There is not enough context information to understand the file content**

**Content does not conform with defined institutional policies**

**Organization staff is not enough or adequately trained to maintain content**

**Actions executed on files (e.g. file format migration) are not having expected results**

**Defined preservation plans (e.g. defined preservation format) became outdated**

Are there any other digital preservation incidents you find important?

5

Please write your answer here:

## Detecting and monitoring preservation incidents

Manual or automatic ways to detect if an incident has occurred or will occur in the near future.

For every question, **vote from 1 to 7** on the more preferable way to detect the incident:

- 1 - Completely disagree
- 2 - Disagree
- 3 - Somewhat disagree
- 4 - Neither agree nor disagree
- 5 - Somewhat agree
- 6 - Agree
- 7 - Completely agree

Also, **at the same time**, state your current practice by defining if you currently **use** the method,

**don't use** it or are **uncertain** if it is used. Finally, add **other ways to detect** that the incident has occurred or is about to occur in the free text field.

**Heads up!** Some questions are more technical than others. **If you don't feel you are the right person to answer some question, just skip to the next one!** Also, forward this survey to whom you know might know the answer to the questions you skip.

File corruption

6

Please choose the appropriate response for each item:

	1	2	3	4	5	6	7	Use	Don't use	Uncertain
<b>Check files manually</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Automatic file fixity checks</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Automatic file fixity check is a program that verifies if the file bytes have changed by processing the file and comparing with existing technical/preservation metadata.

Other ways to detect file corruption?

7

Please write your answer here:

Backup failure

8

Please choose the appropriate response for each item:

	1	2	3	4	5	6	7	Use	Don't use	Uncertain
<b>Manual verification of backup success</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Alerted by backup program when failure occurs</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Notified by backup program on every execution</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Third-party program monitors correct functioning of backups</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Backup is the copy of a file or other item of data made in case the original is lost or damaged.

Other ways to detect backup failure?

9

Please write your answer here:

### Hardware no longer supported or degraded

10

Please choose the appropriate response for each item:

	1	2	3	4	5	6	7	Use	Don't use	Uncertain
<b>Manual analysis of the hardware inventory by IT staff</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Manual analysis of hardware inventory by preservation experts</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Have a close relationship with hardware vendors</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Automatic cross-reference of hardware inventory with a known hardware issues registry</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Hardware like media reader, storage, network or processing components.

### Other ways to detect that hardware is no longer supported or degraded?

11

Please write your answer here:

### Software no longer supported or degraded

12

Please choose the appropriate response for each item:

	1	2	3	4	5	6	7	Use	Don't use	Uncertain
<b>Manual analysis of software inventory by IT staff</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Manual analysis of software inventory by preservation experts</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Subscribe relevant mailing lists or other information channels</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Automatic cross-reference of software inventory with known software issues registry</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Software as the operative system, application server or digital repository system.



- Automatic cross-reference of file formats in your collections with tools that can render formats ○○○○○○○○○ ○ ○
- Automatic cross-reference of file formats in your collections with information available on format registries ○○○○○○○○○ ○ ○
- Automatic cross-reference of file formats in your collections with known file format issues ○○○○○○○○○ ○ ○
- Automatic analysis of consumer trends by inspection of consumer used software (e.g. browsers and operative systems) ○○○○○○○○○ ○ ○

Other ways to detect that a relevant percentage of the consumers cannot read the disseminated file format? 17

Please write your answer here:

There is not enough context information to understand the file content 18

Please choose the appropriate response for each item:

- |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Use | Don't<br>use | Uncertain |
|--|---|---|---|---|---|---|---|-----|--------------|-----------|
| Manual inspection of content on ingest                                       | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○   | ○            | ○         |
| User feedback  | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○   | ○            | ○         |
| Automatic verification that external references still exist (e.g. web sites) | ○ | ○ | ○ | ○ | ○ | ○ | ○ | ○   | ○            | ○         |

Other ways to detect that there is not enough context information to understand the file content? 19

Please write your answer here:

Content does not conform with defined institutional policies 20

Please choose the appropriate response for each item:

	1	2	3	4	5	6	7	Use	Don't use	Uncertain
<b>Manually verify content on ingest and whenever applicable policies change</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Define control policies in a machine readable format and have tools that automatically check the conformance</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Other ways to detect that content does not conform to defined institutional policies? 21

Please write your answer here:

Organization staff is not enough or adequately trained to maintain content 22

Please choose the appropriate response for each item:

	1	2	3	4	5	6	7	Use	Don't use	Uncertain
<b>Manually evaluate staff performance and quality</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Have automatic indicators of staff performance (e.g. objects ingested, described, words written)</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Have consumer feedback on the quality of content description and cataloguing</b>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Other ways to detect that organization staff is not enough or adequately trained to maintain content? 23

Please write your answer here:

Actions executed on files (e.g. file format migration) are not having expected results 24

Please choose the appropriate response for each item:

	1	2	3	4	5	6	7	Use	Don't use	Uncertain
--	---	---	---	---	---	---	---	-----	-----------	-----------



- Name
- Email
- Institution

Information given will be private and only used to contact you back.

Would you like to run our tools to know your file format distribution, along with other content characteristics, and compare it with others? **29**

Please choose **only one** of the following:

- Yes
- No

Would you like to participate in workshops, virtual or presencial, to know how to use the digital preservation tools created in the SCAPE project? **30**

Please choose **only one** of the following:

- Yes
- No

Thank you very much for your participation. If you chose to be contacted you will be hearing from us soon! If not, stay tuned for the developments at the SCAPE site (<http://www.scape-project.eu/>).

In the meanwhile, feel free to forward this consultation to your contacts that might be interested in digital preservation.

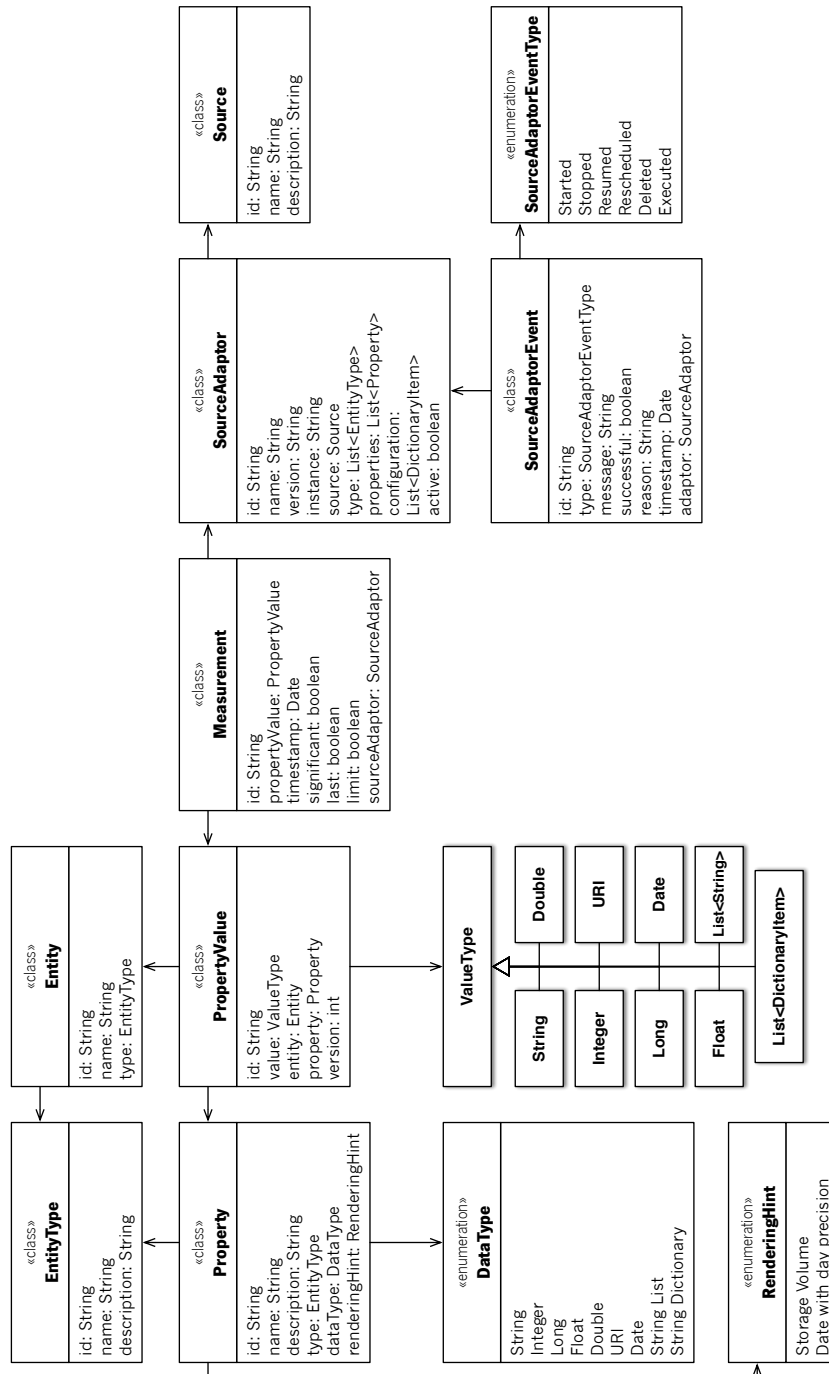
Please fax your completed survey to: +351253067248

Submit your survey.

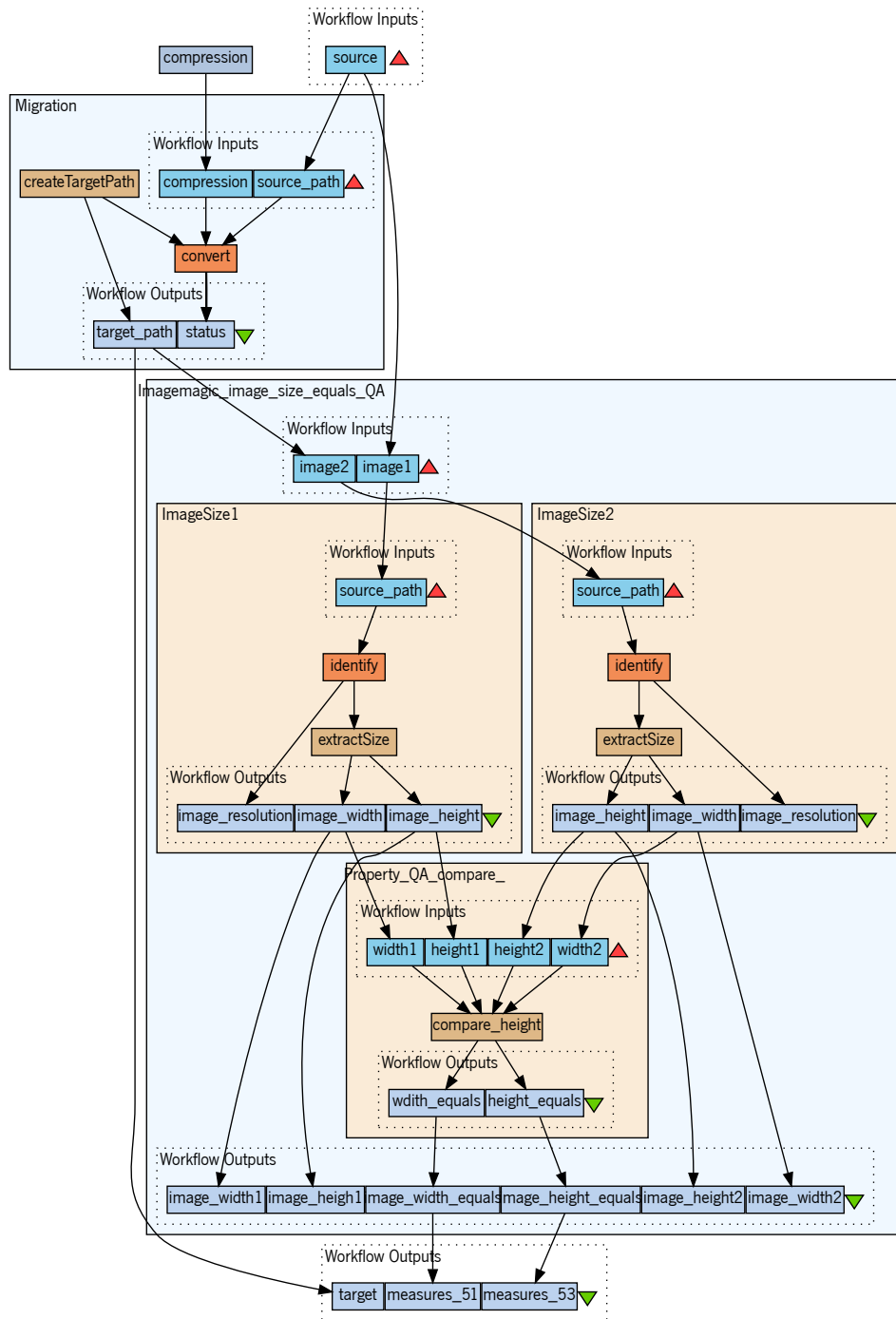
Thank you for completing this survey.



## A.2 Scout knowledge base complete UML class diagram



### A.3 Taverna workflow for a preservation action plan



## A.4 Report API output examples

Listing A.1: Report API identify

```

<?xml version="1.0" encoding="UTF-8" ?>
<OAI-PMH xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns="http://www.openarchives.org/OAI/2.0/"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2015-08-25T18:28:50Z</responseDate>
  <request verb="Identify">http://localhost:8180/roda-core/report</request>
  <Identify>
    <repositoryName>RODA</repositoryName>
    <baseURL>http://localhost:8180/roda-core/report</baseURL>
    <protocolVersion>2.0</protocolVersion>
    <adminEmail>admin@keep.pt</adminEmail>
    <earliestDatestamp>1900-01-01T00:00:00Z</earliestDatestamp>
    <deletedRecord>no</deletedRecord>
    <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
    <compression>gzip</compression>
    <compression>deflate</compression>
    <description>
      <toolkit xmlns="http://oai.dlib.vt.edu/OAI/metadata/toolkit"
        xsi:schemaLocation="http://oai.dlib.vt.edu/OAI/metadata/toolkit http://alcme.oclc.org/oaicat/toolkit.xsd">
        <title>OCLC's OAI Cat Repository Framework</title>
        <author>
          <name>Jeffrey A. Young</name>
          <email>jyoung@oclc.org</email>
          <institution>OCLC</institution>
        </author>
        <version>1.5.61</version>
        <toolkitIcon>http://alcme.oclc.org/oaicat/oaicat_icon.gif</toolkitIcon>
        <URL>http://www.oclc.org/research/software/oai/cat.shtm</URL>
      </toolkit>
    </description><software name='RODA' version='1.1.0'/'></Identify>
  </OAI-PMH>

```

## Listing A.2: Report API list of PREMIS events

```

<?xml version="1.0" encoding="UTF-8" ?>
<OAI-PMH xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.openarchives.org/OAI/2.0/"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2015-08-25T18:34:22Z</responseDate>
  <request metadataPrefix="premis-event-v2" verb="ListRecords">http://localhost:8180/roda-core/report</request>
  <ListRecords>
    <record>
      <header>
        <identifier>PlanExecuted:roda:14</identifier>
        <timestamp>2014-09-04T14:10:51.71Z</timestamp>
        <setSpec>PlanExecuted</setSpec>
      </header>
      <metadata>
        <event xmlns:ns2="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="info:lc/xmlns/premis-v2" xsi:schemaLocation="info:lc/xmlns/premis-v2 http://www.loc.gov/standards/premis/v2/premis.xsd">
          <eventIdentifier>
            <eventIdentifierType>Event ID</eventIdentifierType>
            <eventIdentifierValue>PlanExecuted:roda:14</eventIdentifierValue>
          </eventIdentifier>
          <eventType>PlanExecuted</eventType>
          <eventDateTime>2014-09-04T14:10:51.71Z</eventDateTime>
          <eventOutcomeInformation>
            <eventOutcome>success</eventOutcome>
            <eventOutcomeDetail>
              <eventOutcomeDetailNote>detected properties</eventOutcomeDetailNote>
              <eventOutcomeDetailExtension>
                <actionPlanReport plan="d8395fa1-8e43-416b-9e59-f8db9ca93df9">
                  <qualityReport>
                    <file id="F0">
                      <measure label="image distance SSIM" url="http://purl.org/DP/quality/measures#1">0.998432</measure>
                      <measure label="image width equal" url="http://purl.org/DP/quality/measures#51">true</measure>
                      <measure label="image height equal" url="http://purl.org/DP/quality/measures#53">true</measure>
                    </file>
                  </qualityReport>
                  <warningsAndErrors/>
                </actionPlanReport>
              </eventOutcomeDetailExtension>
            </eventOutcomeDetail>
          </eventOutcomeInformation>
          <linkingAgentIdentifier>
            <linkingAgentIdentifierType>RODAObjectPID</linkingAgentIdentifierType>
            <linkingAgentIdentifierValue>roda:13</linkingAgentIdentifierValue>
            <linkingAgentRole>preservation task</linkingAgentRole>
          </linkingAgentIdentifier>
          <linkingObjectIdentifier>
            <linkingObjectIdentifierType>Plan ID</linkingObjectIdentifierType>
            <linkingObjectIdentifierValue>roda:13</linkingObjectIdentifierValue>
          </linkingObjectIdentifier>
          <linkingObjectIdentifier>
            <linkingObjectIdentifierType>Object ID</linkingObjectIdentifierType>
            <linkingObjectIdentifierValue>roda:11</linkingObjectIdentifierValue>
          </linkingObjectIdentifier>
        </event>
      </metadata>
      <about>
        <agent xmlns:ns2="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="info:lc/xmlns/premis-v2" xsi:schemaLocation="info:lc/xmlns/premis-v2 http://www.loc.gov/standards/premis/v2/premis.xsd">
          <agentIdentifier>
            <agentIdentifierType>RODAObjectPID</agentIdentifierType>
            <agentIdentifierValue>roda:13</agentIdentifierValue>
          </agentIdentifier>
          <agentName>DATA CONNECTOR</agentName>
          <agentType>software:connector</agentType>
        </agent>
      </about>
    </record>
  </ListRecords>
</OAI-PMH>

```