SCAPE

Scalable Preservation Environments

Report on future research challenges

Authors

Kresimir Duretec (Vienna University of Technology), Brian Matthews, Catherine Jones (Science and Technology Facilities Council), Sean Bechhofer (University of Manchester), Rainer Schmidt (AIT Austrian Institute of Technology)

September 2014

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Executive Summary

This report is the final deliverable in the XA.WP.3 Open Research Challenges work package.
It concisely presents the major SCAPE contributions to the digital preservation research landscape. These contributions come from the domain of automation and scalability and can be categorized into the two high level research questions:

- *How do we enable the storing and processing of large amounts of potentially large, complex or varied digital objects?*
- *How do we support the alignment of digital content with a constantly changing socio-technical environment?*

The document furthermore gives a concise report on the Open Research Challenges workshop held at the IPRES 2013 in Lisbon where the discussion was centred on seven research challenges submitted from the community.
Finally a list of nine research challenges is presented. These challenges cover a range of diverse topics from modelling, simulation and benchmarking to some specific challenges rising when dealing with scientific data.
The identified challenges are:

- Information modelling and benchmarking
- Modelling and simulation of technology landscape
- Collaborative preservation infrastructures
- Building a Preservation Infrastructure for research data
- Component Based Workflows
- Research Objects for Workflow Preservation
- Provenance
- Preservation Analysis
- Preserving Context in Preservation Infrastructure

Table of Contents

1 Introduction

The field of digital preservation has received special attention over the last decade in the EU research landscape. This has resulted in a number of research projects addressing different aspects of digital preservation, such as costs[1], web archiving[2] , process preservation[3], workflow preservation[4], and audio and video preservation[5] .

The huge increase in digital content has led the community to identify automation and scalability as important challenges. Therefore, the SCAPE project had a special focus on advancing digital preservation systems and tools in order to make them more scalable and adaptive to increasing amounts of digital content. Advancements have been made in several tools and also in an ecosystem which brings together two dimensions of scalability - decision and control, and storing and processing of large amounts of digital objects - into a one system.

The main goal of this document is to describe a number of research challenges in the digital preservation field which could bring further improvements and new capabilities to the field when properly addressed.

As the process of continuous improvement should never stop, from the implemented solutions and current state in the digital preservation field new research challenges and opportunities are rising. The community understands that more rigorous and repeatable methods are needed to enable validating and improving critical preservation processes and tools. Different aspects of scalability will continue to raise new challenges and will require continuous improvement of current platforms. Furthermore, due to the exponential increase of digital data, we are already witnessing a steady migration of solutions to more distributed systems which has resulted in several initiatives to establish cross-country data infrastructures.

This is especially relevant in the scientific community. Current instruments such as particle colliders and telescopes are producing huge amounts (often measured in TB) of digital data on a daily basis. This brings new challenges in domains which were mostly considered as "solved" (e.g. bit preservation). Even though highly relevant, bit preservation can be seen only as a tip of an iceberg when it comes to the preservation of research data. There is a broader context, such as experimental set-up, that needs to be captured and preserved together with produced data. Thus the problem can be reformulated from how to preserve the scientific data to how to preserve scientific memory.

This document is structured as follows. Chapter 2 provides a short overview of the research contributions from the SCAPE project focused around two top level research questions. Chapter 3 gives a brief report on the Open Research Challenges workshop and the papers presented there.

Finally the Chapter 4 identifies nine open research challenges. Some of the challenges are a direct result of the work done in the SCAPE project and some of them are derived from considering a broader scope outside of the project.

Addressing these challenges could bring new benefits to the community and in the end make information stored in a digital form more secure.

[1] e.g. http://www.4cproject.eu/

[2] e.g. http://liwa-project.eu/

[3] e.g. http://timbusproject.net/

[4] e.g. http://www.wf4ever-project.org/

[5] e.g. http://www.prestoprime.org/

2 SCAPE research contributions

2.1 Top level challenges

The growing amount of digital content, its heterogeneity and complexity have led the community over the last few years to highlight automation and scalability as one of the key research challenges to be tackled in digital preservation.

DPE project[6] recognized the lack of systematic approaches to digital preservation and the need for efficient and cost effective models which are able to scale to different types of digital objects and organizational environments (DigitalPreservationEurope, June 2006). Furthermore, automation of key preservation processes such as characterization and migration was seen as a necessity in order to enable coping with growing amounts of digital content.

In the workshop organized by the EU commission[7] scalability and automation were recognized as important challenges where participants have pointed out the need to make already developed solutions more scalable. Also they have addressed other fields such as machine learning, pattern recognition and rule based system as potential areas for providing new techniques for automating digital preservation systems (European Commision Information Society and Media Directorate-General, 2011) (Strodl, Petrov, & Rauber, 2011).

Finally a seminar with the topic of automation in digital preservation was held in Dagstuhl[8] in July 2010. The topics covered automation outside of aspects such as number and size of digital objects. For example, evaluation and benchmarking in digital preservation poses challenges on providing means to automatically generate annotated data sets for testing digital preservation tools (Chanod J.-P. , Dobreva, Rauber, Ross, & Casarosa, 2010).

The sheer broadness of the challenge requires development of new techniques for addressing automation and scalability from processing digital objects, decision and control to more advanced topics such as automated benchmarking.

In the context of SCAPE, two top level automation and scalability challenges can be distinguished:

1. *How do we enable the storing and processing of large amounts of potentially large, complex or varied digital objects?*
   The sheer size, number and complexity of digital objects call for scalable solutions which will enable storing and processing digital objects in reasonable amount of time. An efficient computation model is needed to be connected with a digital object repository for supporting efficient ingest, access and processing of digital material. Furthermore preservation components (characterization, migration and quality assurance) need to be adapted in order to efficiently run on the selected model.
2. *How do we support the alignment of digital content with a constantly changing socio-technical environment?*
   The continuous changes in socio-technical environment (formats, tools, user needs ...) bring the need for automated means which will support a continuous controlling and decision making process in order to keep digital material aligned with its environment. Scalable solutions for collection profiling, preservation monitoring and planning are required in order to analyse digital collections, link them with potential risks and opportunities from the

[6] http://www.digitalpreservationeurope.eu/

[7] http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-future-digital-preservation_en.html

[8] http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=10291

environment and finally enable making a trustworthy decision to address those risks and opportunities.

Three major areas in SCAPE (scalable platform, scalable planning and watch, and scalable components) have addressed these defined challenges by providing several concepts, models, algorithms and tools. As the scalable platform and scalable components address the first challenge and the scalable planning and watch the second challenge, only an integrated system which combines results from those three areas will be able to provide a fully scalable solution which is able to cope with the two defined challenges in an efficient way.

2.2 Scalable Platform

The scalable platform work has addressed several challenges around questions including:

- *What is the architecture of a scalable preservation platform and which computation model is the most suited for the preservation needs?*
- *How to integrate repositories (such as RODA[9] or Fedora[10]) with such platform?*
- *How do we enable discovery and execution of preservation workflows and tools developed by third parties on such a platform?*

With respect to architecture, the implemented platform solution is based on a MapReduce (Dean & Ghemawat, 2008) computation model where Apache Hadoop [11] and technologies built around it have been taken as an open source implementation of this model. The SCAPE Platform (Schmidt, 2012) integrates and extends existing software platforms like the Fedora digital object repository and the Taverna workflow stack[12]. Integration between the platform and a digital object repository is done via dedicated APIs[13] and enables efficient methods for ingest and access to digital content.
The Taverna workflow stack together with its associated publishing platform, myExperiment[14], has been adapted to enabled publishing and reuse of different preservation components. An ontology was developed to support semantic annotations of those components which provide third party applications, such as preservation planning tool Plato, the ability to automatically discover and use needed components[15]. To enable execution of those workflows in large scale scenarios, methods have been developed to translate them into a form more suitable for the MapReduce model (Akbik, 2014). A major challenge in the context of the Platform's execution environment was the incorporation of the third party software components, many preservation scenarios rely on. As a result, ToMaR[16] (Schmidt, Rella, & Schlarb, 2014), a generic application wrapper, was developed to enable the efficient execution of legacy tools in massively parallel environments such as Apache Hadoop.

[9] http://www.roda-community.org/
[10] http://www.fedora-commons.org/
[11] http://hadoop.apache.org/
[12] http://www.taverna.org.uk/
[13] https://github.com/openplanets/scape-apis
[14] http://www.myexperiment.org/
[15] http://openplanets.github.io/scape-component-profiles/
[16] https://github.com/openplanets/ToMaR

2.3 Scalable Planning and Watch

Scalable planning and watch has addressed several challenges around questions including:

- *How to automatically collect and analyse preservation information from different sources and provide an automatic notification on important events?*
- *How to enable scalable content profiling?*
- *How to enable efficient creation of trustworthy preservation plans?*
- *How to put preservation planning, monitoring and operations in context with institutional policies?*

To monitor socio-technical changes in the environment Scout[17] , a preservation watch service, was developed. It provides an adaptable architecture based on a plugin mechanism in order to gather information from different sources, a data model based on linked data principles to enable linking collected information and a notification mechanism which detects important events and sends a notification to an interested user (Faria, et al., 2014) (Becker, Faria, & Duretec, 2014) .

One source of information for Scout may come from systematic analysis and monitoring of digital objects in a repository. Such analysis is an important aspect of preservation monitoring and a key step for successful preservation planning. The C3PO tool[18] provides a content profiling mechanism which consists of three high-level steps: gathering metadata, processing and aggregation, and meta-data analysis (Petrov & Becker, 2012) (Becker, Faria, & Duretec, Scalable Decision Support for Digital Preservation, 2014). To cope with the scalability challenges, which arise from the fact that metadata can be a large collection in terms of the size and number of objects, the tool is also based on the scalable MapReduce model implemented by the MongoDB[19] database.

In the preservation planning domain, work continued on improving the preservation planning tool Plato[20], where the focus was on how to automate the fourteen workflow steps provided by Plato in order to improve the efficiency of creating a trustworthy preservation plan. Improvements were made in several workflow steps. Content analysis, enabled by C3PO, eliminates manual sample description and representative file selection. Experimentation environment based on Taverna enables automated software testing and can minimize the need for manual input. Integration with control policies enables reuse of already defined objectives for digital collections (Kulovits, Becker, & Andersen, 2013) .

Successful preservation is enabled only when preservation planning, monitoring and operations are put in context with institutional preservation policies. Often, in digital preservation, those policies are expressed as mission statements in high-level strategic documents which make it a challenging task to align preservation planning, monitoring and operations with them. In order to facilitate automation of digital preservation processes while keeping them aligned with high-level statements a three level model for expressing digital preservation policies has been developed (Sierman, Jones, Bechhofer, & Elstrom, 2013) . Those three levels reflect different levels of policy in an organization: from top strategic levels to bottom control (operation) levels. To make the control level understandable to the software components such as preservation planning tool Plato and preservation watch component Scout, an ontology was created which enables the definition of a machine understandable policy model (Kulovits, Kraxner, Plangg, Becker, & Bechhofer, 2013) (Bechhofer, Sierman, Jones, Elstrøm, Kulovits, & Becker, 2013). This ontology is accompanied with a

[17] http://scout.openplanetsfoundation.org/
[18] http://c3po.openplanetsfoundation.org/
[19] http://www.mongodb.org/
[20] http://ifs.tuwien.ac.at/dp/plato/intro/

controlled vocabulary[21] which uniquely defines key elements for building policy models. The policy elements catalogue (Sierman, Jones, & Elstrøm, 2014) provides guidance and often used policy elements to support more efficient definition and translation of high level policies to the control level.

2.4 Scalable Preservation Components

The scalable preservation components work has addressed challenges around questions including:

- *How to ensure large scale applicability of preservation components?*
- *What are the mechanisms to reduce high amount of conflicting results from characterization components?*
- *How to enable accurate web page comparisons?*
- *How to support different aspects of quality assurance in digital image, audio and video collections?*
- *Can a machine replace human judgement in terms of the quality assurance?*

A number of different characterization, migration and quality assurance tools were either identified or developed and adapted to run on a Hadoop platform and published as discoverable workflows on the myExperiment platform. To automate the publishing process a special tool wrapper[22] was developed which, based on a single specification, publishes a component as a Debian package (for installation) and associated Taverna workflow (for discovery and execution). Through this, tools can easily be discovered and combined into Preservation Workflows for execution on a Hadoop cluster; equally, the underlying tools can easily be deployed.
When creating a collection profile (e.g. with C3PO) based on outputs from more characterization tools (e.g., using the FITS tool[23]) a number of conflicts (i.e., the case when two characterization tools provide different value for the same property) can appear. Those conflicts severely affect the quality of results as a significant part of a collection can be marked as conflicted. To cope with this problem (Kulmukhametov & Becker, 2014) proposed a framework based on a rules mechanism which enables resolving conflicts in an automated way. The results have shown that significant improvements can be achieved only by addressing conflicts caused by different output conventions of different tools.
Quality assurance is an important aspect of digital preservation and the SCAPE project has addressed it in several domains: web pages, images and audio. Taking each of these in turn:
For web page preservation it is important to monitor how a page's renderability changes in different browsers over time. To classify a pair of web pages as "similar" or "not similar", the SCAPE developed Pagelyzer tool[24] uses visual and structural features from those web pages as an input for a support vector machine classifier (Pehlivan, et al., 2013).
Often during digitization, unwanted effects such as having duplicate scanned images or overlapping audio sequences in broadcast recordings can occur. Matchbox [25] calculates a similarity estimation between two images based on image fingerprints. Those similarity estimations enable it to detect duplicate images within an image collection. Similarly for sound, xcorrSound[26] detects overlaps in

[21] http://purl.org/DP/quality
[22] http://openplanets.github.io/scape-toolwrapper/
[23] http://projects.iq.harvard.edu/fits
[24] http://openplanets.github.io/pagelyzer/
[25] http://openplanets.github.io/matchbox/
[26] http://openplanets.github.com/scape-xcorrsound

audio files by cross-correlating their waveforms. Furthermore it can compare the whole waveform in order to calculate the similarity between two audio files.

An important question in the quality assurance domain is the accuracy of given tools and if they can actually substitute for human judgement. A number of different experiments have been conducted in order to show the accuracy of developed tools. (Pehlivan, et al., 2013). While promising results are achieved it is clear that more detailed experiments are needed. This will require development of richer datasets with proper annotations (Kulmukhametov, Plangg, & Becker, 2014).

2.5 The SCAPE Ecosystem

In order to have a fully scalable system for digital preservation both the top level challenges (defined in the Chapter 2.1) need to be covered. All parts – the platform, planning and watch, and the preservation components – must be integrated into one working system in which the need for human intervention should be minimized. And yet, individual use of single parts of this system must also be achievable; there may be organizations which would benefit from using only specific tools, for example. A set of APIs was defined which enables loose-coupling between these different parts. They enable the connection of the preservation planning tool, Plato, and the preservation watch service, Scout, to a digital object repository, and connecting that repository to the execution environment. These connections enable a full lifecycle support for digital objects. The lifecycle supports, in a fully automated way, the monitoring of the object's environment, notifying and reacting (by creating a preservation plan) on and to risks and opportunities, and execution of an operation which will address those risks and opportunities (Becker, Faria, & Duretec, 2014).

3 Research challenges from the community

Following the highly successful workshop on Open Research Challenges in Digital Preservation held at the IPRES 2012[27] another workshop with the same topic was organized at the IPRES 2013[28]. The organizers of the workshop were Christoph Becker (University of Toronto and Vienna University of Technology), Andreas Rauber (Vienna University of Technology) and Christopher (Cal) Lee (University of North Carolina at Chapel Hill). The goal of the workshop was to elicit and discuss the Digital Preservation research challenges to be tackled in the next decade.

Seven papers were chosen and presented in the workshop.

- Functional Long Term archiving Group (presented by Dirk Suchodeletz): **Challenges of Remote Access to Emulation of Original Environments**[29]
- Functional Long Term archiving Group (presented by Dirk Suchodeletz): **Challenges to Identify Software for Reproduction of Complex Environments**[30]
- Functional Long Term archiving Group (presented by Dirk Suchodeletz): **Migration of Complex Original Environments – Verification and Quality Assurance Challenges**[31]

[27] an exstensive report can be found in (Becker, Rauber, Paton, Schmidt, Milic-Frayling, & Matthews, 2012)

[28] http://digitalpreservationchallenges.wordpress.com/

[29] http://digitalpreservationchallenges.files.wordpress.com/2013/09/2013freiburg1.pdf

[30] http://digitalpreservationchallenges.files.wordpress.com/2013/09/2013freiburg2.pdf

[31] http://digitalpreservationchallenges.files.wordpress.com/2013/09/2013freiburg3.pdf

- Alexander Schindler and Reinhold Huber-Mörk: **Towards Objective Quality Assessment in Digital Collections**[32]
- Mohammad Raza and Natasa Milic-Frayling : **Leveraging Human Intelligence: Semi – automated Processing in Assuring Access to Digital Content**[33]
- Natasa Milic-Frayling: **Sustainable Computation – Foundation for Long Term Access to Digital**[34]
- Artur Kulmukhametov: **Digital Preservation as a Science**[35]

The seven papers presented have identified challenges which can be grouped around higher level topics:

- enabling new more sustainable models of preserving digital information by considering computation environments
- advancements in migration and quality assurance of digital files and complex software environments
- establishing digital preservation field as a science

There is a rising awareness that to get the maximum value from stored information the associated computation environment is as important as the digital object containing the information. In the workshop three approaches were presented: offering Emulation as a Service (EaaS), establishing software archives and incorporating digital preservation needs into the software development process.

To remove complexities of the emulation process and to make it a more inviting alternative an approach of offering emulation as a cloud based service (EaaS) was explored. However, to successfully implement such service, the quality requirements need to be fully understood. In order to provide a high-quality experience, delivered content (audio and video) will need to be synchronized. In scenarios when content is composed of audio and video streams synchronization might be highly unpredictable due to a number of different buffers which will be present in networked systems used to offer EaaS. Already existing solutions (streaming protocols) are not powerful enough for such complex demands and will need to be further extended or new protocols will need to be developed. Interactivity of emulation environments is another important quality. When those environments are offered as remote services, achieving good interactivity levels will be a challenging task mainly because of the limited bandwidth and latency of the network to be used.

Establishing software archives which would be responsible for preserving computation environments is another approach. Several organizational and technical challenges rise. The main challenge is to identify a proper workflow that those archives should follow when preserving software. Once software is being preserved it will need to be properly described and identified. There is also a need for an analysis of the possible levels on which those archives should run (institutional, national or global) and the type of an archive (central or decentralized). Finally, there is a question whether available licensing options are sufficient for such initiatives or new licensing mechanisms will need to be established.

A challenge was raised to incorporate preservation of digital information and its value into the software development process. In order to enable computation by which digital assets can be used in contemporary environment three approaches are identified: *encapsulation of the original*

[32] http://digitalpreservationchallenges.files.wordpress.com/2013/09/2013schindler.pdf

[33] http://digitalpreservationchallenges.files.wordpress.com/2013/09/2013raza.pdf

[34] http://digitalpreservationchallenges.files.wordpress.com/2013/09/2013milicfrayling.pdf

[35] http://digitalpreservationchallenges.files.wordpress.com/2013/09/2013kulmukhametov.pdf

computation in a virtual machine; porting of the software application to a new version; replacement of a software application by a contemporary application and conversion of the data files into the file format supported by the new application. All three approaches require additional effort usually in a form of software development. Thus this makes preservation dependent on available development skills and available economic models which will support the incurred costs. An approach to this problem is to introduce new properties in software quality models which will refer to the end-of-life of computing systems. Those should be mainly focused on minimizing the expected effort and cost of sustaining digital assets which are produced by a specific system.

Migration and quality assurance are topics often covered in the digital preservation research. Still new research challenges are being proposed to improve the state of the migration and quality assurance tools.
Mark-up languages such as XML are frequently used for storing data in different applications. To be able to preserve such data, methods which can translate from one format to another are needed. A research roadmap based on learning format transformations from human interactions was proposed. The main contribution of the research should be a domain specific language (DSL) which supports the definition of a transformation and an algorithm which can automatically derive transformations. With this approach the main challenge is to develop a proper DSL which should be expressive enough to be able to cover a range of scenarios. As the DSL expressivity raises though, so do computational costs for inferring needed transformations. This will obviously require finding a good balance between the expressivity and the computational costs of such approach.
Quality assurance is an essential process during format migration. Current metrics used are appropriate to detect deviations after lossless migration but unable to do so when a lossy migration is used. As lossy compression is quite often used it is necessary to develop metrics which will be able to measure information loss in such scenarios. A research to identify and apply metrics which include perceptual features is proposed.
Emulation can be considered as a form of migration which takes place on the level of applications, operating systems or even hardware. As it might result in a change of rendering outcome, it requires quality assurance which would prove environment reproduction satisfies certain authenticity requirements. The main challenge is to find a broad enough test which could perform such an operation in an automated way. Another approach could be to formalize such system migrations in order to enable achieving verifiable results.

Finally, there is a need to reconsider the whole digital preservation field and make it more scientific. Current practices, based mainly on empirical knowledge, need to be improved in several ways. Proper metrics, which will enable rigorous evaluation of tools, systems, processes and whole organizations, need to be established. Furthermore, different experiments, many organizations are conducting, need to be based on rigorous scientific approaches and made reproducible. This would in the end result in digital preservation being a well-structured filed where main terms, decisions and processes are proven.

4 Identified research challenges

In this chapter nine identified research challenges are described. These nine challenges were identified by the participants within the XA3 workpackage. While some of them build directly on the work done in the SCAPE project, others are derived from considering broader domains such as benchmarking or e-infrastructures. Furthermore, three of them (Preservation Analysis, Building a

Preservation Infrastructure for research data and Preserving the Context of Research in a Preservation Infrastructure) address the specifics of preserving the memory of science. Challenges are considered independently of each other and each challenge starts with a motivation followed by research questions to be tackled. Due to the different scope of each challenge, the level of detail and number of research questions in each challenge varies.

4.1 Information modelling and benchmarking

Characterization covers one of the three key preservation operations (the other two are preservation actions and quality assurance) and is often considered as the first step for the successful preservation of digital objects. The special attention to this field over the last few years has resulted in a number of different tools covering activities from identification (determining format and format version of a digital object), validation (checking the conformance of a digital object with the format specification) to general characterization (extracting different metadata from a digital object). Among others, DROID[36], Apache Tika[37], PDFBox[38], and JHove2[39] can be identified as well-known examples in the community.

Functional correctness (ISO/IEC, 2011) of such tools is recognized as a key software quality property.

Unfortunately, evaluating the functional correctness is limited by the lack of properly annotated datasets which could be considered as a ground truth to enable the results of each tool to be compared against a known result. For example, an often used corpora for benchmarking, the Govdocs corpora[40], contains only identification data which was generated using a forensics tool provided by Forensic Innovations[41]. This limits benchmarking of characterization tools in several ways. First, the provided annotations can cover a limited set of identification scenarios. Second, and most importantly, the provided annotations are produced by another tool which hasn't proven its functional correctness which makes those annotations and in the end the whole benchmarking process untrustworthy. The community recognizes this problem and understands the need for ground truth datasets which would enable proper evaluations[42].

To address this question (Becker & Duretec, Free Benchmark Corpora for Preservation Experiments:Using Model-Driven Engineering to Generate Data Sets, 2013) proposed a framework which would enable benchmarking of different characterization tools by automatically generating datasets with accompanying annotations (ground truth). The framework is divided into four main parts: *analyse*, *generate*, *evaluate* and *publish*.

The main goal of *analyse* is to analyse already existing collections in order to determine distributions of different file properties (for example understanding distribution of number of pages in government documents). This enables an overview of a real world dataset and serves as an input for generating a dataset which should represent the real world dataset as well as possible.

As the manual annotation of already existing collections is an expensive process the goal of the *generate* part is to automate the generation of digital objects and associated annotations. It is a workflow which is based on model driven engineering concepts. To separate implementation

[36] http://droid.sourceforge.net/

[37] http://tika.apache.org/

[38] https://pdfbox.apache.org/

[39] https://bitbucket.org/jhove2/main/wiki/Home

[40] http://digitalcorpora.org/corpora/files

[41] http://www.forensicinnovations.com/

[42] http://www.openplanetsfoundation.org/blogs/2014-01-27-identification-pdf-preservation-risks-analysis-govdocs-selected-corpus

independent properties (e.g. number of tables in a MS Word document) from implementation specific properties (e.g. representing a table in a MS Word document as a simple Word table or embedded MS Excel table) platform independent and platform specific models are used. A set of transformations is automatically used in order to translate from one model into another. For example, when generating MS Word documents with different tables on the platform independent level the number of tables is controlled and on the platform specific level the table implementation (simple MS Word table or embedded MS Excel table) is determined. The output of the *generate* process is a dataset with annotations for each digital object.

The *evaluate* step takes the generated dataset, the tool which is to be benchmarked and compares the tool output with the expected output which is provided by the *generate* step (document annotations). This enables the calculation of the functional correctness of the tool being benchmarked.

Finally all the data generated in those three steps are *published* and made available online.

In the article (Becker & Duretec, Free Benchmark Corpora for Preservation Experiments:Using Model-Driven Engineering to Generate Data Sets, 2013) the feasibility of approach was shown by presenting two different prototypical implementations based on different technologies (EMF[43] and XML/XSLT). Based on this preliminary work and the proposed framework a list of research questions has been derived. These should be tackled in the future in order to improve the usability and soundness of the approach.

The questions are listed for each part of the framework even though in some cases they might address the connection points between two parts.

How do we identify the sample size which is big enough to support statistics used in the generate part? What is 'big enough?

As our goal is to create a solid ground truth, we have to know what are the limits and scope of a generated dataset. Once this is done, we need to know how to pass the results of *analyse* to *generate.*

How to represent the correlations between properties of interest of a dataset and create platform statistics?

Analysing a single property of content is a straightforward task and results may be presented in a form of a histogram with a property distribution. A simple example is the *wordcount* property for MS Word 2007 documents. But how to effectively represent features when we want to test if a characterization tool correctly identifies *wordcount* in MS Word 2007 documents, which contain tables? In this case, to generate useful platform statistics, we need to consider different ways of creating tables in a Word document (a simple table, an embedded excel table, a complex table with OLE-objects). All of the peculiarities inherent to target properties should be studied and expressed in a way which would allow generation of a dataset.

What is the workflow which will provide the highest level of automation in the generate process?

A number of different benchmarking scenarios with different requirements for the ground truth and generated datasets (in terms of format and the content in a digital object) can be elicited. In order to provide automation in the generate step the main challenge is to find a generic workflow which would be applicable to all of those scenarios. It will be important to enable reuse of already created elements (models, transformations) in order to support efficient inclusion of new scenarios and reducing the effort of creating new datasets.

[43] http://www.eclipse.org/modeling/emf/

Could a search based generation strategy provide benefits over a template based generation strategy?
Search based testing is a well-established field which, due to the complex input space, uses meta-heuristic algorithms, such as genetic algorithms or simulated annealing, to generate test input data. It should be explored if such an approach could be used as an alternative to the template based (proposed in (Becker & Duretec, Free Benchmark Corpora for Preservation Experiments:Using Model-Driven Engineering to Generate Data Sets, 2013)). The main challenge would be to identify a fitness function which would drive the whole search process. Also a question is if there could be some synergy between those two approaches?

How do we introduce controlled errors (invalid states) in generated objects?
Often, when testing tools, we are interested in testing their ability to accurately detect errors or invalid states. The main challenge is to find mechanisms which will enable adding invalid states in a controlled and systematic way. Furthermore, in order to generate error, documents will need to be generated directly and not via some language or framework as this might prevent generating invalid documents. For example, when generating PDF objects, using a latex processor and tools build around that technology will potentially only generate valid PDF documents.

What are the mechanisms to enable systematic ground truth extraction and its representation?
Ground truth will be highly dependent on scenarios and thus could cover several levels of details. Therefore there should be mechanisms in place which will allow systematic extraction of the ground truth. Ground truth will also need to be recorded in some form. In order to support automation in the evaluate step an ontology should be developed which should support querying ground truth and thus enable automation of the evaluate step.

How do we ensure the correctness of the ground truth?
In some scenarios the correctness of the ground truth might be endangered because of the properties which are dependent on the rendering environment of a digital object. For example adding too large an amount of text on a single page might result in two page document. When all the combinations are considered (number and size of different elements) it might be challenging to determine the correct ground truth. Obviously there will be a threshold when certain ground truth might become invalid and the challenge will be to find that threshold. Mechanisms for measuring the probability that the ground truth is correct could be useful in such situation.

To what degree can we automate the evaluate step?
The evaluate step, in its basic form, will consist of running a tool on a specific digital object gathering its outputs, comparing them with ground truth data and providing results. In order to automate this there will need to be a mechanism which will be able to match certain tool output with a certain entry in the ground truth. Also a challenge would be the ambiguity of different properties. For example it might not always be clear what a certain tool considers under number of words in a document. Here the ontology and the measure vocabulary (Kulovits, Kraxner, Plangg, Becker, & Bechhofer, 2013) developed within the SCAPE project could provide pointers towards the solution. Also the Taverna workflow environment could be used to automate the whole process as it already offers an ontology to link tool outputs with measures from the vocabulary.

How to interpret results?
Once results are calculated the challenge will be to generalize those results. As the evaluation will be performed on an artificially created datasets the question is whether those results can be generalized to the real world datasets. Is it possible to measure the level with which a generated dataset is approximating the real world dataset and would that kind of a measure enable calculating the confidence levels of achieved results? How do we calculate benchmark coverage and can coverage be used to improve the confidence in results?

Can we automatically infer probable reasons for errors?
One purpose of benchmarking is to detect bugs and enable improvement of tested tools. In a case of a huge number of objects, providing a functional correctness measure might not be enough to detect potential bugs and cases in which errors appear. This will call for more advanced methods which can cope with a big number of high dimensional data and can group those results according to certain properties. Clustering methods might be an approach which could enable grouping files which produced errors in a way which would allow easy extraction of common features and enable inferring most probable reasons for a certain error.

Which mechanisms are needed to enable citation and search of published results?
Once a larger number of different datasets covering different scenarios have been published it might be potentially challenging to discover appropriate datasets for a given scenario. This calls for mechanisms which would be able to index datasets and automatically infer which scenarios they could cover.

Which aspects could raise problems in reproducibility of results and how to address those problems?
Reproducibility is one of the key requirements for trustworthy benchmarks and will need to be considered. By properly publishing all the results, reproducibility aspects should be improved. Still there might be some hidden challenges. For example the generate part, due to its randomness, might not be able to generate the same dataset twice. Furthermore, in the evaluate step, different results can be achieved on different machines in cases when those machines have different libraries on which a certain tool is dependent.

4.2 Modelling and simulation of technology landscape

The rapidly changing information technology (IT) environment is one of the main drivers of research and development in the digital preservation field. More precisely, format obsolescence is often pointed as a major risk for digital information which calls for effective mitigation techniques (Rothenberg, 1995) . Recently however, a question has arisen over whether format obsolescence is still a threat (Rosenthal, 2010). Rosenthal points at, as one of the arguments for the format obsolescence not being an issue any more, the maturing market of formats. The network effect makes the format market stable with low levels of innovation and thus changes which would make a format obsolete are not probable. These, obviously conflicting view points, require rigorous identification, analysis and prediction of trends in the IT environment in order to support more trustworthy conclusions about the technology change and how it affects (endangers) digital information.

Market research is a well-established field and researchers have over the last few decades developed a number of different models in order to better understand markets of different products, and to enable accurate predictions of their future trends. A well-known example is the Bass diffusion model which models a product’s diffusion in terms of innovators and imitators (Bass, 1969). Several models have built upon that model by incorporating more advanced concepts such as technology generations, social networks, cross country influences and competition (Peres, Muller, & Mahajan, 2010).

The huge amount of different technologies and complex relationships among those technologies makes the IT domain especially challenging fields for predicting future trends. (Adomavicius, Bockstedt, Gupta, & Kauffman, 2008) propose an IT ecosystem model with the main objective to provide a formal problem representation structure for the analysis of IT development trends.

The main purpose of those models is to support decision making in terms of new product development and technology investments.

Unfortunately, there have been no attempts to systematically analyse and model the technology landscape and its evolution for a better understanding of digital information endangerment.

The SCAPE project has made several advancements in characterizing digital objects and the aggregation and analysing of the produced metadata (C3PO & SCOUT). This should present a good starting point for creating such models as it offers feature rich datasets which span over a longer time period.

In the next few paragraphs several challenges around this topic are identified.

How do we model the technology landscape in order to estimate information endangerment?

There are already developed models which model the technology landscape. The question is how those models can improve our understanding of the technology markets in terms of information endangerment? The challenge is to extend those models in order to measure the information endangerment of certain objects which are dependent on a specific technology. For example simple diffusion models could show the evolution of acceptance of formats and their versions by the general population. Also, those models could be used, with a certain level of accuracy, to predict future trends such as peak usage time of a specific technology. The challenge is to establish a link between the obsolescence and usage and whether we can call certain technology obsolete just because it is not used any more, or whether some more advanced measures are needed.

Once properly established those models should provide insight into the market dynamics and how those dynamics affect information.

What kind of data is needed for created models to support accurate predictions?

In order to theorize about markets and to provide accurate predictions about the future, model parameters will need to be estimated based on already existing data. The question is which kind of dataset will be suitable for such a task. The web archives might be the best candidate as they provide datasets that are large in terms of the number of objects, heterogeneous in terms of the formats used and span over a bigger time period (usually 10 to 20 years). In order to show the generalizability of the results it will be important to show how well web archives represent the general population in terms of technology use.

What kinds of simulation models are suitable for simulating technology evolutions in order to project possible future states?

Simulation can show cost effective benefits in order to project future states of the technology landscape. The number of different approaches, such as discrete event, stochastic, agent based, etc., means effort will be required to identify the most suitable approach for simulation.

What are the reproducibility requirements and how do we support those?
Reproducibility of such work will be crucial in order to establish the trustworthiness of the achieved results. Due to the complexity of the approach, which will require mathematical computations and reliance on different software and datasets, a number of reproducibility aspects should be covered. These range from making sure there are no side effects caused by numerical computation to simply publishing all the models (represented as software) and datasets in a form which is easily citable and accessible.

4.3 Collaborative preservation infrastructures

Digital preservation environments like repositories and archiving systems are traditionally designed to operate autonomously and are self-contained. A paradigm shift towards collaborative preservation environments, however, could greatly improve technical as well as economic factors, and provide major benefits to the user community. There are successful and globally distributed systems that have been developed in sciences like High Energy Physics. There are existing distributed systems that allow the reliable storage of data across organizational boundaries. Collaborative Preservation Infrastructure might be a logical next step to enhance preservation infrastructure towards computation and user participation.
In particular, we envision improvements of state-of-the-art preservation systems based on three design goals: (1) collaborative storage and backup, (2) participatory data analysis, and (3) coordinated sharing and outsourcing of hosting infrastructure.

Digital preservation is dealing with the problem of maintaining digitally encoded information so that it remains available and understandable over very long periods of time. While having its origin in the cultural heritage domain, digital preservation targets almost all areas of society and forms of digital information. Applying a suitable preservation strategy for individual (groups of) digital items can be a technically challenging task. The required effort typically depends on the individual objects and characteristics like hardware and software dependencies, data formats, or the integrity of the object, to name a few. However, in recent years, preservation strategies and corresponding technologies have been developed that enable us to preserve a large range of digital materials. Key aspects target the prevention of data loss (e.g. using replicated storage and checksumming) as well as format/hardware obsolescence (e.g. through continuous format migration and software/hardware emulation).

Another increasingly difficult problem, digital preservation is facing, is the management of the steadily growing volumes of data. Analogous to large-scale web analytics (Dean & Ghemawat, 2008)and scientific data management systems (Hey, Tansley, & Tolle, 2009), preservation environments will have to adopt Big Data platforms in order to cope with the volumes of data that are being produced by today's society. The SCAPE project is addressing this topic by developing tools and services for the efficient planning and application of preservation strategies for large and heterogeneous data collections. The SCAPE Preservation Platform (Schmidt, 2012) developed in this context, supports the development of environments that are built upon scalable data management frameworks. We expect that solutions that are built on such data-centric infrastructures have the potential to greatly advance the capabilities of existing preservation systems with respect to robustness, throughput, and scalability. Although a scalable platform will help individual institution in managing and preserving growing amounts of data, it is questionable if individually hosted stand-alone systems will be technically as well as economically viable on the very long term.

Research infrastructures provide geographically distributed networks that enable scientific integration in various domains. Such e-infrastructures can provide scientists with global access to research facilities, data, or computational resources. While a steadily increasing number of e-infrastructures exist in the sciences, the adoption in the humanities is taking place more moderately. Although the need for research infrastructures in the humanities has been well recognized, understanding and implementing the requirements in terms of information policies, organizational and technical requirements is still a grand challenge in the field. E-infrastructures that interlink preservation archives would – besides yielding other benefits - have a great potential to provide unique, scalable, and globally operating research and knowledge sharing environments.

How do we enable collaboration in terms of storage and computation between systems?
In order to provide viable solutions on the long run, there is a fundamental change required in the way preservation environments work. At present, most archival systems are designed as autonomous systems. These systems might be capable of managing data within a distributed environment but hardly know about information that is stored within different systems in the same environment. This design philosophy however causes a number of inefficiencies which can have a significant impact on the overall scalability and cost efficiency of preservation systems. These factors are, in particular, critical if systems are operated within a large-scale environment. An example is the uncontrolled replication of content. One could consider multiple web archives which independently harvest overlapping content, which is stored in different environments, and additionally saved as a backup in different locations. From an overall perspective, such systems lead to an inefficient and expensive use of IT-resources.
Truly collaborative environments would detect replications and, for example, act as mutual backup resources, like that implemented by the LOCKSS system (Maniatis, Roussopoulos, Giuli, Rosenthal, & Baker, 2005). However, archival storage and backup provides only one aspect of collaboration. Computation provides another major aspect. Preservation environments and the institutions that operate them are typically limited in the computations they can run against the archived data. Examples are regular checksumming, identification, and migration of archived data. It will be important to provide means that allow 3rd party users to process archived content, as with a growing amount of content it will simply be impossible for a single institution to understand and curate all of the data it preserves. The major argument here is that there is no point in preserving data if it cannot be analysed and, hence, be discovered by its users.

Which legal aspects need to be clarified and which institutional policies need to be developed for data distribution, access, and community involvement?
Infrastructure providers that offer the hosting of applications and data within globally operating data centers (called clouds) have gained major attention in recent years. The model has turned out to be very cost efficient taking advantage of the economy of scales. It is obvious that institutionally hosted data repositories can only be operated for very limited (handpicked) data sets without hitting an economic barrier. Cloud offerings, on the other hand will usually not meet institutional policies, hindering one to outsource the housing of the data and systems. The same policies will most likely also prevent collaborative storage and computation.
The need for developing an international framework for a *Collaborative Data Infrastructure* has been described by the High-Level Expert Group on Scientific Data (High Level Expert Group on Scientific Data, 2011). The report suggests a conceptual framework for a *Global Collaborative Data and*

Knowledge Infrastructure. The DCH-RP project[44] has stated plans to investigate this model with respect to its applicability to distributed digital preservation.
We therefore see an important research challenge in the development of methods that foster and promote the development and deployment of collaborative, scaling, and globally operating preservation environments.

Which existing systems, frameworks, or platforms provide a sufficiently mature and sustainable technology to implement CPI, and which are the building blocks which cannot be implemented using existing technology?
Grid Computing is a discipline that deals with the coordinated sharing of computer resources over multiple administrative domains. While this concept - in contrast to the cloud model - did not prevail as a general purpose architecture, it has been very successful in providing collaborative infrastructures for large-scale data management and computation in specific domains. One of the most prominent examples is provided by CERN's LHC Computing Grid. An interesting research question is to study the feasibility of building a smaller but similar structured research infrastructure for the purpose of digital preservation and research. The concept of tiers might, for example, be well applicable in building different trust domains of a preservation system.

Which kind of a model would support different trust domains?
A limited number of highly sensitive data sets (for example protected and private data) could be preserved by its owner institution only, while large amounts of less sensitive data (e.g. scans of commercially printed books, newspapers, web harvests) could be hosted in less protected tiers (or circles) of the infrastructure. Such circles could for example include (1) an isolated standalone tier, (2) a tier that is secured but connected and inter-operating with other systems, (3) a tier that manages data using cloud resources. Other parameters that could be relaxed depending on the tier might be related to safety, security, and preservation policies the data is subject to. In general, we expect that collaboration will need strong means to control what and how resources are shared between the stakeholders.

4.4 Building a Preservation Infrastructure for research data

Research is intrinsically a collaborative endeavour, especially within large scientific collaborations, with teams of people engaged in common projects, each contributing their own digital artefacts to the common collection, together with their views and comments. An infrastructure which supports the preservation of the records of research should support a preservation analysis process, have strong core data archiving and cataloguing processes, and support additional layers of contextual information, from information about formats and data descriptions, through contextual information on the data collection, through to the preserving the provenance for a complete picture of the actions under taken to collect and process data. The capturing of both explicit and tacit knowledge should be supported, and the distributed nature of research collaborations needs to be accommodated. OAIS does provide an abstract framework to develop such a framework, with most of these notions finding a place at some level. However, this standard is subject to a good deal of different interpretation and many of the details need to be fleshed out.

Furthermore, in such infrastructures, a good management of data at the "bit level" – that is maintaining the physical identity of the data – is crucial. Much of this aspect is part of the good

[44] http://www.dch-rp.eu/

practise of the management of a data centre, with staff and resources available and procedures in place to maintain availability of data, as part of the active use of live data. This would include the use of quality data storage and management tools, systems and procedures, although there are aspects which are of special value for long term preservation. This is usually known as bit preservation, and involves the following aspects.

- Replication: ensuring that copies of data are maintained, including at different locations.
- Integrity checking: checking the data against corruption, typically via checksums which test whether the physical bits stored have been corrupted of changed
- Media refresh: moving the data periodically onto new (tape or disc) media to mitigate against the effects of physical decay of the media material; also as physical media become obsolete, there is a need to transfer to new storage technology.
- Scaling: all issues of bit preservation are subject to scaling issues; these tasks become harder as data volumes increase, both in terms of total amount of data, and also in number of data units (e.g. files) stored.

Bit preservation is not seen as such a great research challenge particularly in the Science Data domain. In a sense it is “standard business” - (Bicarregui, Gray, Henderson, Jones, Lambert, & Matthews, 2013) discusses how “big science” projects in particular can factor in digital preservation from the beginning as a product of good data management projects; issues here are on long term resourcing, and good planning, rather than specific bit preservation challenges . For “bench” science, there are subject and institutional repositories which are collecting data, and again it is resourcing and managing these collections which are challenges rather than the bit preservation per se. A list of research challenges can be derived.

What are the components for efficient preservation of research in distributed architectures?
Both SCAPE and SCIDIP-ES[45] have built components of a preservation infrastructure. SCAPE has a collection of tools which while powerful, are not specially tailored to preserving science. SCIDIP-ES has taken an OAIS based approach and consequently has a direct approach to capturing Representation Information as “RepInfo objects”. This emphasis on preserving RepInfo is a step in the direction of preserving more of the science context, especially when used for Semantic RepInfo, describing domain parameters and also necessary software. However, defining and describing RepInfo is not straightforward even with these tools. A detailed analysis of the preservation scenario is needed, which is difficult for domain specialists (rather than information specialists) to carry out. There is a need for guidelines and processes for specific domains; European Space Agency’s Long Term Digital Preservation guidelines present an approach to this (LTDP Working Group, 2012), and it needs proving in other domains.

How do we enable bit preservation of very large files?
Scaling means that some of the bit preservation tasks (file integrity checking, file format checking and verification, media refresh) may take a long time. Generating a check sum of a very large file (of up to 100s of GB and potentially larger) may take many hours and may be impractical for large files. Experiments on scientific archives with Hadoop (as part of SCAPE) have demonstrated that while there is utility in using such approaches to support specific preservation actions, there is also a need to tailor the approach to the specific needs of the archive. The overheads of adapting and

[45] http://www.scidip-es.eu/

integrating an established working repository within a Hadoop architecture and using legacy systems and software may overwhelm the advantages gained.

How to enable preserving the record of research in a distributed environment?
As a consequence of using distributed environments, the artefacts may be distributed in different locations with different ownerships. Thus a preservation system for the records of research would need to accommodate:

- The location of artefacts in different locations, potentially with copies and versions of artefacts in different places.
- Maintaining a link structure across repositories in different places, which are under different jurisdictions and may change at different rates.
- Managing the Trust relationships between people and organisations to provide the appropriate guarantees that there can be stability of preservation for the long term.
- Attribution and rights management so that people can be properly credited for their contribution to the scientific activity.

How do we establish trust and sustainability of distributed architectures?
The linked data approach proposed in the work on Research Objects in SCAPE, and elsewhere, also works well within a distributed environment. There is no necessity for artefacts to reside in the same archive, and links can be external as well as internal. However, there remain issues of trust and sustainability in a distributed architecture. If archive managers are going to link to external sources, they require some guarantees. They require that artefacts kept in other archives are:

- **stable**, do not change and maintain their identity (especially in dereferencing of persistent identifiers);
- **accurate**, the information that they offer is truthful and accurate to some specified means;
- **accessible**, the rights to accessing the artefact do not change, and;
- **meaningful**, the artefacts are provided with sufficient context in their own right to be understood as objects of interest to science.

There is also a need for sustainability in the long term, with due consideration for managing archive change and archive migration.

4.5 Component Based Workflows

Scientific workflows support the combination of data and processes into a configurable, structured set of steps that implement computational solutions. They are increasingly used in experimental science, primarily due to their balance of expressivity and ease of programming. The SCAPE project has made extensive use of workflows to manage and enact the preservation process. In particular, the Taverna workflow management system has been used to describe preservation processes, invoking various services for planning and execution of institutional preservation and quality assurance strategies.

Within SCAPE, workflows are constructed using *components*: sub-workflows that can be used without necessarily exposing implementation details (Fellows & Plangg, 2014). In addition to supporting abstraction, the use of components or sub-workflows can facilitate re-use and sharing. In order to successfully apply a component based approach, however, there is a requirement that components are discoverable. To facilitate search and use of components, semantic annotations have been used to provide additional information about the components. Component *families* group together related components with a component *profile* describing the interface of the components within that family. This specification of profiles makes use of terms drawn from an *ontology.* Components are held in a component repository (provided by the myExperiment service which has been enhanced for this purpose).

While there are challenges concerned with the engineering required to use components within a workflow engine such as Taverna and the provision of a component registry (for example, the definition and implementation of APIs to support component repositories and the integration of the Taverna workbench with those repositories), there are wider issues relating to the management and use of components.

How do we manage the evolution of vocabularies?
The terms and annotations that are used to describe the components are all potentially subject to change, presenting problems relating to ontology management and evolution (Zablith, et al., 2013). This is, of course, a challenge faced in many other situations, not just in component-based workflows. For example, the issue of change management is also faced in the use of vocabularies and ontologies within the planning and watch process (Kulovits, Kraxner, Plangg, Becker, & Bechhofer, 2013).

How do we encourage and support the publication of annotation information?
Effective use of a component based approach requires components to be annotated, published and curated. Such annotation and publication is unlikely to be achieved for “free”. Incentives are required to encourage annotation and publication of components in order to support their discovery. Citation mechanisms are needed that allow reference to workflows and workflow components. Measurements of those citations can then provide assessments of the impact that component’s publication has had. What are the (alt-)metrics[46] that could be used to measure their impact, and thus encourage or incentivise annotation and publication. There may be a place here for the gamification (where users are issued with rewards, such as the badge system in Stack Overflow) of curation/publishing activities.

4.6 Research Objects for Workflow Preservation

The term Research Object (RO)[47] is being used to describe an emerging approach to the publication and exchange of scholarly information on the Web. Research Objects are aggregations of content with semantic annotations that describe the aggregations and the resources within them (Belhajjame, et al., 2012).

While the Research Object idea is being applied to a variety of domains and problems (for example supporting reproducibility[48], the exchange of code[49] and models in systems biology[50]), the use of ROs to support preservation of scientific workflows has been an area of particular focus. The W4fEver

[46] http://altmetrics.org/manifesto/
[47] http://www.researchobject.org
[48] http://isa-tools.github.io/soapdenovo2/
[49] http://www.researchobject.org/initiative/code-as-a-research-object/
[50] http://www.researchobject.org/initiative/combine-archive/

project[51] was an EU Specific Targeted Research Project (STREP) concerned with the preservation of scientific workflows. Thus, rather than considering the *use* of workflows to drive the preservation process, Wf4Ever considered the workflows as the *targets* of the preservation process. This is of interest/concern if digital preservation activities are tied to the use of workflows to drive operations as those operations/workflows then need to be the subject of preservation.

Within the Wf4Ever project, enhancements were made to the myExperiment service to support the use of Research Objects. This allows the aggregation of resources such as a workflow along with additional information such as input data files that related to a particular study, investigation or usage. Minimum Information Models or checklists were used to express the constraints or conditions that were required or expected allowing for “health-check” services that identify potential issues with preserved objects or provide assessments of “completeness”. The Planning and Watch architecture used within SCAPE (Kraxner, Plangg, Duretec, Becker, & Faria, 2013) was a source of inspiration for Wf4Ever.

The Research Object notion is not unique to Wf4Ever and is the focus of activity elsewhere[52] – for example within SCAPE, STFC considered the notion of Investigation Research Objects within the Research Data test beds, bundling together resources relating to scientific data sets. Components and Component Families (see above) can also be considered as Research Objects. The use of such an approach brings with it many challenges.

How do we manage aggregations of heterogeneous resources?
A general challenge here is the management of aggregations that involve resources under *mixed stewardship* – ROs can contain pointers to third party resources that are outside the control of the RO creator or owner. Workflows may themselves use third party services that are subject to change or decay. How do we define the boundaries of our Research Objects or aggregations? Which resources are considered to form the aggregation? How do we manage the identity of composite objects? What are the versioning mechanisms needed to support evolution and change and when should aggregations be considered immutable? The mixed stewardship and distributed nature of these aggregations provide particular challenges here.

What does it mean for a workflow to decay?
Preservation is concerned with ensuring continued access to materials for as long as necessary. Workflows are essentially executable objects, so such access may include not just the ability to inspect the object, but also the ability to re-run or re-execute the workflow. Mechanisms are needed to characterise the decay of preserved workflows. For example why is the workflow no longer executable or why is it producing different results? How might such decay be monitored or measured, and what actions can be taken to mitigate it?

How might workflows be “born-preservable”?
Design patterns provide reusable solutions to commonly occurring problems. What are the design patterns that may be applied to Research Objects, workflows or components that support or

[51] http://www.wf4ever-project.org/

[52] See the work of Force11 [https://www.force11.org/] on the future of research publishing, the COMBINE archive [http://co.mbine.org/documents/archive] in systems biology or the Mozilla Science Code as a Research Object Project [http://mozillascience.github.io/code-research-object/]

facilitate preservation? What specific annotations or metadata are needed for "born-preservable" workflows and Research Objects?

What are the policies and guidelines needed to manage workflow preservation?
Policies provide formal statements as to how an organisation will carry out its activities[53], with preservation policies describing approaches to be taken by a repository towards the preservation of objects. These policies will then guide preservation activities. While there is existing work on policies for digital preservation (e.g. the work of SCAPE's Preservation and Watch sub-project (Kraxner, Plangg, Duretec, Becker, & Faria, 2013) (Kulovits, Kraxner, Plangg, Becker, & Bechhofer, 2013) (Sierman, Jones, Bechhofer, & Elstrom, 2013)) less is understood about policies for management of workflows.

4.7 Provenance

Provenance provides "… information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness" (Groth & Moreau, 2013). Within SCAPE, provenance information generated during workflow execution records information about process execution, entities generated, potential failures or errors during execution and so on. This can be used to inform planning of runtime performance and resource requirements and to annotate the resulting data products. Recent community activities have resulted in the PROV Recommendation from W3C (Groth & Moreau, 2013), providing a common model and collection of serialisations of that model that support interoperation of provenance information. A significant eco-system has grown up around the standard.[54]

What to capture?
Indiscriminate capture of provenance information can result in large amounts of data with corresponding problems in interpreting or using that data. How do we determine the appropriate levels of granularity for provenance capture? What contextual information is necessary – details of execution environments, actors involved, implicit knowledge surrounding the intention or purpose of activities?

How do we support interpretation of provenance?
Interpretation of provenance information requires appropriate presentation mechanisms. How do we visualise provenance traces? What are the abstractions or distillations that can support interpretation of the underlying information? How do we navigate around provenance information and the relationships between artefacts?

4.8 Preservation Analysis

The case for preserving data is not obvious. This is particularly the case within many science contexts as not all science data is equally reusable and thus valuable, and it may be the case that the cost of maintaining large volumes of data may outweigh the potential benefit derived. Thus each collection of data needs a separate analysis of the preservation case. This would include:

[53] http://www.alliancepermanentaccess.org/index.php/consultancy/dpglossary
[54] http://provenanceweek.dlr.de/

- **Developing a Preservation Policy**: an analysis of the collection to be preserved to determine criteria for retaining artefacts, who the target audience is and what would be their expected level of competency ("designated community"), and for how long the data would be expected to be retained?
- **Developing a Business Case**: what are the costs and benefits associated with the preservation of data, what future technological and social risks can be anticipated in preserving the science with associated costs?
- **Developing a Preservation Strategy**: a detailed description of the approach taken for preservation, including hardware and support, replication strategy, what and how related representation information is collected and managed, tools and services used, and processes and procedures to maintain the archive?
- **Preservation watch and managing conceptual shifts :** what process and procedures for maintain the accessibility and usability of the archive in the face of changes in technology and in the designated community?

In organisations whose focus is on the creation and management of data, the business case for preservation as a discrete activity is not yet accepted. There are costs and benefits associated with preserving research data; the benefits in particular are not well explored in the literature, so need further empirical evidence. Further, while the importance of bit preservation is a case which is accepted and well-understood in good data management practise, the importance of maintaining understandability and identifying what that entails, and hence its cost is still under exploration.

The concept of the Preservation Network Model (PNM) (Conway, Giaretta, Lambert, & Matthews, 2011) (Conway, Matthews, Giaretta, Lambert, Wilson, & Draper, 2012) was developed in the European project CASPAR[55] and subsequently, was developed as part of the preservation analysis within a wider establishment of the preservation case. This method in based on the OAIS model and considers the dependencies between digital objects and the representation information components which give them context and how these dependencies impact the cost of preservation, and their maintenance over the long terms. This has been supported by the federated preservation tools and services developed in SCIDIP-ES.

SCAPE has considered how to ensure that the policies used to control the preservation actions within the data infrastructure reflect the business goals of the organisation. Specifically it has looked at a mechanism to map business guidance via a model of common preservation procedures to a specific control policy expressed in an ontology (Kulovits, Kraxner, Plangg, Becker, & Bechhofer, 2013).

A list of research questions can be derived.

How do we establish practical applicability of PNM?
The practical application of this technique in a variety of scenarios which are tailored to the particular needs of the domain community to be explored further to make it a practical approach, and also to manage different preservation strategies (e.g. emulation and migration). Additionally,

[55] CASPAR - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval - an Integrated Project co-financed by the European Union within the 6thFramework Programme 2006-2009. http://www.casparpreserves.eu/index.html

the formal modelling of the PNM needs to be established; with a formal notion of preservation objectives can be satisfied to determine whether a particular chosen preservation strategy is sufficient and complete enough to achieve the objective. PNMs also do not cover the contextual environment; further analysis is needed to determine whether sufficient contextual information has been collected to preserve the records of research (see the discussion on context in the Chapter 4.9).

How do we improve understanding of significant properties in the scientific data?
Creating high-level preservation policies is a complex and time consuming business. To be able to write an effective policy, the key characteristics, or significant properties, of the object need to be identified, and the environment required in order to ensure that these are maintained needs to be described. This is complex for all objects, but especially so for the data within a scientific data, where the potential compromises are not yet identified. For example, one may decide that having a photograph/image is acceptable in black and white as it is known what is being lost is the definition provided by colour and this may not be vital to the information content of the image; does anyone know what the compromise is for a specialised data file from a neutron spallation source? Of course, depending on the designated community, only having images in black and white may not be acceptable. It is these sort of significant properties that need to be captured in the form of a machine-readable policy.

How do we ensure that machine readable policy statements are consistent with high level policy statements?
On a policy making front, one of the open issues is the ability to check whether the derived machine readable policy statements can be reverse engineered back into the original policy statements. This is needed to ensure that the conversion process has not introduced inconsistencies and to be able to assure one's self that if the natural language policy changes, then the machine readable policy also is adapted in line.

4.9 Preserving the Context of Research in a Preservation Infrastructure

Preserving the context in which research has been undertaken requires a different point of view on the preservation problem. A broader view is needed, from the maintenance of the digital artefacts themselves to the collection and maintenance of information which provide insight into how the data should be interpreted, and thus preserve the record of the research activity. This requires the selection, elicitation, capturing and linking of the appropriate information, which could include the following:

- Information about instruments, sensors, samples, data sampling conditions, parameters measured, coverage, units and data rates.
- Information on the intention of the observation, its methodology, and the actors involved in the data collection.
- Information on the environment in which the data has been collected which may have an influence on the interpretation and calibration information on the instruments so that data can be normalised against a reference measurement.
- Information on errors, tolerances and biases known to affect the data.
- Tacit knowledge concerning the data collection, which may be captured in (Electronic) Laboratory Notebooks, websites, blogs, social media, annotations etc. This tacit knowledge may document the information considered above.

Preserving science provenance extends the notion of preserving the research context to cover a wider portion of the research lifecycle, so that how the research develops to generate intellectual outputs is recorded and made available. Thus, we would need to preserve a number of different types of research artefacts (for example, raw and derived data, software, workflows, visualisations, various documents), and also the relationships between them to record the full picture of how research results are derived. This relationship structure should be considered separately from the artefacts themselves; the same artefacts can occur in different relationship structures representing (different views on) different scientific points of view or results. To preserve the full context of research, we need to consider:

- How to best represent and practically capture the dependencies and relationships between artefacts generated and used in the research process.
- The specific preservation needs of different types of digital artefacts, particularly: software, visualisation, documents, and workflows as well as data.
- Navigating through provenance structures to address particular digital artefacts in context.
- Aggregating and packaging aggregations of artefacts as digital objects in their own right
- Tools for handling aggregations of artefacts as digital objects in their own right.

Further, not only do we need to capture the explicit knowledge of the data collected, encapsulated in databases, file-stores, documentation, registries, ontologies etc., but we also need to keep implicit or tacit knowledge which is kept informally in peoples' minds or within the dialogue which goes on between people. This knowledge is vital for a true understanding of why the data collection was undertaken in the way it was. It uses the prior knowledge and experience of the data collector, their developed intuitions, and their observations on the conduct of the experiment which lead to intervene in the way they do.

Software is a class of electronic object which is by its very nature digital, and the preservation of software is often a vital prerequisite to the preservation of other electronic objects. However, software has many characteristics that make preserving it substantially more challenging than for many other types of digital object. Software is inherently complex, normally composed of a very large number of highly interdependent components and often forbiddingly opaque for people other than those who were directly involved in its development. Software is also highly sensitive to its operating environment, with the typical software artefact depending on a large number of other items including compilers, runtime environments, operating systems, documentation and even the hardware platform with its built-in software stack. Preserving a piece of software thus involves preserving much of its context as well.

Thus said a list of research questions can be derived:

What kinds of models are needed for capturing context and provenance of research in preservation infrastructures?
The SCIDIP-ES approach has limited support for context; RepInfo is a very general concept which needs developing for particular domains. It could be used to represent context, as could parts of the accompanying Preservation Description Information, but it has not as yet been developed for that purpose. Further, the notions in OAIS are not tailored to support the links and dependencies between items needed to support capturing contextual information.

This equally applies to capturing Provenance, part of context which has been discussed in more detail earlier. Again, there is a need to capture networks of relationships between artefacts which are not well supported in current preservation architectures. Thus there is a need for networks of relationships to be captured and stored to record research; the Research Object approach discussed above, using Linked Data techniques, developed in Workflow4Ever and further explored in SCAPE is well-suited for this. Thus an approach which combines the SCIDIP-ES approach to OAIS with a linked data approach would be a strong candidate to extend the preservation infrastructure discussed above to support a more complete record of research.

This approach would bring the SCIDIP-ES information model into the Research Object world, using the Ontology for OAIS for the basis of this approach, but also combine this with other relevant linked-data vocabularies, including: Prov-O, Open Annotation format, DCat and OAI-ORE, and PREMIS, together with domain specific ontologies. This Research Object view allows us add science context, so that AIPs can be generated which capture the relationships between entities rather than treat them in isolation.

This linked data approach would also allow the tools to be more loosely coupled in a linked data framework, thus exposing RepInfo and other information via linked data endpoints. Tools such as SCOUT should then be adapted for use in this framework.

What constitutes a "complete" research object?

Research objects try to encapsulate a scientific objective, bringing all the items of interest together and grouping them. This raises the issue of what constitutes a "complete" research object. In a particular domain, we could reasonably expect that research objects of a particular type (in the case of the SCAPE analysis, an Investigation Research Object) would have particular artefacts and relationships present. This would be the output of a preservation analysis in the particular context of the domain of under study. This would allow an assessment of the completeness of ROs to be established.

How do we manage (im)mutability of the Research Object?

The notion of immutability of the Research Object is not clear: there are some which are immutable (experiment and raw data) – and there are some which are extensible (supplementary data & publications) – is there a point when these extensible items should be present? The Workflows4ever project Evolution model explored this and it will be interesting to see how this work can be transferred to a wider context.

Further, this issue of change means that the IRO, with its unique identifier becomes so different that it needs to be considered to be a new entity with a new identifier. An example of this would be when the underlying experimental data is migrated from one format to another, is this same IRO or a different one? Should there be links to both versions even though the fact that a migration has occurred means that there was some preservation risk to the original data?

Using the notion of research object as a more open ended bounded object raises the question of what exactly is being published persistently in this case. If we add additional information are we maintaining stability? Research Objects are well suited to notions of versioning, where we can relate objects together as they change, thus keeping the old boundary stable.

How do we capture, link and preserve tacit knowledge?

Tacit knowledge is notoriously hard to capture. It may be written in tools such as blogs, social media, Electronic Laboratory Notebooks etc. Work is required to manage the preservation of these types of record and link them appropriately to the explicit data objects. Research in business knowledge management could be of particular use here, with its emphasis on the elicitation of tacit knowledge, via techniques such as interviews (which may include media such as video), via storytelling and after action review, or communities of practise.

We wish to move from a point of view for preservation which moves from preserving artefacts, such as documents, data or digital objects, to preserving *a record of the human activity*. It is the knowledge of the reasons behind the data which makes the artefacts useful in the future, both to understand and validate the work undertaken in the past, and to give sufficient understanding of these artefacts so that they can be reused in the future.

Thus we see that preservation should be seen as *knowledge management*. A vision of a preservation system should try capture and preserve both the explicit knowledge of the activity, embodied in data, documents and other artefacts, but also the implicit knowledge, trying to capture the experience and intuitions behind the decisions made in the data collection process.

How do we integrate software preservation in preservation infrastructures?

There have been models developed for the systematic preservation of software. In particular, (Matthews, Shaon, Bicarregui, Jones, Woodcock, & Conway, 2009) (Matthews, Shaon, Bicarregui, & Jones, 2010) (Matthews, Shaon, & Conway, 2012) outline the issues which arise when considering the preservation of software, including: the motivations for its preservation; the complexity of software leading to the question of what should actually be preserved; different strategies which are undertaken in the preservation of software, and criteria for judging whether software has been preserved to an adequate level of quality.

However, these have not been systematically supported and integrated into a preservation infrastructure, especially in regard to a migration rather than an emulation strategy for preserving software. This would best be done during the development of software – good software engineering practise should support good preservation, and research in software engineering, especially software reuse is particularly relevant to apply to preservation. In practise, preservation needs to be undertaken for legacy system, so not all parts of the detailed software description may be available.

5 Conclusion

This document reports on research contributions from the SCAPE project in the domain of automation and scalability, research challenges identified by a broader community at the IPRES 2013 workshop and nine research challenges identified based on the SCAPE project but taking into account a much broader scope.

Automation and scalability in the digital preservation field have been addressed by the SCAPE project in two main areas: 1) addressing the storage and processing of large, continuously growing, amounts of digital objects and 2) adapting to continuously changing socio-technical environment. These have been addressed by the three main areas of the project - scalable platform, scalable planning and watch and scalable components - where different concepts, models, algorithms and tools have been developed to address specific problems.

While the project has advanced the field in many ways, new challenges are continuously rising and requiring future attention.

The community has continued defining new research challenges at the Open Research Challenges workshop, held at IPRES 2013. By now, there is a clear understanding that to get the full value from digital information it is equally important to consider the computation environment surrounding digital object, as well the digital object itself. Therefore, new and more sustainable models for digital preservation which consider computation are proposed. Emulation-as-a-Service (EaaS), for example, when successfully implemented is considered to be a solution which might trigger a wider use of emulation in the community.

Also, incorporating digital preservation needs in the software engineering practices could make the whole intent of information preservation more sustainable.

Continuous improvements in preservation components such as migration and quality assurance is still required so new approaches which would bring more automation in migration and improved capabilities of quality detection are envisioned.

Finally there is a recognized need to make the current practices in digital preservation field more rigorous, repeatable and based on a well-defined metrics.

Taking the more broad scope from the community but also building upon the results from the SCAPE project nine research challenges were identified.

Those challenges mainly try to bring future improvements in the field by:

- providing models and methods for better understanding of technology landscape
- considering different preservation infrastructures for preserving scientific memory but also cultural heritage.
- providing means for preserving preservation components and new repeatable benchmarking capabilities in order to enable continuous improvements of those components

Once successfully addressed, these challenges would bring future improvements to the field and thus make the digital information more secure.

6 Bibliography

Adomavicius, G., Bockstedt, J. C., Gupta, A., & Kauffman, R. J. (2008). Making Sense of Technology Trends in the Information Technology Landscape: A Design Science Approach. *MIS Quarterly* .

Akbik, A. (2014). *D6.3 Optimization of Preservation Processes.*

Bass, F. M. (1969). A New Product Growth for Model Consumer Durables. *Management Science* .

Bechhofer, S., Sierman, B., Jones, C., Elstrøm, G., Kulovits, H., & Becker, C. (2013). *D13.1 Final Version of Policy Specification Model.*

Becker, C., & Duretec, K. (2013). Free Benchmark Corpora for Preservation Experiments:Using Model-Driven Engineering to Generate Data Sets. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital Libraries*, (pp. 349-358).

Becker, C., Faria, L., & Duretec, K. (2014). Scalable Decision Support for Digital Preservation. *OCLC Systems & services* .

Becker, C., Rauber, A., Paton, N., Schmidt, R., Milic-Frayling, N., & Matthews, B. (2012). *D3.1 Open Research Challenges and Research Roadmap for SCAPE.*

Belhajjame, K., Corcho, O., Garijo, D., Zhao, J., Newman, D., Palma, R., et al. (2012). Workflowcentric research objects: First class citizens in scholarly discourse. *Proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web.* Heraklion.

Bicarregui, J., Gray, N., Henderson, R., Jones, R., Lambert, S., & Matthews, B. (2013). Data Management and Preservation Planning for Big Science. *International Journal of Digital Curation* , 29-41.

Chanod, J.-P., Dobreva, M., Rauber, A., Ross, S., & Casarosa, V. (2010). *Automation in Digital Preservation.* Leibnitz-Zentrum fuer Informatik.

Chanod, J.-P., Dobreva, M., Rauber, A., Ross, S., & Casarosa, V. (2010). *Issues in Digital Preservation:Towards a New Research Agenda.*

Conway, E., Giaretta, D., Lambert, S., & Matthews, B. (2011). Curating Scientific Research Data for the Long Term: A Preservation Analysis Method in Context. *International Journal of Digital Curation* .

Conway, E., Matthews, B., Giaretta, D., Lambert, S., Wilson, M., & Draper, N. (2012). Managing Risks in the Preservation of Research Data with Preservation Networks. *International Journal of Digital Curation* .

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, (pp. 107-113). New York.

DigitalPreservationEurope, c. (June 2006). *Deliverable D7.2 - Research Roadmap.*

European Commision Information Society and Media Directorate-General. (2011). *The Future of the Past - Shaping new visions for EU-research in digital preservation.*

Faria, L., Duretec, K., Kulmukhametov, A., Moldrup-Dalum, P., Medjkoune, L., Pop, R., et al. (2014). *D12.2 Final version of the Preservation Watch component.*

Fellows, D., & Plangg, M. (2014). *D7.3 Design and implementation of the preservation component catalogue.*

Groth, P., & Moreau, L. (2013). *PROV-Overview An Overview of the PROV Family of Documents. W3C Working Group Note*. Retrieved from http://www.w3.org/TR/prov-overview/

Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Redmond: Microsoft Research.

High Level Expert Group on Scientific Data. (2011). *Riding the wave: How Europe can gain from the rising tide of scientific data.*

ISO/IEC. (2011). *Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Systems and software quality models (ISO/IEC 25010).*

Kraxner, M., Plangg, M., Duretec, K., Becker, C., & Faria, L. (2013). The SCAPE Planning and Watch suite – Supporting the preservation lifecycle in repositories. *Proceedings of the 10th International Conference on Preservation of Digital Objects.* Lisbon.

Kulmukhametov, A., & Becker, C. (2014). Content Profiling for Preservation: Improving scale, depth and quality. *ICADL.* Chiang Mai.

Kulmukhametov, A., Plangg, M., & Becker, C. (2014). Automated Quality Assurance for Migration of born-digital Images. *Archiving.* Berlin.

Kulovits, H., Becker, C., & Andersen, B. (2013). Scalable preservation decisions: A controlled case study. *IS&T Archiving.* Washington.

Kulovits, H., Kraxner, M., Plangg, M., Becker, C., & Bechhofer, S. (2013). Open Preservation Data: Controlled vocabularies and ontologies for preservation ecosystems. *10th International Conference on Preservation of Digital Objects.* Lisbon.

LTDP Working Group. (2012). *Long Term Preservation of Earth Observation Space Data: European LTDP Common Guidelines.*

Maniatis, P., Roussopoulos, M., Giuli, T., Rosenthal, D. S., & Baker, M. (2005). The LOCKSS peer-to-peer digital preservation system. *ACM Transactions on Computer Systems* , 2-50.

Matthews, B., Shaon, A., & Conway, E. (2012). How do I know that I have Preserved Software? In *The Preservation of Complex Objects. Volume 1 Visualisations and Simulations.*

Matthews, B., Shaon, A., Bicarregui, J., & Jones, C. (2010). A Framework for Software Preservation. *International Journal of Digital Curation* .

Matthews, B., Shaon, A., Bicarregui, J., Jones, C., Woodcock, J., & Conway, E. (2009). Towards a Methodology for Software Preservation. *6th International Conference on Preservation of Digital Objects.* San Francisco.

PARSE Insight, c. (June 2010). *Deliverable D2.2 Scienece Data Infrastructure Roadmap.*

Pehlivan, Z., Lechervy, A., Sanoja, A., Pitzalis, D., Jurik, B. A., Schindler, A., et al. (2013). *D11.2 Quality Assurance Workflow, Release 2 + Release Report.*

Peres, R., Muller, E., & Mahajan, V. (2010). Innovation Diffusion and New Product Growth Models: A Critical Review and Research Directions. *Intern. J. of Research in Marketing* .

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

Rosenthal, D. S. (2010). Format Obsolescence: Assessing the Threat and the Defenses. *Library Hi Tech*.

Rothenberg, J. (1995). Ensuring the longevity of digital documents. *Scientific American* .

Schmidt, R. (2012). An Architectural Overview of the SCAPE Preservation Platform. *Proceedings of the Ninth International Conference on Preservation of Digital Objects.* Toronto.

Schmidt, R., Rella, M., & Schlarb, S. (2014). ToMaR -- A Data Generator for Large Volumes of Content. *14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, (pp. 937-942).

Sierman, B., Jones, C., & Elstrøm, G. (2014). *D13.2 Catalogue of preservation policy elements.*

Sierman, B., Jones, C., Bechhofer, S., & Elstrom, G. (2013). Preservation Policy Levels in SCAPE. *10th International Conference on Preservation of Digital Objects.* Lisbon.

Strodl, S., Petrov, P., & Rauber, A. (2011). *Research on digital preservation within projects co-funded by the European Union in the ICT programme.* Vienna University of Technology, Vienna.

Zablith, F., Antoniou, G., d'Aquin, M., Flouris, G., Kondylakis, H., Motta, E., et al. (2013). Ontology evolution: a process-centric survey. *The Knowledge Engineering Review* .