Open research challenges and research roadmap for SCAPE

Grant Agreement Number	270137
Full Project Title	Scalable Preservation Environments
Project Acronym	**SCAPE**
Title of Deliverable	Open research challenges and research roadmap for SCAPE
Deliverable Number	D3.1
Work-package	XA.WP.3
Dissemination Level	**PU** = Public
Deliverable Nature	**R** = Report
Contractual Delivery Date	2012-11-30
Actual Delivery Date	2012-11-29
Author(s)	Christoph Becker (TUW) Norman Paton (UNIMAN) Rainer Schmidt (AIT) Natasa Milic-Frayling (MSRC) Andreas Rauber (TUW) Brian Matthews (STFC)

Abstract

This roadmap identifies research topics to be addressed within the SCAPE project between M24 and M38. It provides a general overview of research goals addressed in the R&D work packages in the project, and analyses topics emerging in the cross-section of these and the open issues identified in the general roadmap. We furthermore provide a report on the outcomes of a workshop on Open Research Challenges in DP organised at IPRES 2012.

Based on this, we identify a number of key future research topics that will be focus areas for research. These are: (1) information models and benchmarking, (2) advanced simulation and prediction models, and (3) future preservation infrastructures.

Keyword list

Digital Preservation, Research Questions, Open Research Challenge

Authors

Person	Role	Partner	Contribution
Christoph Becker	Lead author	TUW	Authoring, editing contributions
XA.WP.3 participants	Contributors	WP participants: AIT, MSR, STFC, TUB	

Document Approval

Person	Role	Partner
Ross King	Coordinator	AIT
Christoph Becker	WP lead	TUW
Jose Carlos Ramalho	Reviewer	KEEPS / University of Minho

Distribution

Person	Role	Partner
XA.WP.3 participants	Contributors	WP participants: AIT, MSR, STFC, TUB
User Group	Comments	All content partners

Revision History

Ver.	Status	Author	Date	Changes
0.1	Draft	Christoph Becker	2012-06-26	Initial outline
0.2	Draft	Christoph Becker	2012-07-18	Extending background and references
0.3	Draft	Christoph Becker	2012-09-14	Extending section 2
0.4	Draft	Christoph Becker	2012-09-27	Extending section 2, adding results from the September meeting on gaps and challenges
0.5	Draft	Christoph Becker	2012-09-28	Added Norman's and Rainer's contribution on the future preservation infrastructure task
0.6	Draft	Christoph Becker	2012-10-08	Added section on IPRES workshop, including summaries of topics based on input by table hosts Brian Matthews, Rainer Schmidt, and Andreas Rauber.
0.7	Draft	Christoph Becker	2012-10-10	Added PC research goals with some compression on the text, revisions and comments with Miguel Ferreira. Added the section on IPRES workshop table "experimentation simulation, prediction"
0.8	Draft	Kresimir Duretec, Christoph Becker	2012-10-10	Added advanced simulation and prediction chapter
0.9	Draft	Christoph Becker	2012-10-12	Partially fixed references, added general sections, included RDS goals
0.10	Draft for review	Christoph Becker	2012-10-15	Included input from Natasa Milic-Frayling into sections 3.1 and 4.3. General clean-up, removed comments.
0.11	Revised draft	Christoph Becker	2012-11-23	Revisions based on review comments, integrated workshop report section from Cal Lee.
1.0	Final	Christoph Becker	2012-11-26	Final version for release

Executive Summary

This report outlines the research roadmap of the SCAPE project, focused on the scalability of preservation systems in terms of storing and processing as well as decision making and control. It positions the research carried out in SCAPE within the European research landscape focused on digital preservation research. It further outlines the key goals of the R&D work packages in SCAPE, grouped according to sub-projects (preservation components, preservation platform, and preservation planning and watch). Each research goal shortly outlines the state of art, key contributions, and open issues.

Broadly speaking, these goals strive to

1. advance the state of art in scalable preservation components and processes for preservation actions, content analysis, and quality assurance,
2. provide flexible mechanisms for constructing powerful preservation workflows based on such components,
3. advance the start of art in flexible, scalable, distributed parallel execution of such processes based on paradigms such as MapReduce, and
4. provide scalable mechanisms for decision making and control.

The document furthermore reports on the results of a workshop on Open Research Challenges, organized at IPRES 2012, which received strong participation from the global DP community. Discussions were grouped in six topics:

1. Digital information models
2. Value, utility, cost, risk and benefit
3. Organizational aspects
4. Experimentation, simulation, and prediction
5. Changing paradigms, shift, evolution
6. Future content and the long tail

Based on the collected research goals and the broad involvement of the DP community, we identify and outline common gaps and openings for future research and finally, three emerging critical research topics that arise from the cross-section of identified open problems and point to fundamental research questions. These are, broadly speaking:

1. Future preservation infrastructures.
2. Advanced simulation and prediction models.
3. Information models and benchmarking.

We furthermore conclude that it is paramount to continue analysing emerging research topics and challenges throughout the project and beyond. This will also provide crucial input for the final research roadmap that will be delivered by SCAPE in 2014.

Table of Contents

1 Introduction

Digital Preservation has emerged as a key challenge for information systems in almost any domain from cultural heritage and eGovernment to eScience, finance, health, and personal life. The field is increasingly recognised and has taken major strides in the last decade. However, key areas of research are often limited to applying solutions to existing problems rather than proactively investigating the challenges ahead and probing for innovative break-through approaches that would radically advance the domain.

The work package XA.WP.3 Open Research Challenges focuses on innovative and emerging research having the potential to dramatically improve our capabilities. A series of focussed research activities will contribute to emerging challenges arising from the cross-section of problems posed in the project, and introduce innovative approaches from other domains to cross-fertilize applied research. At the end of the project, this work package will deliver a research roadmap that will lay foundations for advanced research and upcoming issues and potentials looking towards more long-term solutions for the future. This forward-looking nature of the work package opens up a broad perspective of questions relevant to digital preservation. The limited resources of the work package, on the other hand, force us to focus on a selected few key questions to address within the frame of SCAPE. Hence, the roadmap outlined in this deliverable will identify new research topics to be addressed within the SCAPE project. We will outline overall research topics currently in focus of the DP community and discuss in more detail the research questions addressed in SCAPE. This will lead to an overview of open questions and topics that arise from identified open questions. Additionally, we report on a workshop at IPRES 2012 where we engaged with the broader, global DP community. This discussion sets the basis for identifying particular advanced research topics to investigate in SCAPE until 2014.

This report is structured as follows. Section 2 gives an overview of research goals addressed in SCAPE. We start with a short high-level introduction that positions SCAPE in the European DP research landscape and then focus on the research roadmap of SCAPE. We outline collected research goals that are on the roadmap of the Platform, the Planning and Watch, and the Preservation Components subprojects, as well as the testbeds. Based on these goals and the open issues identified in each of them, we set out in Section 3 to discover emerging issues reaching out the community, by looking at other projects and reporting on a workshop conducted at IPRES 2012. In Section 4, we combine this view with an outward look, by analysing the issues and gaps within the SCAPE roadmap and identifying promising, yet challenging future topics. Section 5 sums up the discussions and points forward.

2 Research in Digital Preservation

2.1 Digital Preservation Research roadmaps

Visions about the future of digital preservation are outlined in a number of research roadmaps such as the DPE [1] and Parse.Insight reports [2]. A previous SCAPE report on research in European Digital Preservation projects summarizes key goals and application areas of current R&D [3].

One of the key observations from these reports is a slow shift from addressing questions that help to fix problems in maintaining digital information over time to ensuring that the problem will not appear in its full complexity in the first place, reducing the need for specific ex-post fixing.

With the progress made in DP research so far, the community has developed a solid understanding of the problems and the approaches needed to fix them, turning DP activities in some areas into a challenging engineering task that requires further attention. Beyond that, however, more fundamental research is required in order to ensure that the information artefacts and information systems of the future pose less of a challenge in terms of preservation.

This can be seen in research challenges focussing on the development of DP-ready systems, integrating DP requirements in any system design and development process. A higher level of resiliency against technological changes on all levels will not only make preservation easier, it will also offer benefits in the operation of information systems.
A further area of focus is automation on all levels to be able to deal with the increasing amounts as well as growing levels of complexity of objects that have to be dealt with. While the focus of the former will be on scalable architectures, the focus of the latter will need to involve a more solid understanding of the fundamental concepts of digital information including entire systems and distributed processes.
We also observe a shift in the community recognizing the need for preservation solutions and thus also stakeholders in DP related research and development. While originally being strongly based in the cultural heritage and scientific data domain, stakeholders from a range of other disciplines involved in e-* activities (e-health, e-government, e-commerce) realize their dependency on electronic information and processes beyond legal retention requirements for their core operations. This will have an impact on the type of solutions expected, as well as the approaches taken to meet these, broadening both the interdisciplinarity as well as the methodological approaches to be taken.
With digital preservation having evolved into a dedicated and highly specialized discipline in its own right, a further challenge now will be to reach out again to other disciplines to bring in know-how from highly specialized domains. Within the ICT domain, this will require attracting input from groups in areas such as information systems, software engineering, embedded systems design, algorithm and compilers, theory of computing, security, semantic technologies, IT Governance, and Enterprise Architecture, among many others. To address the technological challenges in digital preservation specifically within the broadening application domains, where solutions are direly needed, will require teams integrating experts from a range of ICT disciplines, organizational and legal experts and domain experts to cover the entire lifecycle and operational context of an information system.
In this context, the SCAPE project is focused on scalability of preservation systems in terms of storing and processing as well as decision making and control. This context guides, but hopefully does not constrain, the scope and vision of this document.

2.2 Current Research Questions in DP

Current research in DP is expanding the notion of content to be preserved beyond the preservation of static artefacts, documents and data structures. The extended focus includes interactive objects, embedded objects, ontologies and ephemeral data. Examples for this development in Europe are the LIWA project[1] addressing the dynamic nature of Web Archiving, the TIMBUS project[2] focusing on the preservation of business processes, *Wf4Ever*[3] working on workflow preservation, and *BlogForever*[4] focusing on blogs.
Much research and development in digital preservation focuses on scalable preservation systems. This need stems from the user communities requesting tools, methods and models that perform on realistic, heterogeneous large collections of complex digital objects. A second aspect of handling vast amounts of objects effectively is the automation and decision support in a number of stages, ranging from object selection and tool performance to validation criteria.
In the past, a number of conceptually well designed modules for digital preservation tasks were developed that required human intervention. Current research is focused on taking these modules to

[1] http://liwa-project.eu/
[2] http://timbusproject.net/
[3] http://www.wf4ever-project.org/
[4] http://blogforever.eu/

the next level and providing a high degree of automation of preservation processes as well as assist decision making.
Examples are the SCAPE project that is primarily addressing the scalability issue, and ARCOMEM[5] that is using the social web for automated information creation and supported appraisal. The ENSURE project will research on scalable pay-as-you-go infrastructures for preservation services for integration into workflows.
The third issue addressed by current projects is networking. An achievement of past projects with intensive outreaching and publication activities is the broadening of the digital preservation community. Awareness about digital preservation stretches far beyond the traditional archive, library and museum sector (ALM), now reaching the academic sector as well as the industry and enterprise domains. This development is well reflected in current project consortia with increasing participation of industry players as solution providers as well as problem owners.

2.3 Research Goals in SCAPE

2.3.1 Overview and method

To enable mapping out the landscape of research goals that are on the individual roadmaps of the subprojects, a common template structure was used and sent to all work packages. The individual subprojects then provided a number of research goals. These are not meant to exhaustively reflect all work items that are being carried out in the work packages, but instead to reflect the key research goals to enable analysis of their key relations and identification of common issues in relation to on-going work outside the project. The next sections list the research goals collected.

2.3.2 Scalable platform[6]

A common model for implementing parallel preservation actions

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   MapReduce provides a programming model and execution framework for processing structured data at large-scale using a parallel system. In SCAPE, we are seeking ways to apply this methodology to the domain of digital preservation. While it is certainly possible to develop map-reduce applications that solve individual preservation problems, it is challenging to make these applications reusable and interoperable. In particular, the implicit and undeclared handling of data IO, the implemented data models, and runtime dependencies hinder the interoperability of such applications. This in turn carries the risk of developing highly specific and monolithic applications that are short-lived and difficult to sustain beyond a particular experiment. It is therefore required to adopt appropriate software design principles that can tackle these challenges. The goal addressed here is the development of a component-oriented approach, which allows us to create non-trivial parallel preservation applications that are reusable, modular, and independent of specific input (container) formats.
2. *Current state-of-the-art*.
   On-going research is dealing with the classification and specification of preservation components taking into account different aspects like interfaces, semantic descriptions, or performance characteristics. Modularized preservation components have been implemented using technologies like object-oriented languages, web services, or shell scripts. A number of meta-data standards for modelling and serializing data objects as exchangeable records exist.

[5] http://www.arcomem.eu/
[6] This section is comprised of the input gathered from the PT subproject.

3. *Research contributions*.
   In order to efficiently leverage data-intensive environments for digital preservation applications, it will be important to develop a framework for implementing scalable preservation components that conform to a defined programming and data exchange models. In SCAPE, we will develop the required abstractions to develop scalable preservation components as well as a service-oriented environment to configure, execute, and monitor the parallel preservation applications.
4. *Open issues*.
   This approach will provide a generic model that allows a user to easily attach different input sources and output sinks to preservation components that operate in a parallel environment. The model will however rely on the mechanisms provided by the underlying framework to distribute and balance the workload among worker nodes. Issues and improvements with respect to data locality, data (re-)distribution, communication, and the involved distributed data structure are not addressed.

Scalable preservation platform architecture

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   Executing preservation scenarios on large-scale datasets requires, besides careful workflow preparation, also an execution of such workflow on a platform that will allow scaling to large datasets. By designing the software architecture of the large-scale preservation platform we aspire to provide a solution overcoming problems when passing from small scale preservation to large-scale.
2. *Current state-of-the-art*.
   Taverna as a model of a small-scale preservation platform is studied. It provides means to exploit all participating modules, e.g., repository, execution platform and result presentation.
3. *Research contributions*.
   In SCAPE, we will design and develop the architecture of a large-scale preservation system. The envisaged contributions are: (i) collection of necessary interchanges between the independent modules (APIs). (ii) iterative design and validation of the required scenarios on the designed architecture and platform in scale. (iii) challenges related to the deployment of the architecture on the hardware instance.
4. *Open issues*.
   Performance oriented integration of already existing modules might involve re-implementation of certain APIs and endpoints of third party systems. This effort might not always succeed since the particular developers are not always part of the project.

Cloud workload management

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   In SCAPE, individual preservation tasks are represented using workflows, which are executed over a cloud platform that includes preservation actions that can be applied to different data sets. However, planning activities may give rise to a diverse collection of workflows over different collections, which raises questions such as: (i) how much resource should be allocated to different workflows or users; (ii) how should the cloud be configured and the workflows compiled to ensure effective performance across the entire workload; and (iii) when there is contention for resources, how can these resources be allocated in ways that meet user expectations.

2. *Current state-of-the-art*.
   There has been significant effort on workload management in different settings, and this is a currently active research area for clouds. In practice, a wide range of techniques can be applied, but individual results often make quite strong assumptions that may not be applicable to preservation scenarios. One goal may be to support autonomic workload management, with a view to minimising systems management overheads
3. *Research contributions*.
   In SCAPE, in PT.WP2, there is a task on resource management, but this is difficult to separate out from workflow compilation in practice. The specific opportunity in SCAPE may be to take high level descriptions of preservation plans into account when developing and evolving workload management strategies. Almost any approach that is highly self-managing, for example by using learning to revise policies, is likely to be pushing the state-of-the-art.
4. *Open issues*.
   This work has not really started in the project. There will be a need for simple default policies, but we should develop the architecture in such a way that these can be evolved to allow more ambitious capabilities. It may be that effective workload management will be every bit as important as workflow compilation for achieving scalability.

Hadoop as a Storage Backend for Fedora-based repositories

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefit.*
   The increase in volume of digital material poses new challenges for digital preservation in terms of performance and scalability. Traditional repository architectures do not meet the requirements for such situations particularly well. Computation Clusters like Hadoop are built to process large amount of data in a very short period of time. Using Hadoop as a storage backend of repositories would enable the user to run preservation tasks in a performant and scalable way over large amounts of data.
2. *Current state-of-the-art.*
   Fedora Commons is a widely used and well known repository. In SCAPE, three out of four repositories are built on top of Fedora Commons. For writing digital objects representations and managed datastreams in a persistent storage the Fedora Commons architecture offers a plugin called Akubra. By exposing an API, developers are enabled to create implementations of Akubra for arbitrary storage systems (e.g. Content Addressable Storage, GRID, and Cloud). Concrete implementations are for example with Dell DX Object Storage Platform and iRODS. The SCAPE computation platform builds on Hadoop, and as such the repositories must be able to store digital objects and datastreams on Hadoop via Akubra.
3. *Research contributions*.
   In SCAPE we are seeking for ways to wire the Hadoop Cluster, as a scalable computation and storage platform, with Fedora Commons repositories in an efficient way. By implementing the Akubra API for the HDFS and / or HBASE, data processing within the Hadoop framework via Map-Reduce becomes feasible without exporting the whole corpora for processing beforehand in order to perform compute-intensive tasks in distributed way.
4. *Open issues.*
   Since Akubra is only responsible for persisting serialized objects and their datastreams, only the storage layer can profit from the distribution via the Hadoop framework. Other resources Fedora depends upon for its services like the database or the web application itself remain non-scalable. In order to be able to turn Fedora itself in a distributed system, another layer

of abstraction in Fedora's management module as described by the High-level proposal Storage [4] is needed.

Scalable execution of workflows for clouds

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   Scientific workflows provide high level, declarative techniques for describing recurring application requirements. In SCAPE, we are deploying the Taverna workflow language and development environment to write workflows that coordinate the application of preservation actions. In so doing, we aspire to maintain ease of authoring, while supporting scalable execution over cloud platforms.
2. *Current state-of-the-art*.
   Several scientific workflow systems have been compiled to execute on parallel platforms, and there are proposals that allow map/reduce programs to be written explicitly using workflow languages and that allow the writing of workflows that call cloud services.
3. *Research contributions*.
   In SCAPE, we will develop techniques that execute Taverna workflows transparently over map/reduce, so that workflow authors are insulated from the execution environment where their workflows run. Thus the expected contributions are: (i) techniques for scalable implementation of scientific workflows on clouds; (ii) evaluation of these techniques with workflows in digital preservation; and (iii) techniques for generating comprehensive provenance records with low overheads.
4. *Open issues*.
   It is certainly possible that this activity will leave some performance challenges unaddressed and that scalability will require some manual tuning. This specific goal is silent on workload management.

2.3.3 Scalable planning and watch[7]

Efficient creation of trustworthy preservation plans

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   A preservation plan nowadays is constructed largely manually, which involves substantial effort. This effort is spent in analysing and describing the key properties of the content that the plan is created for; identifying, formulating and formalizing requirements; discovering and evaluating applicable actions; taking a decision on the recommended steps and activities; and initiating deployment and execution of the preservation plan. When automating such steps, trustworthiness must not be sacrificed for efficiency. Still, the goal is to substantially increase the efficiency of planning, so that the effort to create a plan is reduced, for example, to a couple of hours.
2. *Current state-of-the-art*.
   The preservation planning framework and tool Plato provide a well-known and solid approach to create preservation plans. However, the Planets-based planning tool needs some rework to be fit for SCAPE and interoperable with Taverna, the reference repositories, etc. Most importantly, on the basis of a prototype, automation heuristics and modules need

[7] This section is comprised of the input gathered from the PW subproject.

to be developed to automated manual steps and hence increase the efficiency of using Plato to create preservation plans. The effect of such improvement should be measured.

3. *Research contributions*.
   We will address the bottleneck of decision processes and processing information required for decision making. We build on a clear workflow based on well-established and proven principles, and automate now-manual aspects such as collection profiling, constraints modelling, requirements reuse, measurements, and continuous monitoring. The starting point is a baseline prototype of the SCAPE planning component and a roadmap for manual aspects to be automated. The resulting operations will be validated for compliance with criteria for trustworthy repositories.
4. *Open issues*.
   It will not be possible to deliver trustworthy planning in a fully *autonomous* way, without intervention of a decision maker. Furthermore, we will likely not be able to conduct research into new paradigms for services in the cloud, such as the creation of preservation plans as a service offered to repositories.
   Plan portfolio management would include plan optimisation of decisions across plans to achieve a certain strategy. This might be out of scope.

Automated mechanisms for collecting and analysing preservation-related information

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   For successful preservation operations, a preservation system needs to be capable of monitoring compliance of preservation operations to specifications, alignment of these operations with the organisation's preservation objectives, and associated risks and opportunities. This requires linking a number of diverse information sources and specifying complex conditions. Doing this automatically in an integrated system should yield tremendous benefits in scalability and enable sharing of preservation information (especially risks and opportunities).
2. *Current state-of-the-art*.
   Isolated strands of systematically collecting information that can be used to guide preservation decision making have been developed. Well-known examples include registries of file formats or emulation environments. However, these are far from being complete in the information they cover, and there are few links between the islands of information.
3. *Research contributions*.
   We will systematically identify sources of information that need to be monitored. Based on this, we will develop a 'Watch component' that collects information from a number of sources, links it, and provides notifications to interested parties when specified conditions are satisfied. This entails an information model of the domain, a system architecture and design, and the development of such a system.
4. *Open issues*.
   While the envisioned sources to be included cover a substantial part of the world of interest, we will certainly not be able to cover all interesting and relevant information sources. For example, valuable information about preservation risks is hidden in the web in extremely diverse and partially implicit forms. Similarly, this research stream cannot invest into quantifying the correctness of the information provided by a source and is thus silent on reliability. Finally, fully automated reaction to identified conditions is out of scope.

Scalable Content Profiling

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   Systematic analysis of digital object sets and the identification of sample objects that are representative of a collection are critical steps towards preservation operations and a fundamental enabler for successful preservation planning: Without a full understanding of the properties and peculiarities of the content at hand, informed decisions and effective actions cannot be taken.
   Content profiling essentially consists of three high-level steps: gathering (primarily technical) metadata, processing & aggregation and meta-data analysis.
2. *Current state-of-the-art*.
   Approaches and tools demonstrated thus far are often focused solely on format identification. Still, automatic characterisation and meta data extraction is done by numerous tools such as Apache Tika and JHove/JHove2. The FITS tool follows a different approach that unifies many different characterization tools, but instead provides a normalized output of their results and gives indicators for their validity. These features provide a solid basis for preservation analysis and a complete content profile.
3. *Research contributions*.
   One key argument against the usage of in-depth characterization is that the analysis of metadata produced is extremely time-consuming. This stems from the observation that even the amount of metadata itself may be substantial. However, scalable approaches for content characterization can build on parallel architectures such as map-reduce to increase the processing speed in the analysis itself. We will develop and evaluate a prototype tool to generate content profiles in a scalable fashion as a key source of information for Watch, and evaluate its scalability on large real-world collections.
4. *Open issues*.
   Several advanced questions arise on the basis of this tool that might be outside of the scope of this work package. This includes dynamic automated partitioning into homogeneous subsets based on multi-dimensional views of content and sophisticated mechanisms for finding representative sets from massive data collections.

Simulation and prediction

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   When hosting a small collection of files, the capacity and computational load needed for preservation it is not of big importance. On the other side, hosting large amounts of data requires insight in storage and computational requirements. To gain such insight, observing the current situation is not enough – a “look into the future” is needed. The reason for that lies in several facts: For example, collections can grow because new files are inserted and certain actions need to be taken to keep the collection accessible. Most importantly, interactions and dependencies between possible actions and their outcomes are complex and often defy direct human assessment.
   Providing a simulation environment that simulates a collection and its evolution through time will enable us to predict (with some level of certainty) the state of a collection in the future and therefore enable the user to make a better decision in the present. Such a simulation environment can also deepen insights into causal relations of influence factors, actions and their effects, and the longevity of content.

2. *Current state-of-the-art*.
   There is little knowledge and no formal models about causal effects of preservation actions and their effects. Simulation, however, is a mature field with existing approaches, frameworks and tools.
3. *Research contributions*.
   It becomes obvious that a number of dimensions and aspects have to be considered for meaningful simulation. This ranges from content lifecycles to categories of preservation actions, formats, and content profiles.
   A key question will focus on the question how to model whole repositories, i.e. what the level and scope is on which we should do the simulation. This ranges from large collections and their feature distribution to single complex objects and their internal construction. Further on, we will investigate how to model those collections and objects, their complex relations, and aspects such as ingest and file format obsolescence. Finally, we will investigate ways to evaluate the results of simulation and prediction and quantify prediction confidence.
4. *Open issues*.
   Instances of complete information loss could be simulated, but currently this is considered out of scope for this work. In general, simulation is a very early topic in DP and will start focussing on a narrowly defined set of phenomena, gradually expanding and refining the underlying models to represent more complex cause-effect relationships.

Formalized preservation policy representation

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   While there is an increasing awareness and understanding of the interplay of technology, business goals, strategies, and policies for digital preservation, there is no standard model for formalizing preservation policies to provide the required context for preservation planning, monitoring and operations. This context is key to successful preservation; however, so far, it is provided implicitly by decision makers. This not only puts additional burden on decision makers, but also threatens the quality and transparency of planning and actions. What we need is a policy model that relates general human readable preservation policies to a more refined level of preservation policies that can be understood by automated processes and enables decision makers to formulate policies so they can be understood by automated processes.
   With preservation policies, we refer to elements of governance that guide, shape and control the preservation activities of an organisation.
2. *Current state-of-the-art*.
   The term “policies” in DP is used very ambiguously; often, it is associated with mission statements and high-level strategic documents. Representing these in formal models would lead to only limited benefit for systems automation and scalability, since they are intended for humans. On the other hand, models exist for general machine-level policies and business policies. However, a deep domain understanding is required to bring clarity into the different levels and dimensions at hand.
3. *Research contributions*.
   We will clarify the different levels of control involved in DP, from strategies to operations; collect aspects of policies that are relevant from both a top-down strategic view and a bottom-up operational planning view; and clarify the key elements of policy statements that can and should be formalized and fed into systems. This will lead to an iteratively refined machine-understandable policy model. This model will be related to a higher level intended

for decision makers: a policy elements catalogue. Both will be evaluated and refined, and their elements will be set in relation to each other to clarify how the different levels of control interact. The result will support organizations to define their own preservation policies and to better understand the need to describe them.

4. *Open issues*.
   While providing a machine-understandable model for policy specification is a key goal of the work package, the description of work is silent on how users should be supported in their policy creation activity. That means that sophisticated tooling for manipulating such a machine-understandable model may be out of scope. For example, we will not develop a framework that lets decision makers write their policy statements in natural language.

Loosely-coupled preservation systems

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   Preservation planning focuses on the creation of preservation plans; Preservation Watch focuses on gathering and analysing information; Policies focuses on the representation of organisational goals, objectives, and directives. These methods and tools will in general be deployed in conjunction with a repository environment. This requires open interfaces and demonstrated integration patterns in order to be useful in practice.
2. *Current state-of-the-art*.
   Preservation plans are specified following a published XML schema, but there are no standards for policies, monitoring specifications, Service Level Agreements for preservation operations, or system interfaces.
3. *Research contributions*.
   We will specify APIs for all key interface points between systems and PW, i.e. between Planning and Repositories; Planning and Watch; and Repositories and Watch. Finally, we will develop ontologies for policy specification.
   For all APIs, we will provide Reference Implementations.
4. *Open issues*.
   Evolution and extension over time (including after SCAPE). Repository migration is not considered as in scope for this goal.

Preservation planning as a continuous management activity

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   Through its data-centric execution platform, SCAPE will substantially improve scalability for handling massive amounts of data and securing quality assurance without human intervention. But fundamentally, for a system to be truly operational on a large scale, all components involved need to scale up. We need an approach to planning, monitoring, and operating a repository on a terabyte-scale. Only scalable monitoring and decision making enables automated, large-scale systems operation by scaling up the decision making and QA structures, policies, processes, and procedures for monitoring and action.
   Apart from automated systems and interfaces, this also requires us to improve organisational processes. Preservation planning is a decision making process in an organisational setting, supported by methods and tools. While the frameworks and tools developed in SCAPE can be deployed in different settings, it is often hard for organisations to assess where they stand in their capabilities, so that they could target specific improvements. Currently, there are no agreed and tested mechanisms to help organisations to improve their preservation planning

and monitoring capabilities.
Numerous organisations are investigating approaches and tools for preservation planning. Providing them with a mechanism for assessment and improvement would enable them to advance their preservation planning and monitoring capabilities. This can for example be measured on typical maturity scales from 0 (non-existent) to 5 (optimizing).

2. *Current state-of-the-art*.
The ISO 16363 criteria on trustworthy include certain criteria that are related to preservation planning and management. However, they are focused on compliance to the OAIS model for the purpose of audit and certification. As such, they are not meant to be actionable and do not provide advice or guidance on how an organisation can improve what it does to better meet its goals. Maturity models and governance frameworks, however, provide the necessary mechanisms for such assessment and improvement.
3. *Research contributions*.
We will develop a framework for clarifying required capabilities, responsibilities and roles, and for assessing the maturity of preservation planning and monitoring in an organisation.
4. *Open issues*.
Standardised public benchmarking of organisations and approval of such a maturity model by a standard body would be tremendously valuable, but is clearly out of scope.

2.3.4 Scalable components[8]

Identify and select existing digital preservation action tools & services

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
This goal aims at identifying, assessing and selecting currently available action tools that are compatible with the SCAPE platform and necessary to solve the problems portrayed by the SCAPE testbed scenarios.
2. *Current state-of-the-art*.
There are some reports from previous projects that list off-the-shelf commercial and open-source migration tools. However, these do not assess tools on the grounds of whether these are compatible with SCAPE requirements.
3. *Research contributions*.
A registry of useful tools for digital preservation.
4. *Open issues*.
Format coverage is always an issue. A tool registry becomes obsolete pretty quickly if no one cares to update it. A strategy towards a collaborative effort is likely to be necessary. Several niche formats are difficult to migrate as no open-source tools are available, nor through format documentation.

Identifying Preservation Actions regarding research datasets

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits.*
Further preservation actions may be required in the case of research datasets to maintain the contextual information in which the research dataset has been collected. These include:
a. *Risk acceptance and monitoring*. Rather than take definite preservation actions that alter the content, the repository records specific instructions about external information sources,

[8] This section is comprised of the input gathered from the PC subproject, with only minor edits.

the nature of what needs to be monitored, and considered in terms of risk to the long-term reusability of the information.
b. *Migration*. This may or may not involve the loss of information but should always force the re-evaluation of the Preservation Network Model (i.e. the representation information dependency graph).
c. *Description*. This may use textual or formal data description languages such as DRB or EAST to provide supplementary representation information. Thus the service may incorporate some sort of automated mechanism for (re-)checking the preservation decisions made in the representation information record originally used to define the preservation actions for a AIP, and relinking and augmenting the existing.

2. *Current state-of-the-art.*
Migration, integrity checking and syntactic validation well understood and included as preservation actions.
3. *Research contributions*.
Classifying preservation actions for research datasets. Identifying how the preservation action service for Research Datasets (RD) can be controlled by the preservation plan to maintain representation information dependencies. Prototyping of preservation action prototypes specific for RDs which monitor and maintain the representation information dependencies.
4. *Open issues*.
Preservation actions which manage the representation information dependency graphs are not well understood. How to manage compound objects is not well understood. Software packages as representation information are a complex area.

Ensure large-scale applicability of preservation action services

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
Preservation actions, especially migration tools, have been extensively analysed and employed in experimental digital preservation systems. However, current approaches are often not capable of coping with real size collections. This goal is focused on the applicability of such tools to large collections of complex digital objects in a timely manner, by focusing on analysing and improving the interfaces and internal functionality of existing preservation action tools, extending and creating new large-scale preservation functionality and enabling tools to deal with not only single file formats but also with compound objects (container objects with a set of related files in different file formats).
2. *Current state-of-the-art*.
Current tools are often not capable of handling large-size digital object collections. The tools need to be adapted to be able to run on parallel execution platforms.
3. *Research contributions*.
The ability to process millions of files in a short period of time by making use of all available computing power, and not a single machine.
4. *Open issues*.
There is still quite a lot of uncertainty about which platform architecture will best support this goal. It is certainly possible that this activity will leave some performance challenges unaddressed and that scalability will require some manual tuning.

Ensure interoperability between service clients and cloud services providers

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   As cloud computing services are becoming more prevalent and distinct execution platforms are available, interoperability becomes an issue. Sometimes, service execution paths cross the boundaries of a single execution platform (e.g. tools can only run on a different platform), so transparent platform interoperability is something to attain.
2. *Current state-of-the-art*.
   Azure cloud services based on Windows operating system and Hadoop parallel execution platform are currently incompatible.
3. *Research contributions*.
   Transparent execution of action services workflows over two or more distinct execution platforms (e.g. Hadoop vs. Azure, Linux vs. Windows).
4. *Open issues*.
   There is still no consolidated strategy on how to attain this goal. Depending on the approach taken by the PT SP, it might not be possible to run Taverna workflows on Azure network.

Data publication platform

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   While open data sources, such as *PRONOM*[9], *Software Conversion Registry* (CSR)[10] and *govdocs*[11] are excellent examples of publishing re-usable data (to some extent) there is still a big problem with gaining access to other sources of data. This is mainly due to projects and organisations not focussing on re-usability of data, rather just their own internal aims. PRONOM and govdocs are great examples of where this is not the case but other valuable sources of data are disappearing due to many of the wrong reasons. At the other end of the scale the currently published datasets are missing valuable information relating to their context, version and provenance.
   Jeni Tennison puts some of the issues very succinctly: "It's fairly obvious that high quality data, supplied in a timely and consistent fashion, is going to be easier to use and more accurate than low quality data, supplied as and when, using different formats and coding schemes within each release".
2. *Current state-of-the-art.*
   Not enough data sets are available as 20th century open data (let alone 21st century). More high quality, small and easy to maintain datasets are needed. Of the 21st century linked dataset, very few are maintained with full provenance information (not that they were before). It is the second point that is particularly relevant when it comes to analysing risk related to changes. This is also a problem not just for the preservation community but also in the wider area of web and semantic web research. Indeed the problem of provenance information is known to this community (http://www.jenitennison.com/blog/node/141), who we appear not to be working very closely with.
3. *Research contributions*.
   As part of the SCAPE/OPF/University of Southampton work, the LDS3 Specification for managing fully provenance aware datasets was constructed. This specification was then implemented in order to be a publication platform for digital preservation data and also a potential way of solving the PRONOM problem with provenance information. In SCAPE, we will develop techniques that execute Taverna workflows transparently over map/reduce, so

[9] http://www.nationalarchives.gov.uk/PRONOM
[10] https://isda.ncsa.illinois.edu/NARA/CSR/php/search/conversions.php
[11] http://digitalcorpora.org/corpora/files

that workflow authors are insulated from the execution environment where their workflows run. Thus the expected contributions are:

(i) Techniques for scalable implementation of scientific workflows on clouds.
(ii) Evaluation of these techniques with workflows in digital preservation.
(iii) Techniques for generating comprehensive provenance records with low overheads.

4. *Open issues.*

*H*ow do we get people to create more high quality and maintainable preservation datasets, do they even exist? Where? And how do we get at them? In the most part these are not technological problems. What is the business of open datasets (http://www.jenitennison.com/blog/node/172)?
There is still work to be done with the wider community on how to enable clear discovery of current and historical data. Further can we dynamically query historical data using protocols such as Memento[12]? Within the preservation community, how do we build provenance aware services for users? Where do these fit into the current situations? Do they scale?
Proof of concepts are coming along, but integrated platforms are still more silos than integrated solutions.

Support the growing use of web content for analytical purposes by allowing analysis of large scale collections of web pages

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   The digital preservation community generates a plethora of mineable information in diverse forms and media. We want to harness this information for preservation.
2. *Current state-of-the-art.*
   The field of text mining is highly active, but the topic is still fairly new within the digital preservation community.
3. *Research contributions*.
   In SCAPE, we will develop techniques that use techniques from the large-scale text analysis field for enhancing information gathering.
4. *Open issues.*
   None identified by the work package.

Highly accurate visual aspect based web page version comparison

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   Web page version comparison is of great interest in Web archiving (check the quality of the archive, adjust crawling strategy, emulation, control migration …). Combined with machine learning techniques, it helps in automating the decision making for the mentioned tasks.
2. *Current state-of-the-art*.
   Existing approaches are limited. Hash based comparison is simple but very inaccurate (in the sense of “understanding” the differences between subsequent versions). Structure based comparison allows for locating the important changes but do not fully take into account the visual aspect of snapshots. Image based comparison is accurate but does not exhibit the semantic of the content that is changed.

[12] http://mementoweb.org/

3. *Research contributions*.
   In SCAPE, we will develop a new approach that combines structure based and image based techniques, as well as learning strategies to produce fully automatic decision systems. Page versions are segmented in order to compare versions of semantically homogeneous blocks and detect changes in both the content and the block structure. This leads us to also address the issue of a new hybrid (structure and image) segmentation tool.
4. *Open issues*.
   Our approach is certainly the most accurate, but this comes with a performance cost. Some tasks may require faster but less accurate processing, which can lead us to study the way we can derive simplified versions of our tool.

Interactive end-user conversion of XML based documents

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   XML has become the dominant data representation language and XML-based formats, as well as being the basis of web content, have become the default for many office-productivity tools, including Microsoft Office, OpenOffice, LibreOffice and Apple's iWork. However, there is very little support for helping end users to use or convert between *arbitrary* XML based formats. Given a document in a specific format, if the rendering software for that format is obsolete or unavailable, a non-expert end user would lose all utility of the document.
2. *Current state-of-the-art*.
   There exist numerous XML based formats, and various tools for conversion between specific formats exist, which may be proprietary or free depending on the formats in question. Converters between certain formats may not always exist, especially for older, less popular or obsolete formats, and we are not aware of any general purpose tool that helps an end-user to view, process and convert arbitrary XML based formats.
3. *Research contributions*.
   We will investigate the design, development and evaluation of an interactive end user tool that, given a document in an arbitrary XML format, would aid the user to interactively discover the original formatting, layout or other properties of the document, or to reconstruct it with the goal of preserving the original intended semantics of the document. Any inferred conversion templates generated using this programming-by-demonstration approach may also be saved an applied to other documents, possibly allowing large scale conversion of documents without requiring programming expertise.
4. *Open issues*.
   How much of the utility of a document can be salvaged through such an interactive interface combined with various inference techniques? Can this process inspire any metrics for defining the "preservation cost" of using a given format?

Quality assurance for digital image collections

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   Currently many institutions are carrying out large-scale digitization projects. Resulting collections contain millions of image documents. Furthermore, many digitized collections are constantly improved with new versions. In that case collection operator has to select between old and new version of document, since only one version should be stored. Therefore, automated solutions for quality control are required in order to manage and to maintain such collections. Such solution should help collection operator to detect duplicated,

missing or added images.
Secondly, assessing the quality of migration processes such as TIFF-to-JPEG2000 can be challenging because the tools to do this do not exist, are not of sufficient quality, or do not support JPEG 2000. As a result, shortcomings of the migration workflow may go largely unnoticed. The benefits of addressing this goal would be twofold: a better migration path, and better ways to assess the quality of the produced images.

2. *Current state-of-the-art*.
Most existing approaches use global image descriptors in order to compare images in large collections. Optical character recognition approach is typical for information extraction from text documents but performs with insufficient accuracy and flexibility.
TIFF to JP2 migration workflows are now commonly used in operational settings, but the degree to which colour fidelity is preserved (or even important to begin with) is often left unspecified. Scalable solutions for assessing the quality of the resulting images appear to be largely non-existent. Examples are solutions that would establish whether i) an image is valid according to format specifications, ii) it conforms to a characteristics profile (e.g. progression order, number of quality layers, etc.), iii) pixel values are unchanged relative to the source image.
3. *Research contributions*.
In SCAPE, we develop an image comparison tool "Matchbox" that reduces digitization costs, improves quality of stored collections, runs automatically or semi-automatically and increases efficiency of human work. Thus the expected contributions are: (i) techniques for analysis of image collections applying modern image processing algorithms; (ii) evaluation of typical use cases for these techniques in modern digital preservation processes; (iii) techniques for finding duplicates in collection, comparison of digital collections and for comparison of particular two images; and (iv) support of scalable multithreaded processing for "Matchbox" jobs.
We will also develop tools to assess the quality of the generated JP2s (i.e. validation against format specifications), image comparison tools, and methods to test whether images conform to a pre-defined set of characteristics. The expected contributions are: (i) improved migration workflows, and (ii) new or improved tools and workflows for analysing and assessing the quality of JP2 images.
4. *Open issues*.
It is certainly possible that this activity will leave some quality assurance tasks in digital preservation unaddressed and they will require some manual tuning. Thus, one specific goal is to give an operator a quality assurance tool at hand that supports not only automatic but also human inspection in order to compare old and new instances of the corresponding documents and decide which version should be overwritten.
One limitation is that the current scope is restricted to RGB images (CMYK colour spaces are not covered, and are in fact not allowed in JP2), but within the project's context (which mostly involves digitised content) this is unlikely to be important.

Video and Audio Format Migration Quality Assurance.

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
The goal is to develop Quality Assurance for video (moving images) format migration. In particular the QA should be able to confirm that sound and video is still synchronic in the migrated files. The Danish State and University Library (SB) will then be able to migrate a 4TB Windows media video collection to a format better suited for preservation. Earlier attempts

at migration using ffmpeg failed on some files. Some of the migrated files had sound and video out of sync.

2. *Current state-of-the-art*.
   There is no established standard for preservation quality digital video. CARLI Digital Collections Users' Group (DCUG), Standards Subcommittee recommend MXF (.mxf) file format (best practice) [5]. The "Preferences in Summary for Moving Image Content" state, in the section on formats for professional moving image applications, that "Clarity and fidelity characteristics (bitstream encoding) should be used as the primary consideration; choice of file formats as secondary" [6]. The SB Danish TV broadcasts video collections are mostly in MPEG2 (various dimensions and video and audio bitrates, sampling rate: 48kHz, bit depth: 16). This format was chosen as it can be used for both recording, ingest, preservation and dissemination, thus minimal need for transcoding. The examples of digital video format migration for preservation are sparse, as are any uses of quality assurance in this context. Transcoding is however used widely for dissemination. The question is how much quality assurance is done in this context. Also quality assurance in digitization context should be considered.
3. *Research contributions.*
   In SCAPE we will develop video format migration quality assurance that will be able to catch faulty migrations such as sound and video out of sync. We will put the QA into workflows that can run on the SCAPE platform thus ensuring scalability and performance.
4. *Open issues*.
   In quality assurance it is always an open question, how much is enough? Through large scale heterogeneous testing we should over time be able to give some statistic "guarantees", such as 'the quality assurance catches any serious migration error with 98% likelihood'. Note that this also requires a definition of 'serious'. The statistical analysis as well as algorithm improvements will still be open for further research.

Matching Metadata with Data using Audio Indexing.

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   The Danish State and University Library (SB) Radio Broadcast Collections consist of 2 hour recordings. The metadata of each recording is the channel id, the start time and the end time of the recording. SB however also has program listings and even some news broadcast manuscripts in other collections. For preservation purposes we would like to match the program information to the recording where possible. We would like to extend the xCorrSound sound wave comparison tool [7] to search for jingles indicating the start of a certain program. This would make an indexing possible, and we could then match the metadata with the data.
2. *Current state-of-the-art*.
   There are audio fingerprinting algorithms used for identifying a song in a large archive of songs. And there are the current tools in the xCorrSound tool suite, which find the best offset for a match between two audio files based on computing the cross correlation.
3. *Research contributions*.
   In SCAPE we will develop a tool, which finds the offset(s) of a short sound wave piece in a large sound wave file. We will write workflows that can run on the SCAPE platform prioritising scalability. This will be used to match metadata to Danish radio broadcast recordings for preservation purposes.

4. *Open issues*.
   Performance improvements. A usability study of the most important metadata for researchers?

2.3.5 Additional research data testbed goals

The specific nature of the research data testbed brings about a number of additional research goals in relation to core goals of the R&D work packages.

Value proposition for research data

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   Establishing the value of long term preservation of research data is not straightforward; it is not always clear that all data should be kept, considering its cost of adequate preservation, cost of re-collection and potential for reuse. Guidance on establishing this value proposition is needed.
2. *Current state-of-the-art*.
   Existing cost models (e.g. LIFE) are not tailored to consider research data. A number of studies (such as KRDS) consider costs of research data.
3. *Research contributions*.
   In SCAPE, we will consider the factors which establish a value proposition for the testbed, and consider how to generalise them.
4. *Open issues*.
   It may not be possible to establish common guidelines for all research data. Cost information is hard to establish.

Preservation analysis for research data

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits.*
   Preservation analysis and planning for research data requires the description of the dependencies of the data to other digital objects providing representation information, forming a graph (Preservation Network Model, PNM). This is a complex process, and requires tool support.
2. *Current state-of-the-art*.
   Methodology for PNMs developed in CASPAR[13] and other projects.
3. *Research contributions*.
   Use of PNMs within RDT scenario. Development of prototype tools.
4. *Open issues*.
   Developing a formal model for PNMs; using PNMs to drive preservation watch and actions. How best to provide tools to support PNMs. Methodology for undertaking preservation analysis. Managing the scale and diversity of information objects requires making the preservation analysis: feasible in terms of the amount of work (and cost) to do the analysis; within a reasonable skill level of an analyst to undertake the work of analysis.

Developing a scalable platform for research data

The following are the key features of this goal:

[13] http://www.casparpreserves.eu/

1. *Motivation for addressing this goal and anticipated benefits*.
   Preservation tools and services need to be established and integrated to exercise and test the research data scenario as a prototype.
2. *Current state-of-the-art*.
   Tools such as Safety Deposit Box developed to support preservation and being adapted for research data. Prototype preservation platform in CASPAR.
3. *Research contributions*.
   Development of an architecture identifying the services required to support research data preservation. Integration with SCAPE platform based on Hadoop.
4. *Open issues.*
   Research data has often established systems in place for data management, especially in "big science" projects. Research data platform needs to take into account the legacy platform into which it is being introduced. Scalability is key: data sets are typically very large (TB in some cases). Data files within data sets may be very large (many GB per file); number of files in data sets may very large (1000s).

Preservation workflows for research data

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   Workflows for the stages of the research data testbed need to be established to automate processes at scale.
2. *Current state-of-the-art*.
   Simple workflows included in tools such as Safety Deposit Box.
3. *Research contributions*.
   Defining workflows which can be executed using Taverna.
4. *Open issues.*
   How to establish workflows which are specified by the PNM (and which may involve human intervention).

Persistent identifiers and links

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   In order for dependency networks of representation information and for compound objects, persistent identifier schemes need to be used to uniquely identify objects, and provided (semantically meaningful) links between them.
2. *Current state-of-the-art*.
   Many existing persistent identifier schemes (e.g. DOI, ARK, Handle, PURL). Linked Open Data provides schemes for linking and describing links between objects.
3. *Research contributions*.
   Using a persistent identifier service to identify objects; considering how to carry out re-linking and recombining links as data items change over time.
4. *Open issues.*
   Interacting between persistent identifiers services (APARSEN is looking at this). How to persist links over time.

Preserving complex research objects

The following are the key features of this goal:

1. *Motivation for addressing this goal and anticipated benefits*.
   Research data objects are rarely single digital objects (or homogeneous collections of objects), but rather collections of related objects, dataset, documents, raw, analysed and aggregate data, metadata, software components, images, visualisations etc. These need to be managed over time as a whole.
2. *Current state-of-the-art*.
   Some work on using OAI-ORE within a preservation context; development of provenance standards (e.g. W3C); frameworks for preserving software; and preserving workflows (e.g. Workflow4Ever).
3. *Research contributions*.
   An initial consideration of how to manage complex research data.
4. *Open issues.*
   Preserving software, preserving provenance, preserving workflows are all open questions. Preserving context.

Identified gaps and opportunities

The identification of gaps as part of the research goal description enables us to draw together the perspectives from the diverse work streams, identify common issues and opportunities, and use these as guidance on identifying, phrasing, positioning, and prioritizing challenging research topics and questions.

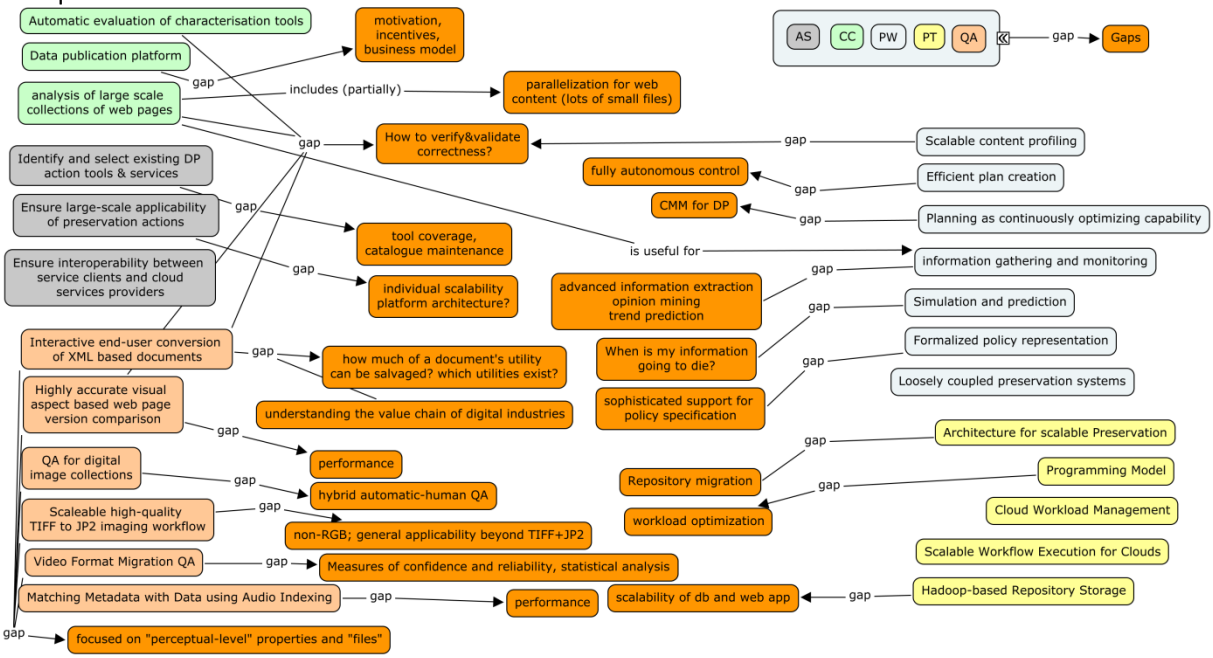

Figure 1 Research goals and identified gaps

To guide the discussion, Figure 1 illustrates, in condensed form, the research goals described above and the key gaps identified (in orange). It can be seen that some common gaps are identified by several work packages. This includes open issues of performance that go beyond the planned improvements and development innovations, but also the issue of verifying and validating the correctness of results obtained by characterization and QA processes and, in turn, analysed in content profiling. Other issues arise more from the cross-section of identified issues and gaps. This includes the notion of QoS fulfilment in a distributed environment: It is not possible to state a priori with complete certainty that a certain preservation action plan, i.e. workflow including complex components, can be fulfilled completely in a given environment in a given state. Even assuming that

full experimentation is conducted on a well-chosen set of sample objects, the environment in a given configuration will have finite resources and may not be able to carry out all tasks successfully. The question arises if we can address notions of varying degrees of QoS fulfilment and flexibility in the platform.

In the next section, we group these identified gaps and use this to guide the specification of challenging research topics for the roadmap.

2.4 Emerging Topics

Figure 2 shows the identified gaps from above, grouped into a set of related topics. It furthermore highlights a number of critical areas where unsolved questions and potential opportunities are identified.

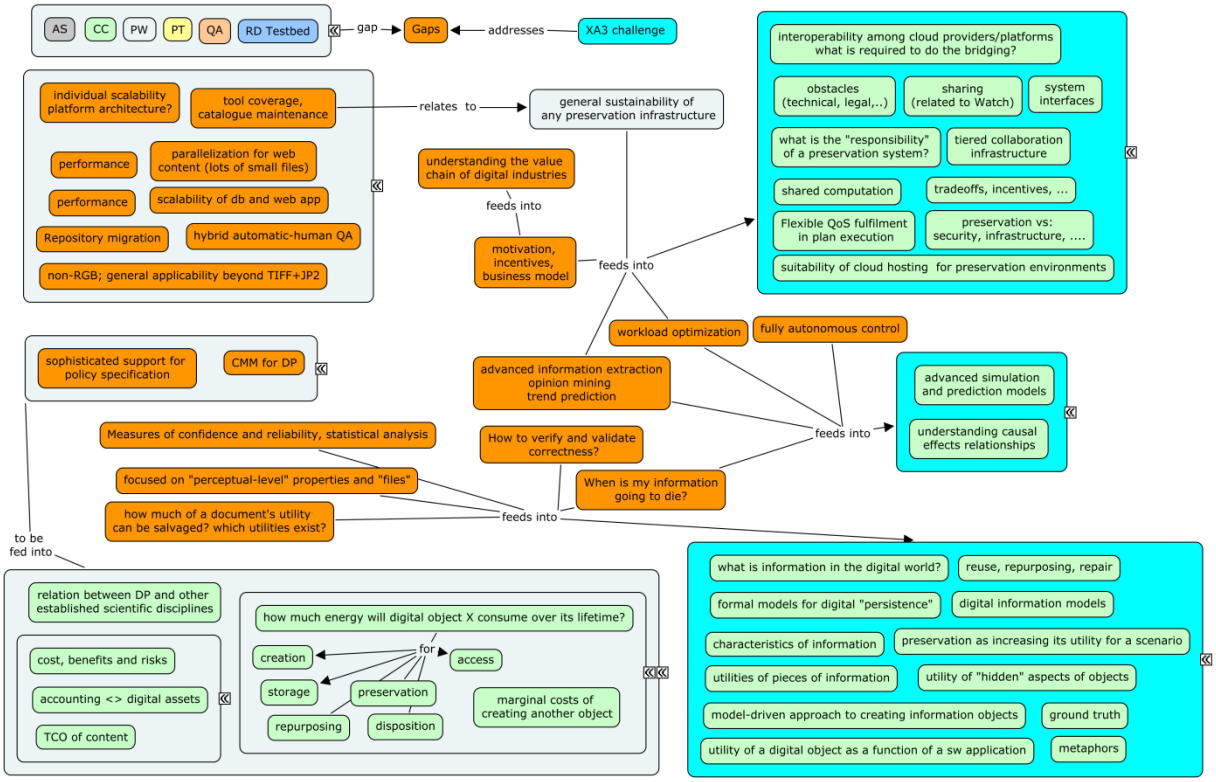

Figure 2 Research gaps grouped and associated challenges on the roadmap

We can broadly identify a number of categories.

- One set of gaps clearly points to *engineering challenges.* These can be found on the top left of the diagram. The gaps identified cover areas such as tool maintenance; the coverage of tools in comparison to the total desirable set of tools that could be covered; individual scalability of tools; and specific issues raised by the large numbers of small files encountered in web archives.
- A second, smaller set of gaps that belongs to *organizational challenges* is shown below the engineering field. It identifies topics such as opportunities to providing more sophisticated methods and tools for policy specification and management, as well as the area of capability maturity models that can provide tremendous help in assessing and improving organization's capabilities in digital preservation.

- Finally, there is a set of gaps that points to the need for a deeper understanding of several key issues. These are not grouped together since they show complex relationships. Instead, it can be seen that all of these gaps feed into one of three research challenges (shown in blue):
    - The top right challenge covers the notion of **future preservation infrastructures**. Questions covered include flexible QoS fulfilment, shared computation models, but also legal and technical obstacles to preservation cloud environments.
    - Below that, **advanced simulation and prediction models** are required to enable a deeper analysis of causal relationships and effects in the domain. This should take into account, and use as a starting point, the simulation environment being developed in Automated Watch, but reach further and allow deeper insight into complex issues such as workload optimization, autonomous operations, and cause-effect relationships pertaining to key risks, activities, events and outcomes in preservation operations.
    - On the bottom right, a number of challenging questions arises in the area of **information modelling and benchmarking**. There is an utter lack of benchmark approaches that combine public data sets with full ground truth. There is a clear need for innovative approaches to benchmarking in order to enable verification of correctness of preservation processes. Questioning and analysing in detail the notion of information in the digital world needs to dive into fundamental issues pertaining to the models we need for even thinking properly about digital information and its preservation. Aspects of reuse, repurposing and repair intersect with the needs for more formal models for digital "objects" and their characteristics.
    - Finally, a number of relevant challenges are identified on the bottom left that we may not be able to cover in depth within the project, but that are seen as relevant and emerging areas of interest. These need to be monitored over time as the project progresses.

3 Community involvement

In order to facilitate a broader discussion on emerging research challenges within the DP community, a workshop was conducted at IPRES 2012 that focused on future challenges in DP research.[14] The workshop was conducted by Christoph Becker and Andreas Rauber, TU Vienna, and Christopher A. Lee, UNC, US. We solicited – and received – wide-ranged contributions on challenges from Europe, North America, and New Zealand. Ten position papers were invited for presentations:

- Rene van Horik: How can research data archives benefit from digital preservation research?
- Peter McKinney, Lynn Benson, Steve Knight: From Hobbyist to Industrialist. Challenging the DP Community
- Rainer Schmidt: Collaborative Preservation Infrastructure
- Fiorella Foscarini,Gillian Oliver: The Information Culture Challenge: Moving Beyond OAIS
- Claudia-Melanie Chituc: Requirements Engineering Research and Long Term Digital Preservation
- Jose Borbinha: The value dimensions of Digital Preservation
- Dave Tarrant: Let's get physical!
- Natasa Milic-Frayling: What is a digital object, really?
- Simone Sacchi, Karen M. Wickett: Taking modeling seriously [in digital curation]

[14] The workshop website can be found at http://digitalpreservationchallenges.wordpress.com/
A third-party report about the workshop can be found at http://www.ncdd.nl/blog/?p=3117

- Kresimir Duretec: When is my information going to die?

These presentations set the stage for a day-long intense discussions on six tables in rotating involving about 60 digital preservation experts from very diverse fields. The tables were chaired by the organisers as well as Natasa Milic-Frayling (MSR Cambridge), Rainer Schmidt (AIT), and Brian Matthews (STFC).
While some of the tabled discussions are closely related to the challenges identified above, the discussions were intentionally somewhat broader and more inclusive to ensure unbiased discussion and elicitation of stakeholder challenges. The topics in particular were

7. Digital information models
8. Value, utility, cost, risk and benefit
9. Organizational aspects
10. Experimentation, simulation, and prediction
11. Changing paradigms, shift, evolution
12. Future content and the long tail

The following sections will summarise some of the key discussion points for each of these tables, largely based on the original summaries from each table chair. The questions raised are clearly complex and can provide impetus for further thought throughout the SCAPE project and beyond. It is further envisioned to continue this workshop to evolve into a series, and to feed the content of the challenges into the *DP?* Research challenges wiki [8].

3.1 Digital information models[15]

Introduction

Our objective is to reflect on the general aspects of information and information models in order to understand whether there are any inherent properties of information that need to be taken into account when devising strategies and methods for preserving digital media.

Furthermore, our considerations of digitally encoded information may involve tacit assumptions, inherited from our dealings with the physical information artefacts that are not appropriate for addressing digital preservation. Uncovering and assessing how these assumptions affect our practices and decisions are important steps towards creating effective preservation methods.

Discussions of information models revealed multiple-perspectives at all levels, from the basic concepts and terminology such as the meaning of information and information objects, to the issues of information meaning, context, and structure.

In the following sections we discuss several aspects of information models, including

- What information is and how it relates to information objects and knowledge transfer
- Information, metadata, and format of information resources (structured data vs. unstructured content).
- Shared meaning and interpretation of information
- Expectations in information management, archiving, and preservation.

Basic Notions

It is important to differentiate between information and the objects that encode and transfer information. Information is not a thing. Information is propositional. It results from observations.

[15] This section is based on input by the table host Natasa Milic-Frayling.

Indeed, data and assertions lead to information. In fact, the information is often contained in metadata and the boundary between data and metadata becomes blurred.

Information does not decay. Those are the properties of the information 'containers'. Similarly, preservation does not apply to information. However, information can be lost.

Can information live without embodiment? Is the persistence of information its inherent property? Is transmission of information its inherent property? No, they are not. But information in the digital form is persisted and often encoded with the expectation that it will be transmitted and shared.

In what form is information consumed and perceived? One proposal is to consider information in terms of information objects where the meaning of the term "object" is in the broadest sense, as an entity that can be named and referred to. Encoding and transfer are fundamental aspects of information and they relate to the properties of the information object.

Generally, there are at least three aspects of information: physical, syntactic and semantic. Each of these aspects is complex and diverse. One may try to create a unified model of information that cuts across these three layers. However, it is expected that such an exercise would be substantial yet of little practical use because of its complexity.

Nowadays we observe increased tendency towards structured forms of information. On one side, the users workflows are becoming less structured—individuals use information from different resources in a rather flexible manner. On the other side, the traditional documents that are compact and self-contained are now replaced with structured information forms that are linked to databases and other resources.

Adoption and preference for new information technologies and media have implications for the ways information is transferred. For example, there is a trend towards using video recordings to convey information and a question arises how effective that medium is for acquiring knowledge. Studies have shown that it is not the format but the way the information is absorbed that affects our conceptual models and thus our knowledge. Enabling rich interaction with digital content is an important aspect of knowledge transfer through digital media.

Management of Information

Technological advances have enabled us to create, store, and transfer information at unprecedented speed and scale. However, they have also raised expectations that all digital information can and should be stored.

At the same time, there is a sense of different value attached to different types of information. One view is that the value of information cannot be determined in the absolute sense. An information object is always interpreted within the context. Similarly, its value is dependent on the audience and the context of use. Producer of information creates information with an intent which may or may not be well aligned with the objectives and motivation of the information consumer.

Deciding the importance of information is a challenge. We have already seen examples of unexpected reuse where combining seamlessly useless information into a valuable combo provides new value (e.g., in the climate change studies). Thus, one approach is to store and preserve as much as possible. However, that puts a strain on resources.

At the moment, there are no standards for preservation practices and, therefore, the decision process is often implicit, reflected in the action or no-action, rather than prop. Perhaps, it is still too early to make recommendations for the preservation practices. Throughout history, we can observe emerging phenomena that occur through spontaneous engagements rather than planned actions. We may be still in the early stage, learning as we go, until the requirements and specification

crystalize. Until then, inaction is essentially an implicit decision for no planned presentation at a given point in time.

A practice that has already emerged involves eliminating data that can be regenerated. For example, for a given collection, it may be advisable to provide metadata about the information objects it comprises. Such metadata is generated computationally. Since it can be computed on the fly, no attempts are made to store it, particularly for data that changes over time. At the same time, such metadata may be essential to access the collection and therefore needs to be enabled at all times. A decision not to keep a copy of the metadata but, instead, preserve a program that generates it, suggests an approach to where preservation is achieved without data persistence.

Besides the technical aspects, there is the issue of responsibility and accountability. In effect, in many instances it is not clear whose responsibility is to preserve the produced data. Scientists and their organizations are leading in the requirements for storing raw data but are lacking policies on preservation.

Models of Digital Information

The fundamental question is where the modelling of information should happen. It seems that, in discussing preservation issues, we assume the models that we inherited from the physical world. Are they applicable to digital?

Persisted aspects of digital, the files and the software, are used to create a digital artefact that is experience by the user through rendering, audio, or tactile interfaces. In the case of simple documents, we are mostly dealing with document formats and format rendering. The information object is created through software and the human interaction and experience of the object are mediated through technology.

The tight dependence on software raises the concern about our ability to characterize the digital artefacts. Namely, there is a danger that we cannot document the specifications of the involved software. Thus we cannot perform appropriate verification of the displayed information.

Particularly critical are the requirements on the authenticity of digital documents. The question is how to define authenticity and characterize its degradation and loss. Furthermore, could we use redundancy to help us with the information transfer? Could redundancy help with the recovery of information? Could the same principles be applied to preserving authenticity?

Information and Shared Understanding

It is not surprising that, throughout the history, a prolific encoding of information is aligned with the development of human societies and the need to establish rights and enforce responsibilities. Early documents were essentially establishing the shared context and understanding of the rules and rights of individuals.

Digital medium has increased the indirectness and the level of abstraction expressed in information artefacts. Generally, one has to be able to decode the symbols that are used to convey information in order to begin to interpret them. Understanding higher levels of abstractions, e.g., communication in social networks or consuming aggregated data from social analytics requires shared background. Unfortunately, with many new technologies, the shared experience and understanding are still in infancy. At the same time, at the lower level, the encoding and rendering of information to create a digital artefact are based on agreements within different layers of technologies.

The question arises whether, in the effort to preserve information artefacts, the digital preservation experts should take responsibility for interpreting the meaning of digital information. One view is to

defer that responsibility to the users and focus preservation activities on enabling the reuse and repurposing of information.

Concluding Remarks

Any attempt to create information models for digital media would need to take into account specific aspects of digital. First, the general issue of subjective vs. contextual meaning has to be carefully addressed to ensure that adequate context is captured for digital media.

Furthermore, we need to clearly separate the aspect of the digital embodiment of information from the intrinsic aspects of information. Most of the preservation efforts have dealt with the persistent parts, the files and software. However, the ultimate objective is the successful transfer of information.

Finally, our observation that information resources are becoming highly structured suggests a trend from content towards information. This leads to another important aspect: if the ultimate goal is transfer of knowledge, and the knowledge acquisition is primarily determined by the action and engagement, how digital information should be presented to encourage user practices that enable effective assimilation of information.

3.2 Value, utility, cost, risk and benefit[16]

Appraisal is seen as one of the major upcoming problems dealing with the assessment and prioritization of factors like value, benefit, and risks for digital content. The fact that - in contrast to physical materials - it will not be possible for an institution to keep all deposited digital materials within the archive is causing a change of the preservation policy for many institutions. While data is commonly treated as equally important, once it has passed the appraisal process, it is very difficult to assess the value of digital information before as well as after ingest.

Economic value, for example, is typically short lived and does not provide appropriate selection criteria for a preservation archive. An important aspect is therefore to consider the material as well as immaterial value of information for its appraisal. Value, however, is highly dependent on the designated user community, the memory institution and its mission statement, and/or the parameters of a value system in general. Examples are cost, cultural value, relevance for a society and its identity. The biggest challenge here is to deal with uncertainties that lie in the future. Examples are changes within the memory institution (like ownership, funding, organizational and technical processes), the designated user community, and/or global developments.

Another difficulty for appraisal is to estimate the change of value over time. While information that is presently top ranked (e.g. by a search engine), might lose its value dramatically over time. However, often information gets value because it is being archived and available for future research. In general, it will be important to assess the (anticipated) value of data based on defined and comparable models. This will provide the bases for future evaluations and improvements of these models. The discussion on assessing the value of information ended with the following statement: “Value is relative, dynamic, and contextual”.

Risk is also seen as a multidimensional issue in the context of digital preservation. Risk is in general associated with sustainability. However, in order to asses risk in general, it will be important to distinguish between risk factors that pertain to the data and those pertaining to the institution. Cost, an inherent factor of digital preservation, plays an important role in this respect. In order to ensure sustainability it will be important to develop corresponding cost models for digital preservation. One

[16] This section is based on input by the table host Rainer Schmidt.

important goal is to budget digital preservation activities based on informed economic decisions. It will be important to move away from one-time investments and budget preservation based on long-term and running costs. This implies that digital preservation is an on-going activity, which – in contrast to storage costs – cannot be measured in magnitudes of Bytes.

Arising research questions include:
Digital appraisal strategies: Comparison of a selective approach (creating collections) with one that is based on sampling (creating representations), e.g. for long-term Web archiving. The present (traditional) understanding of institutions is that that collections are of higher value than samples. Also, should memory institutions collaborate with Internet companies and compare their archives, e.g. for assessing coverage?

Value of digital information: Develop a survey to understand the value generated from existing digital collections. Is periodic re-appraisal required to understand (and rank) the information contained within large and complex archives? How can this be supported by tools capturing context semantics of the content (e.g. similar to a citations index for academic papers)?

Commercial value of digital preservation: Commercial companies, for example in the medical and financial, sector may be interested in preservation their (research) data for commercial reasons (e.g. measured by the ROI they gain). A report on cost models is presently being prepared by the APARSEN project. What are the drivers for commercial institutions to invest in long-term digital preservation?

Cost of preservation: It will be important to compare different cost models (developed by different institutions) after certain periods (e.g. after 10 years) and to evaluate how well they have done. How do quality aspects (e.g. resolution) affect preservation costs? Study and compare parameters like cost for machinery, storage, digital curation, environmental costs. Compare the cost of digital archaeology (manual reconstruction) to digital preservation.

Preservation models: Evaluate different digital preservation, management, and threat assessment models/methods over time based on the development of concrete institutions, like for example the National Library of New Zealand. Examples are models like DRAMBORA, TRAC, ICPSR or methods developed by the PARSE.Insight project.

3.3 Organizational aspects[17]

The discussion focused on organizational aspects of digital preservation, including (but not limited to) governance; assessment and improvement; policy; control and decision making; managing risk, value and cost; capabilities, competences, skills and expertise; and changing paradigms and roles.
Contributions to the conversations can be broken into ten broad categories:

1. Characterization and comparison of institutional factors
2. Vocabulary
3. Business models, marketing and incentives
4. Collaboration models and strategies
5. Arrangement of work across institutional boundaries
6. Digital preservation outside the context of large centralized repositories
7. Understanding and facilitating tool adoption
8. Managing organizational change
9. Workforce and staffing issues

[17] This section is based on input by the table host Christopher A. Lee.

10. Process definition and improvement

Characterization and comparison of institutional factors

Participants expressed interest in identifying the characteristics and communication mechanisms/strategies of organizations that have successfully preserved digital objects over long periods of time. These could potentially be compared with organizations that have had demonstrated problems or losses (though data from such cases can be difficult to obtain). A related investigation would be to identify the main factors that are inhibiting some organizations or individuals within organizations from more actively engaging in digital preservation activities.

There was recognition within the groups that the digital preservation community would benefit from more formal was to characterize and compare the social and organizational contexts within which digital preservation is being carried out. In order to analyse and compare digital preservation settings, there is further work to be done on identifying the units of analysis and levels of granularity that one is applying in a given case. One participant suggested that digital preservation research could benefit from applying the concept of design patterns – asking what design patterns for digital preservation are shared or could be shared across organizations. Another participant suggested that, in addition to the OAIS Reference Model, perhaps we need another reference model that more thoroughly elaborates the organizational setting in which digital preservation is performed.

Vocabulary

Despite having made significant progress over the past decade in defining common vocabulary for describing digital preservation processes and entities, language issues persist. This is due in large part to the diversity of institutional and professional context in which terms are being applied. A related study could be an investigation of how the terminology of the OAIS Reference Model are being interpreted and applied in different places – are there any patterns in these applications of the terms (e.g. perhaps social science data repositories applying them differently from art museums or manuscript repositories). Another idea that came out of the group discussions was better identifying what vocabularies/terms are appropriate for different purposes, i.e. not just (descriptively) what terms are being used in different settings but (normatively/pragmatically) what terms are most useful for meeting given goals.

Business models, marketing and incentives

Digital preservation requires an on-going investment of resources, so mobilizing those resources is an essential aspect of digital preservation work. One of the discussions was related to investigating the motivations and implied behaviours of different stakeholders – academics, industry, cultural institutions – and different disciplines; not all parties will have incentives to carry out the same activities, e.g. developing prototypes vs. providing on-going services, and it would be beneficial to have a more solid empirical understanding of who is likely to be doing what parts of the work.

Several discussion participants raised the importance of better elaborating business models for digital preservation, as well as models for marketing the value added of digital preservation. One participant pointed out that archivists have had difficulty conveying to vendors what specifically they should change about their software in order to better support digital preservation goals, which raises not just the question of how to express the needs but also why archivists have so much trouble with articulating them. One of the challenges that would benefit from further research is how to demonstrate the value of digital preservation in a short timeframe. It is also important to understand the incentives for information creators to engage in better curation of the information. For example, professional discussions of electronic records management often focus on the risk of lawsuits; but does fear of lawsuits actually serve as an incentive for digital preservation?

Collaboration models and strategies

Digital preservation is a multifaceted endeavour that can benefit greatly from collaboration. Group 3 discussed various aspects of collaboration. A major research question is what aspects of digital preservation parties can actually collaborate on, and which aspects do they simply need to carry out locally, subject to the requirements and constraints of their own contexts. Another important area of investigation is models for collaboration or networks with disparate contributions, e.g. small vs. large institutions that are replicating each other's data.

Arrangement of work across institutional boundaries

Closely related to issues of collaboration are questions about how to arrange digital preservation work that spans institutional boundaries. This includes how to organize roles and responsibilities across institutions and across nations. A related question is what the proper mixture should be of institutional investment and more distributed curation (e.g. by users)? The group also discussed implications of carrying out tasks when data are not stored locally (e.g. processing data stored by a cloud provider) and other noncustodial arrangements.

Digital preservation outside the context of large centralized repositories

Digital preservation activities must be carried out in all sectors of society and by a diversity of individuals. Not all of those activities will take place within repositories that are dedicated to preservation as part of their core mission. Digital preservation goals could be advanced significantly with more understanding of how are business processes are carried out (and resulting data generated and managed) in organizations that do not do DP for a living.

There are a variety of questions that focus on institutions that are devoted to preservation but have limited resources. What is the right infrastructure for small libraries, archives and museums to do digital preservation? What are the appropriate arrangements for digital preservation in developing countries? Assuming that many organizations will rely on outside parties to perform at least some of the digital preservation activities, what are the appropriate arrangements for "digital preservation as a service"? What are the appropriate social/institutional arrangements for the preservation of personal digital artefacts?

Understanding and facilitating tool adoption

There are many existing tools that can potentially support and enhance digital preservation activities. Participants discussed the importance of identifying how and why some tools are diffused and adopted, while others are not. Why are existing tools often not being used in Producer environments? It would be beneficial to develop better models and methods for identifying existing tools to perform given tasks (e.g. mapping an existing application/microservice to a given set of desired actions) rather than reinventing them.

Managing organizational change

Many of the group discussions related to organizational change and change management. Participants expressed the desire to have better models for continuity of operations, including managing cultural change. Related questions are how to optimize organizational transformation, and whether transformation processes are completely different across organizations (or instead have commonalities across contexts).

Workforce and staffing issues

There has been considerable progress in digital preservation professional education and identification of relevant competencies, but there is still a great deal of room for further research. The group discussed the desirability of having better staffing models, in order to know to classify people based on skills, competencies and responsibilities. One participant proposed an investigation of whether people who were hired into given jobs were still performing the duties identified in the job postings some time (e.g. one year) after they were hired, and if their job duties were significant different, what accounts for the changes.

Process definition and improvement

Some of the discussion related to definition and improvement of specific types of digital preservation processes. One participant was interested in determining how to audit change of data within large, diverse collections. A related goal would be defining processes for validation of digital objects after they have been transformed. Finally, there was a discussion of developing better automated or computer-assisted methods for implementing appraisal, retention and disposition actions.

3.4 Experimentation, simulation, and prediction[18]

Starting from the observation of a workshop participant that "*digital preservation is striving to become a science, yet does in large parts not yet behave like a science*", this topic analysed the models and goals of experimentation, model building, simulation, prediction, and hypotheses testing as some of the key aspects that characterise systematic scientific approaches.

As an interdisciplinary field, experimentation in digital preservation will take many forms. This can range from psychological experiments on how individuals perceive difference, in order to analyse the notions of authenticity and content value, as well as questions of cultural differences, to the level of large-scale simulation and prediction that can be found in fields such as meteorology and physics. Clearly, the rigor that can be found in the latter is a challenging aspiration and may in fact not be an appropriate goal to strive for. Quantitative research also needs to be always embedded in a context, with a clear understanding of the stakeholders and goals.

A lot of research and development in digital preservation (and in fact a key goal for SCAPE) is focused on improving the reproducibility of experiments. This can also be seen in a number of the research goals in SCAPE.

However, to a large degree, experimentation in our field is still in an arts-and-craft stage or, as one speaker at the Research Challenges workshop put it, in the "hobbyist" stage. Moving it to an "industrial" or more scientific stage requires fundamental advances. We observe that well-founded and formally grounded verification and validation approaches are almost non-existent to date. If we aspire to "compare apples with apples", with known variables, we need to develop a more systematic and coherent approach and address a number of key building blocks.

On the one hand, a clear understanding of the different levels and dimensions of experimentation is required, ranging from a micro-level of digital objects and their constituent "information elements" to the macro-level of content collections, organisations, technologies, and communities.

It is clear that the community direly misses proper frameworks for benchmarking on all these levels. Going back to the dictionary, we see *"1: usually bench mark: a mark on a permanent object indicating elevation and serving as a reference in topographic surveys and tidal observations. 2 a: a point of reference from which measurements may be made. b: something that serves as a standard by which others may be measured or judged. c: a standardized problem or test that serves as a basis for evaluation or comparison (as of computer system performance) "*[19]

Key building blocks for benchmarking hence need to (at least) include

[18] This section is provided by the table host Christoph Becker.

[19] http://www.merriam-webster.com/dictionary/benchmark

- A clear, unambiguous understanding of the **processes** that shall be "benchmarked",
- A clear set of **metrics** and indicators for taking measures,
- A well-defined **value system** for judging and assessing measures,
- Solid **hypotheses** that can be tested and falsified,
- Public, open available **data sets** that can be shared and referenced,
- **Ground truth** that annotates these data sets with useful measures corresponding to the metrics above, something which is almost entirely absent in the data sets currently available,
- A means for **publication** of benchmark results and all of the above elements.

A key focus today, and an envisioned focus and starting point for future experimentation, is clearly the verification of correctness of tools and processes such as migration, emulation, interpretation, characterisation, etc. This is maybe the most active area of R&D in digital preservation today, but is hindered by a combination of barriers that still prevent us from achieving the above-mentioned building blocks:

- Legal: constraints on sharing existing data,
- Technical: It is technically challenging to develop robust benchmark data
- Economic: Resource constraints on data collection, annotation, sharing, and developing systematic and coherent approaches
- Fragmentation: The lack of a central reference point or body to coordinate such benchmarking.

However, it is only with the above that we can start not just to build models, but also to systematically test and falsify them in order to enable *better* models. This is clearly an area of research that requires a long-term perspective, since early models will inevitably be too crude to yield accurate and useful predictions. However, it is only through such systematic approaches that we can truly advance the state of art, rather than merely trying incremental product development and testing for immediate problem solving.

On a pragmatic note, as one participant put it in the workshop, *"We want carefully variant crap*", i.e. data sets of objects that are carefully designed, and thoroughly annotated, to violate specific constraints and produce specific expected behaviour. These would represent meaningful test data sets that can be used to benchmark correctness, completeness and robustness of existing preservation tools and processes. The rationale behind such testing is that only through systematic validation, improvements can be requested, delivered, evaluated, and approved.

Carefully annotated test data collections, however, are just one of the building blocks of experimentation. Systematic experimentation and simulation should include (as a non-exhaustive suggestion):

- *Loss*: Can we simulate loss on different levels? How do users recognise, locate, perceive and assess loss?
- *Authenticity, significant properties, and equivalence*: How do we need to conduct experiments on authenticity - can we combine and correlate the equivalence of an "information artefact", the meaning of an information artefact, and the representation? Can we model the cause-effect relationships between these to build predictive models?
- *Aging*: Can we invent mechanisms for accelerated aging, mechanisms that simulate accelerated decay processes?

- From the perspective of users, the predictions that would provide tremendous value include questions such as the following. When is this object going to die?
- What are the total costs of ownership (TCO) of keeping this object alive for X years?
- What are the expected costs of preservation at time X?
- What are the marginal costs of increasing life expectancy by X years?

- At time X will my collection be preserved successfully?
- Can we simulate relationships between funding streams and project level results, enabling content holders to claim "If you cut this budget by X, object Y will be in accessible in Z years“?

Several of these questions point to the life expectancy of objects. The comparison is made between life expectancy of digital objects and of living beings. In Actuarial science, "actuary tables" are used for example to calculate life expectancy based on a number of factors. The question arises: What is the equivalence of this in preservation? What are the genetic, social, economic, cultural, and other factors contributing to life expectancy? And when can we say that a piece of information is "dead"? Obsolescence needs to be understood in a contextual form: An object is obsolete to a user if the user does not possess the means to access the content of that object. However, if it is possible to *recover* these means, it becomes accessible to the user. The question, most of the time, will be that of the costs of doing so. Hence, we also acknowledge the limitations of such allegories.

It is clear that experimentation, simulation and prediction have yet a long way to go in this field, but also hold tremendous insights. Much more systematic approaches are required, and there is a clear need to build, publish and falsify models that attempt to explain key cause-effect relationships of the phenomena in the domain.

3.5 Changing paradigms, shift, evolution[20]

The theme of this area was the change in the DP landscape, emphasised by two parallel changes identified by speakers:

- A change in preservation objectives, from a reactive capture of endangered content, through aware preservation process to keep identified content safe, through to adding additional value through aggregation, data mining, additional context.
- A change in the professional and organisational approach from an amateurish trial and error approach, through an artisanal approach, with centres of excellence managing DP for their own purpose based on common tools and share best practice, to an Industrial approach, based on division of labour, service based provision, and with quality assured processes.

A number of key topics were discussed: Big Data, Adding Context for value-added services, Standards, Tools, and Professionalization.

Big Data

The changing business world in computing has led to an explosion in “big data” services, provided by cloud service providers, with effectively unlimited capacity, and added-value web based content service providers, providing context aware services in particular domains. This increase in capacity and expertise means that data management in some circumstance can be delegated to these commercial organisations. These service providers have quality standards of replication, integrity and backup, and in the case of contextual aware service, enough knowledge of the data application, that they can be relied upon to provide a quality data management solution.
This leads to the question that if it is all big data now, what added value is there in DP? After all, we should stop doing research in areas where companies have been doing it for years. This leads us to consider:

- How can we provide specialist context based services which may be of less interest for commercial providers?

[20] This section is based on input by the table host Brian Matthews.

- How can archive holder maintain trust in such an environment: e.g. control, privacy, ownership, integrity?
- Added value services, e.g. Data Mining – preserved data needs to be accessible via computational means.

Identified research challenges include

- Infrastructure for DP which includes cloud technology, with added value for Preservation (e.g. context awareness, integrity, privacy etc.)
- Cross cloud services – federated clouds, moving data from one cloud service to another
- Computational processes (e.g. searching, aggregation, data mining), across cloud providers.

Adding context for value added services

As preservation practice evolves through the maturity phases, more context for digital resources needs to be added. Such context allows the archive to evolve from a trustworthy and aware, but essentially passive object store, to one which provides added value services in terms of data access, analysis and reuse. Such contextual information would include

- Context of creation: in what environment and assumptions was the object created.
- Preservation intent: why is it
- Information on the Designated community

Such context may not always be explicit, leading to a need to capture the tacit knowledge of the community, including culture and social factors, and factors which are specific to particular application domains. This may require anthropological and social studies, and then the methodology to feed such studies to influence preservation practice.

Another aspect is the value of preservation for a community, and how to make the case for preservation within a community? This is especially important outside the context of a memory institution, in a business context, where the case to preserve is not clear cut, and indeed, risk management may lead to data destruction. There needs to be consideration of the added value in terms of a business case, in terms that different communities understand and appreciate. For example, data interoperability, aggregation and repurposing in new business contexts may be attractive. Further there it may appropriate to have different levels of quality in preservation for different business cases.

In a web based world, the value and context of preservation more dispersed, and it could be argued that there is an evolutionary context in the most literal, Darwinian sense. Useful digital resources will be kept and replicated on the web by parties that value them, while less useful objects will die out. Future scenarios of preservation via crowd sourcing on the web could be explored.

Identified research challenges include

- Novel ways of collecting contextual information (video, crowd sourcing, automated)
- Collecting tacit knowledge of context and designated community
- Maintaining information on the designated community as it changes over time
- Accommodating the social and cultural assumptions.
- Sharing contextual information to allow interoperability and access.
- Value and business case of preservation in non-memory institutions.
- Web-based preservation.

Standards

As preservation matures, there is a need for standards to be established and shared. OAIS forms a good grounding of common terminology. However, there is a lack of agreement on what terms mean and especially how this rather specialised vocabulary translates into different contexts. In these cases, we would want local interpretations in different domains and for specialist practitioners (e.g. Archivists, librarians, IT specialists, domain specific curators, user communities).
OAIS also only covers some core areas of the archival information system. Other areas such as: Pre ingest, organisational information, domain specific information and much context information are not so well covered and need further exploration and standardisation.
This would also need to reconcile a different between much academic theory, taking a holistic and idealised view of preservation, with a pragmatic bottom up approach where complete preservation may not be achieved, but good enough is indeed good enough.
Identified research challenges include

- Providing profiles of OAIS and other standards appropriate for specific practitioners and communities, in a language appropriate to those communities.
- Providing approaches and standards to cover other aspects of preservation not covered by OAIS.

Tools

As part of the industrialisation of preservation, so that it become standard practice to consider the integrity and maintainability of digital objects across their lifespan, more support for digital preservation could be provided as standard practice by the information technology which is used., but getting the tools to more as standard practice. For example:

- Preservation aware operating system and file stores could maintain the integrity of objects, manage file formats and systematically manage provenance.
- Storage systems could be "AIP aware", maintaining dependencies with representation information.
- Management information systems, customer relationship systems and other common business systems and tools could maintain more contextual information and be used within a preservation context.

Further Digital Preservation could be considered as part of best practice of systems development, and managed in such a manner, thus leading further to the Industrialisation of DP. This would modify the process to build good DP into the information management system, thus making it natural and ubiquitous
Further tools can add value to the preserved data, including tools for: aggregation, data mining and exploration, data browsing and access.
Identified research challenges include

- Building and modifying standard tools to make them more Preservation aware to provide a higher integrity computing infrastructure
- Developing software engineering processes and best practise to make them more preservation aware.
- Developing easy to use tools to provide added value services on top of preservation archives.

Professionalization

As the domain evolves from being based on preservation to a value-based business case, the profession of digital curator also has to develop. Is such a role part of a library/archivist role extended to digital objects, or are there special skills over and above this? Is there a new profession

of Digital Conservator and what would be the additional skills? What would be there career path? There is a distinct split between the differing expertise of computing scientists, Library and Information Scientists, Archivists and domains specialists. If we are to get the added value from digital preservation from interoperability, data mining and aggregation, then we will need a combination of skills. As a consequence we will need to change the nature of preservation education from the traditional library and archivists setting to include computing and information technology, business information systems and indeed as an integral part of research methods for all disciplines.
Identified research challenges include

- Developing the profession of information preservation specialists.
- Changing the educational profile of preservation from the library and archiving courses to a broader range of disciplines.

3.6 Future content and the long tail

The long tail phenomenon consists of a number of dimensions. Apart from the long tail of objects, we have to consider the long tail of storage, location, access, and use, as well as the long tail of approaches, philosophies and views.
It is observed that we are moving back from a clear separation between hardware and software to a much tighter coupling (content built for specific devices, tight coupling of hardware and software, ebooks, eScience Data tightly coupled to machinery, sensors and visualizations, art based on sensors and actuators, ...). This is completed by a trend towards a much tighter coupling between application and digital object: It is no longer easy to separate these. Consider the increasingly omnipresent apps and emerging content that is targeted differently for different applications, or the fact that traditional content-oriented forms such as books or magazines are becoming apps as well.
This kind of "appification" causes huge challenges in terms of rights management (DRM), while counter-acting standardization and challenging economics of scale.
How can we prepare for this increasing flood of heterogeneous objects? Can we push back towards a nicer separation of data and processing? Can we model hardware and software simply as nice logical abstractions and transformations, thus clearly separating data from processing logic?

We may need to shift our DP focus from data- centric to process-centric: As we are not always able to keep multiple versions of (large) data sets that are continuously integrated, transformed, analysed and mashed-up across analytical processes, we may be forced to keep the transformation/integration processes instead. How can we preserve these? This leads to questions concerning process preservation and software preservation rather than the challenge of numeric data preservation.
How can we preserve processes? What are the boundaries of a process? How to handle the human in the loop of a process? How can we evaluate if preservation of a process was successful? What are the significant properties of a process rather than an object?
It will, however, also help us in answering questions of versioning in data.
How do we handle the shift to virtual research centres? In the traditional setting the physical lab book served as evidence, how does this work in virtual labs?
However, we may need to move even beyond processes (which are linear), toward an experience-centred holism. This raises the fundamental question of experience: How can we capture this? Where are the boundaries?
We observe a shift from the possession of physical objects to being able to access them, somewhere in the cloud. This is starting to hit severe limitations, partially due to legal/regional characteristics of

clouds (There is no guarantee that data is not moved out of a specific jurisdiction.) Thus, local storage seems to be gaining importance again.
Do people really worry only about accessibility rather than owning it? Is "possessing" an outdated model? Will it be coming back? Can we trust the digital continuum?
But also: Many users no longer have files, since they are only part of the application. If there are no files to maintain, what can and should we collect and preserve? How to select it?
When defining what to collect, we need to establish clear goals also in terms of coverage. What is completeness? How can we establish what level of coverage we have achieved? Who has the responsibility? What different types of information are there and where is completeness an important criterion? (For some data and tasks, sub-sampling is perfectly fine, for others anything but complete information is useless.)
With the new challenges in collecting information, does the deposit model still hold? Can it work? Will there be only information brokers? ("Grey literature" doesn't exist anymore, as everything is considered published.)
The globalization of content also means that nationality/regions-based approaches for collection will not work anymore (legal, collection policy, responsibility...). How do we need to adapt policies/responsibilities to match these new setting of globalized information? Should we shift to discipline-specific? But then, many projects don't fit into a single discipline. How to ensure coverage of all "areas", ensuring nothing is left out?

Considering disciplines, the DP research community maybe suffers from a lack of geographic coverage: what about other philosophical backgrounds? Might that lead to different approaches to DP if we integrated views on preservation, value of information and forgetting, as well as means to do preservation if we included philosophical perspectives from Africa, Asia,...?
What is important? How to use it? (As an example, consider a shrine in Japan that is rebuilt identically every 20 years, preserving the object but not the media, and preserving the building skills.)
How to structure content in the new world? (What about ebooks, do they belong to the multimedia department? Currently, every new type of content automatically goes into "audiovisual").
How can we define the boundaries of an object and its context? Do we need to preserve the culture? Do we need to reach out to philosophy/sociology in order to understand DP?
This may apply specifically to the role of forgetting in the digital world: can we model it as natural selection in information preservation? Are there other approaches? What were the losses so far? What losses do we accept/want/need? How to automate abstraction/summarization? Should we have decay based on non-use?
When we consider privacy, the focus is usually on selected projects collecting obviously privacy sensitive data. But in the long-tail, the collection of privacy sensitive information will be increasing, especially when considering the integration of data from different sources, which becomes untrackable. Do we simply have to give up on trying to preserve privacy? Do we have the means to safeguard it? Can random/structured permutation help in ensuring privacy, protecting society from perfect memory, while still serving all goals of information preservation for defined goals? Shall we simply embargo access? (What about value, then? What about incentives for DP?)

Everything is driven by market forces. Is there (can there be?) an incentive for long-term thinking? There are no short-term incentives for long-term planning. Furthermore, the speed-up in technological evolution hinders planning, development, use, and it increases the long tail. There is a limit to how fast humans/society as a whole can adapt to changing environments. Are we hitting these limits?

The overall challenge begins with the question how to measure and understand/describe the long tail: Should we look at the long tail of objects, usages, institutions? How can we understand the full scope of challenges the long tail poses?

4 Digital Preservation Challenges

This section will attempt to explore the four areas identified above and outline possible approaches that should be taken towards each of them.

4.1 Future preservation infrastructures

This research challenge deals with the advancement of preservation environments towards collaborative infrastructures that enable a more open and efficient use of IT resources, data and derived information.

SCAPE is developing solutions that are built on scalable data management frameworks that advance the capabilities of existing preservation systems with respect to robustness, throughput and scalability. However, although a scalable platform will help individual institutions in managing and preserving growing amounts of data, we argue that individually hosted stand-alone systems might not be technically or economically viable in the long term. We question if future preservation archives that are hosted and maintained by individual institutions, will be able to scale beyond very limited (hand-picked) amounts of data before hitting an economic barrier. A more collaborative approach may help to overcome a number of technical limitations (e.g. storage and backup), and to improve the overall scalability and cost efficiency of the preservation environment. We therefore envision Future Preservation Infrastructures that operate across different organizations creating a large-scale, collaborative and distributed preservation environment. Furthermore, we expect that large parts of these infrastructures will take advantage of IT resources provided by commonly accessible and globalized data centres.

Archival storage and backup, however, just provides one aspect of such collaborative environments. Computation provides another major aspect. Preservation environments and the institutions that operate preservation archives are typically limited in the number and complexity of computations they perform against the archived content. Examples are regular checksumming, identification, and migration of archived data. We argue that it will be important to provide means that allow 3rd party users to process archived content in a generic and scalable fashion. This is motivated by the fact that with growing amounts of content it will simply be impossible for a single institution to understand and curate all of the data it preserves. Cloud hosting models on the other hand usually do not meet institutional policies, hindering more economic outsourcing of the housing of the data and systems. The same policies will most likely also prevent collaborative storage and computation. We see an increasing need for Future Preservation Infrastructures to provide generally available services to a broad range of institutions allowing them to remotely preserve diverse data sets with minimal overhead. In order to develop trustworthy preservation infrastructures and services, it will be important to develop policies that are compliant with emerging infrastructure technologies as well as means to validate if those policies are met.

In general, we see an important research challenge in the development of methods that foster and promote the development and deployment of collaborative, scalable, and globally operating preservation environments. The aim of this task is to understand and formulate technical

requirements, restrictions and implications of collaborative preservation infrastructures. It will be particularly important to consider conceptually the applicability of relevant emerging technologies like data preservation networks, computational infrastructures, or cloud hosting models. Work in this task will mainly include the development of a problem statement as well as the identification of primary issues and research questions. Attention will also be given to the benefits, costs and risks that are expected for the involved stakeholders.

As an initial practical step, it is planned to develop a conceptual model for flexible Quality-of-Service (QoS) fulfilment with respect to the execution of Preservation Plans in different runtime environments. The model will also have to take into account the stakeholders preservation needs with respect to organizational issues like complexity, cost and automation of the preservation process. The aim of this model is to make assumptions on the feasibility of executing a plan under QoS constraints within a particular environment. A concrete experiment that evaluates the model based on the technologies developed within the SCAPE Preservation Planning Sub-project and Preservation Platform Sub-project is planned.

4.2 Advanced simulation and prediction models

On a general perspective, simulation is useful when we cannot measure the real outcome of a process. For example, accelerated aging is used for many purposes since it is not practicable to simply wait for a certain time span and then take a measure. Similarly, Monte Carlo simulation is used to enable predictions where multiple uncertainties overlap, and climate model simulations enable weather prediction and forecasting. The question then is, which are the processes we need to simulate to better understand preservation?

The work in this task will directly build up on the work done on simulation in Automated Watch. The simulation environment built in that work package will offer some insights into possible future state of a repository and resources needed to get to that state. This information can be very valuable for content holders. However, the condition to that is the level of confidence: If the simulation models are insufficient to represent reality, simulation results will not be meaningful.

The simulation environment, however, will also provide the necessary framework in which we can build simulations of other aspects of interest. Yet, some of the key aspects are highly complex: Is it possible to model *obsolescence*? Are we able to predict it as a single number, when it is known that it is a result of result of complex changes in many facets of technology? New technologies (like smart phones and tablets) are constantly introduced on the market, and each one will result in new formats. To have a better understanding of potential repository evolution through time, we need models which are powerful enough to represent the world outside of a repository to a meaningful degree.

The watch component developed in Automated Watch will collect valuable data, which can be used as an input to simulation. Information such as the types of files that are ingested and the types that are accessed, and statistics about how repository content changes over time can be used as input to models for predicting future trends. If we know how the ingest changed over the last 5 years, how can we predict it for the next year?

In predictions, it would be possible to go even further. Early electronic documents usually contained only text, while today documents can be complex objects. With the analysis of past documents and how their structure evolved over time, to which degree would it be possible to predict the future? Clearly, this aspect is also challenged by the absence of tools that are good enough to analyse current content in detail.

We will investigate possible approaches to advanced modelling techniques and models that we can use to simulate and predict challenging questions such as those raised in Section 3.4. If feasible, we will attempt to validate such models statistically against historical data collected in Watch and other sources [9].

4.3 Information models and benchmarking

The development and improvement of current characterisation techniques is fundamentally hindered by a non-existence of benchmarks. Annotated benchmark data are needed to support the objective comparison of new approaches and quantify the improvements over existing techniques. This lack of baselines is partly due to the fact that the creation of such benchmarks in the way the problem has been approached proved effort-intensive, and the ex-post annotation of content is impossible to verify automatically.

A baseline benchmark needs to rely on known ground truth. However, for many object types such as databases or electronic documents, this ground truth is never known beforehand, but instead extracted from the objects themselves. Since the variation in objects, their features, and formats and variants is so high, this approach is error-prone, and there exists no safe ground on which to create a baseline for quantitative improvements. The fact that it is currently impossible to quantitatively compare tools and methods means that any incentives to improve these tools are vague, and any improvements to the tools are unsystematic.

Instead of characterising objects taken from real collections, a solid bootstrapping approach could rely on generating test data from a fact base, from a content model with specified properties. The starting point of a document in such a benchmark would not be a file representing it for example in the Word 97 format, but instead the document model as it is created in the text editor by the user.

Correspondingly, automated creation of test data could rely on a domain-specific property definition language and generative approaches such as model-driven engineering and code generation frameworks. This would support the creation of truly 'origin' documents – documents that are created in almost the same manner as if a human user would write them, by essentially simulating the creation process. This could eventually lead to perfectly specified data sets and tackle the challenge of benchmark set stratification, since it could support the explicit and exact configuration of the desired variation of properties. The approach furthermore would make it easy to create representations in different formats supported by one program and analyse the exact variations in the produced byte streams, and supports the inclusion of stochastic processes to vary features. Given the right processes, it should be possible to generate fine-tuned benchmark sets for specific scenarios, where the statistical properties are well-defined and all properties of interest documented in the ground truth. However, this is clearly a complex and challenging task.

A major obstacle is that a large part of the semantics is often contained not in the objects themselves, but only realises itself in a rendering environment. This makes it difficult to arrive at meaningful results by applying merely static analysis means, a symptom which increases correspondingly with the structural complexity of files.

To truly understand the metrics required, however, we need to dive deeper into the nature of digital information, reflecting on and questioning the notions of "digital objects" commonly used as metaphors. These seem to be not only inadequate, but in fact actively misleading in many facets.

Hence, in order to define the notion of preservation for digital media, we first observe that no digital artefact exists without software processing on an adequate hardware platform. In other words, neither persisted files nor a persisted program would suffice without the ability to execute the program and thus instantiate a digital artefact.

In the case of electronic documents, application files contain information that is necessary to display content for reading and editing across application sessions. If the application cannot be activated in the computation environment the content is not accessible. If there is an application that provides similar affordances but, for example, uses a different file format, then one may create a transformer from one format to another.

In order to ensure that the authenticity of the digital artefacts is preserved when a transformed files is used by the new application, we would need to observe the properties of both software packages on their corresponding files. Furthermore, we would need to characterize the format transformation mapping and the effects it may have when composed with the new software application.
We propose to investigate research questions that deal with the notion of fundamental file format properties and the basic software features that manipulate these properties. The fundamental problem is to establish the equivalence between two software packages with the corresponding input formats. We expect that such research questions can be formulated with formal model verification.

4.4 Identification of emerging topics

From the extensive discussions on emerging topics and their manifold relationships, it becomes clear that the identification of emerging challenges, opportunities and research challenges should not stop here, but merely has found a suitable start. Correspondingly, we will continue to engage with the digital preservation community and reach beyond established core groups of research to ensure a continuous dialogue on emerging and future research roadmaps in the domain.

5 Conclusions and Outlook

This document outlined the research roadmap of the SCAPE project. It positioned the research carried out in SCAPE within the European research landscape focused on digital preservation research and outlined representative key goals of the R&D work packages in SCAPE.
Broadly speaking, these goals strive to

1. advance the state of art in scalable preservation components and processes for preservation actions, content analysis, and quality assurance,
2. provide flexible mechanisms for constructing powerful preservation workflows based on such components,
3. advance the start of art in flexible, scalable, distributed parallel execution of such processes based on paradigms such as MapReduce, and
4. provide scalable mechanisms for decision making and control.

Analysing the open issues identified by the R&D work packages in SCAPE, we identified a number of gaps and emerging topics that merit further investigation. These were broadened and discussed with a wide audience in a workshop on Open Research Challenges, organized at IPRES 2012, which received strong participation from the global DP community. Discussions were grouped in six topics, closely related to the above challenges but more broadly defined.

1. Digital information models
2. Value, utility, cost, risk and benefit
3. Organizational aspects
4. Experimentation, simulation, and prediction
5. Changing paradigms, shift, evolution
6. Future content and the long tail

Based on the collected research goals and the broad involvement of the DP community, we identified and outlined common gaps and openings for future research and finally, three emerging critical research topics that arise from the cross-section of identified open problems and point to fundamental research questions. These are:

1. **Future preservation infrastructures**.
2. **Advanced simulation and prediction models**.
3. **Information models and benchmarking**.

We furthermore conclude that it is paramount to continue analysing emerging research topics and challenges throughout the project and beyond. This will also provide crucial input for the final research roadmap that will be delivered by SCAPE in 2014.

6 Bibliography

[1] DigitalPreservationEurope consortium, “Deliverable D7.2 - Research Roadmap,” June 2006.

[2] PARSE Insight consortium, “Deliverable D2.2 Scienece Data Infrastructure Roadmap,” June 2010.

[3] S. Strodl, P. Petrov and A. Rauber, “Research on digital preservation within projects co-funded by the European Union in the ICT programme,” Vienna, 2011.

[4] A. Birkland and A. A. Blekinge, “Re-thinking Fedora's storage layer: A new high-level interface to remove old assumptions and allow novel use cases,” in *The 5th International Conference on Open Repositories (OR2010)*, Madrid, Spain, 2010.

[5] CARLI (Consortium of Academic and Research Libraries in Illinois) Digital Collections Users’ Group (DCUG), Standards Subcommittee, “GUIDELINES FOR THE CREATION OF DIGITAL COLLECTIONS Digitization Best Practices for Moving Images,” 1 August 2010. [Online]. Available: http://www.carli.illinois.edu/mem-prod/contentdm/guidelines_for_video.pdf.

[6] Library of Congress, “Preferences in Summary for Moving Image Content,” 18 November 2011. [Online]. Available: http://www.digitalpreservation.gov/formats/content/video_preferences.shtml.

[7] J. S. Nielsen and B. A. Jurik, “Audio Quality Assurance: An Application of Cross Correlation,” in *IPRES*, Toronto, 2012.

[8] “The DP? Wiki,” Vienna University of Technology, [Online]. Available: http://sokrates.ifs.tuwien.ac.at/wiki/index.php/Main_Page. [Accessed 12 10 2012].

[9] A. N. Jackson, “Formats over Time: Exploring UK Web History,” in *IPRES*, Toronto, 2012.

[10] B. Jurik and G. Elstrøm, “IS13 wmv to Video Format-X Migration Results in Out-of-sync Sound and Video,” [Online]. Available: http://wiki.opf-labs.org/display/SP/IS13+wmv+to+Video+Format-X+Migration+Results+in+Out-of-sync+Sound+and+Video. [Accessed August 2012].

[11] J.-P. Chanod, M. Dobreva, A. Rauber, S. Ross and V. Casarosa, “Automation in Digital Preservation,” Leibnitz-Zentrum fuer Informatik, 2010.

[12] C. Becker, A. Rauber and C. A. Lee, “IPRES 2012 workshop on Open Research Challenges,” October 2012. [Online]. Available: www.digitalpreservationchallenges.wordpress.com.