Evolving Domains, Problems and Solutions for Long Term Digital Preservation

Orit Edelstein
IBM Research - Haifa
Haifa, Israel
edelstein@il.ibm.com

Michael Factor
IBM Research - Haifa
Haifa, Israel
factor@il.ibm.com

Ross King
AIT Austrian Institute of
Technology GmbH
ross.king@ait.ac.at

Thomas Risse
L3S Research Center
Hannover, Germany
risse@L3S.de

Eliot Salant
IBM Research - Haifa
Haifa, Israel
salant@il.ibm.com

Philip Taylor
SAP (UK) Ltd.
SAP Research
Belfast, Northern Ireland
philip.taylor@sap.com

ABSTRACT

We present, compare and contrast new directions in long term digital preservation as covered by the four large European Community funded research projects that started in 2011. The new projects widen the domain of digital preservation from the traditional purview of memory institutions preserving documents to include scenarios such as health-care, data with direct commercial value, and web-based data. Some of these projects consider not only how to preserve the programs needed to interpret the data but also how to manage and preserve the related workflows. Considerations such as risk analysis and cost estimation are built into some of them, and more than one of these efforts is examining the use of cloud-based technologies. All projects look into programmatic solutions, while emphasizing different aspects such as data collection, scalability, reconfigurability, and full lifecycle management. These new directions will make digital preservation applicable to a wider domain of users and will give better tools to assist in the process.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software

Keywords

Preservation, Web Archives, Software as a Service, Business Processes

1. INTRODUCTION

This paper presents the directions of the newly started major efforts on long term digital preservation partially funded by the European Union's FP7 initiative. There are four largest projects (overall budget of above eight million Euro each) funded by the EC on long term digital preservation that started in the last year.

While all four project address digital preservation, they differ in what data are being preserved, how the data are identified, and how the data are preserved. All of these projects consider and, when appropriate, use results of previous digital preservation projects.

We discuss the motivation and objectives of the four efforts, the target communities, and the respective stakeholders. The solutions chosen are presented and alternatives are discussed. By comparing the four projects, highlighting the areas where they complement each other, where they contrast, and what they cover, we are reporting the extent of the current effort within FP7 and their expected contribution to the domain of long term digital preservation.

The four projects are:

- ARCOMEM[1] - From Collect-All Archives to Community Memories - is about memory institutions like archives, museums and libraries in the age of the social web. Social media are becoming more and more pervasive in all areas of life. ARCOMEM's aim is to help to transform archives into collective memories that are more tightly integrated with their community of users and to exploit Web 2.0 and the wisdom of crowds to make web archiving a more selective and meaning-based process.
- SCAPE[2] - SCAlable Preservation Environments - will address scalability of large-scale digital preservation workflows. The project aims to enhance the state of the art in three concrete and significant ways. First, it will develop infrastructure and tools for scalable preservation actions; second, it will provide a framework for automated, quality-assured preservation workflows; and, third, it will integrate these components with a policy-based preservation planning and watch system. These concrete project results will be driven by requirements from, and in turn validated within, three large-scale testbeds from diverse application areas: web content, digital repositories, and research data sets.

[1] http://www.arcomem.eu/

[2] http://www.scape-project.eu/

- ENSURE[3] - Starting with the philosophy that ”one size does *not* fit all”, ENSURE (Enabling kNowledge Sustainability, Usability and Recovery for Economic value) is building on existing tools, processes and approaches to create a flexible, self-configuring software stack. The solution stack will pick both the configuration and preservation lifecycle processes in order to create a financially viable solution for the given preservation requirements, trading off the cost of preservation against the value of the preserved data over time. The requirements and validation of ENSURE are driven by health-care, clinical trials, and financial use cases.
- TIMBUS[4] - Digital Preservation for Timeless Business Processes and Services. The digital preservation problem is well-understood for query-centric information scenarios but has been less explored for scenarios where the important digital information to be preserved is the execution context within which data are processed, analyzed, transformed and rendered. It is this scenario which TIMBUS addresses. The industrial case studies — addressing business processes that include sensor hardware through to large enterprise software services — focus on: (1) engineering services and systems for digital preservation; (2)civil engineering infrastructures; (3)e-science and mathematical simulations.

In rest of the paper is organized as follows: Section 2 presents the motivations of each project, followed by comparing and contrasting them. Section 3 and 4 does the same for the objectives and approaches of the projects respectively. Section 5 discuss related work. Section 6 summarizes.

2. MOTIVATIONS OF PROJECTS

2.1 ARCOMEM

The report *Sustainable Economics for a Digital Planet*[5] states that “the first challenge for preservation arises when demand is diffuse or weakly articulated”. This is especially the case for non-traditional digital publications, e.g., blogs, collaborative space or digital lab books. The challenge with new forms of publications is that there can be a lack of alignment between what institutions see as worth preserving, what the owners see as a current value, and the incentive to preserve as well as the rapidness at which decisions have to be made. For ephemeral publications such as the web, this misalignment often results in irreparable loss. Given the deluge of digital information created and this situation of uncertainty, a first necessary step is to be able to respond quickly, even if preliminarily, by the timely creation of archives, with minimum overhead that would enable later engagement in more costly preservation actions. This is the challenge that ARCOMEM is addressing, relying on the ”wisdom of the crowds” for intelligent content appraisal, selection, contextualization and preservation.

The Social Web not only provides a rich source of user generated content. It also contextualizes content and reflects content understanding and appraisal within society. This is done by interlinking, discussing, commenting, rating, referencing, and re-using content. The ARCOMEM project will analyze and mine this rich social tapestry to find clues for deciding what should be preserved (based on its reflection in the Social Web), to contextualize content within digital archives based on their Social Web context, and determine how to to best preserve this context. The Social Web based contextualization will be complemented by exploring topic-centered, event-centered and entity-centred processes for content appraisal and acquisition as well as rich preservation.

Two application scenarios are used for validating and showcasing the ARCOMEM technology. The first application will target the Social Web driven event and entity aware enrichment of media-related Web archives as they are, for example, required by broadcasting companies. This showcase will be driven by the broadcasting companies Sudwestrundfunk (SWR) and Deutsche Welle. The second ARCOMEM application will validate and showcase the use of ARCOMEM technology for the effective creation of Social-web-aware political archives based on Web archives and other digital archives. This will be driven by the Hellenic Austrian Parliaments.

2.2 SCAPE

The fact that the volume of digital content worldwide is increasing geometrically demands that preservation activities become more scalable. The economics of long-term storage and access demand that they become more automated. Unfortunately, the present state of the art fails to address the need for scalable automated solutions for tasks like the characterization or migration of very large collections. Standard tools break down when faced with very large or complex digital objects; standard workflows break down when faced with a very large number of objects or heterogeneous collections. Even the preservation systems used in the largest memory institutions lack the necessary automated quality assurance tools for detecting and reporting errors in a preservation process, and thereby fail to fully mitigate preservation risks.

SCAPE will use these large testbeds to define its requirements and validate its results. The **Web Content Testbed** highlights the challenges presented by heterogeneous collections and a rapidly changing delivery environment. The sheer volume of content in web archives requires fully automated, scalable preservation solutions. Web content covers many diverse file formats in multiple versions, including obsolete formats, and also associated rendering tools. The **Digital Repositories Testbed** highlights the challenge of carrying out preservation actions within an institutional context where there are legal and policy requirements and substantial investments in legacy systems. Preservation challenges coming from large scale digital repositories include issues of scalability along several dimensions: number, size, and complexity of digital objects as well as heterogeneity of collections. Furthermore, collection profiling is an integral part of planning. Finally, the current generation of digital library and preservation environments are often based on network service oriented architectures that do not scale. The **Research Data Sets Testbed** is concerned with the pressing need for preservation in scientific communities in the face of threats to long-term access and usability of scientific data. Particular aspects of this testbed include potentially very large data sets, wide variety of practices and unique requirements to preserve the original context of the experiment which generated the data in the first place.

[3] http://ensure-fp7.eu/

[4] http://www.timbusproject.net/

2.3 ENSURE

As opposed to the preservation of cultural heritage information, which it is presumed needs to be retained forever, it is neither economically viable, nor in some cases even desirable, to preserve all data managed by business forever. The value of such data tends to decrease over time, although on the other hand, there may be legal or regulatory reasons why aged data must still be retained.

In addition to the business value of data changing over time, the appropriateness of an originally chosen preservation solution can be affected over time by changing regulations, or new advances in underlying technologies influencing price-driven solutions. ENSURE is researching how current Lifecycle Management tools can be used to control the *preservation lifecycle* amidst these shifting conditions.

In addition to examining trading off cost versus quality, ENSURE is looking into the use of emerging ICT technologies to enable solutions which are not only economical, but also capable of scaling over time to meet ever expanding amounts of data. Cloud storage is seen as a primary candidate for the underlying storage services, but this introduces additional challenges, e.g., the migration of data from cloud to cloud, security issues, and the ability to perform preservation-related computing near the storage.

The ENSURE solution will be motivated and validated by three real world use cases. Selected for their relevance to data preservation, the requirements elicited by these use cases cover a wide spectrum of topics, ranging from maintaining data privacy over time, evolving ontologies, and being able to view data stored in proprietary formats decades from now. More specifically, the use cases for ENSURE are:

- Healthcare - where enormous quantities of scientific data are tied to individuals, but managed by an organization controlled by strong regulations for privacy and traceability.

- Clinical Trials - where data has both scientific and business value with strong regularity restrictions, requiring special concern for patient privacy issues.

- Financial Services - which emphasizes the long term retention of data after the regulations mandated period only as long as it has business value.

2.4 TIMBUS

A primary motivation for TIMBUS is the declining popularity of centralized, in-house business processes maintained and owned by single entities. The presence of Software as a Service (SaaS) and Internet of Services (IoS) means business processes are increasingly supported by service oriented systems where numerous services, provided by different providers, located in different geographical locations are composed to form value-added service compositions and service systems which will continue changing and evolving. Besides the advantages of SaaS and IoS, there is the danger of services and service providers disappearing (for various reasons), leaving partially complete business processes.

TIMBUS endeavors to enlarge the understanding of digital preservation to include the set of activities, processes and tools that ensure continued access to services and software necessary to produce the context within which information can be accessed, properly rendered, validated and transformed into context based knowledge. This enlarged understanding brings DP clearly into the domain of Business Continuity Management (BCM). BCM, as standardized by the British Standards Institution (BSI), is defined as:

> A holistic management process that identifies potential threats to an organization and the impacts to business operations that those threats, if realized, might cause, and which provides a framework for building organizational resilience with the capability for an effective response that safeguards the interests of its key stakeholders, reputation, brand and value-creating activities. [8]

2.5 Comparison and Contrast

Obviously, within the context of the EU call, each project has digital preservation as a motivation. However, ACROMEM stands alone in dealing with publically available and non-regulated data and in harnessing the "Wisdom of the Crowds" to help decide what to preserve. TIMBUS focuses on the environments that produce the data rather than the data itself. ENSURE and TIMBUS are motivated in part by accurate risk assessment and preservation lifecycle issues related to regulations. Together with SCAPE, they also address the scalability of technology infrastructure and software infrastructure for digital preservation. While there is some overlap in use cases,the projects as a whole cover a broad cross section of scenarios from tradition memory institutions (SCAPE), web (SCAPE, ACROMEM), engineering (TIMBUS), scientific (SCAPE, ENSURE, TIMBUS), health care (ENSURE), and finance (ENSURE).

3. OBJECTIVES OF PROJECTS

3.1 ARCOMEM

ARCOMEM's goal is to develop methods and tools for transforming digital archives into community memories based on novel socially-aware and socially-driven preservation models. This will be done (a) by leveraging the "Wisdom of the Crowds" reflected in the rich context and reflective information in the Social Web for driving innovative, concise and socially-aware content appraisal and selection processes for preservation, taking events, entities and topics as seeds, and by encapsulating this functionality into an adaptive decision support tool for the archivist, and (b) and by using Social Web contextualization, as well as extracted information on events, topics, and entities for creating richer and socially contextualized digital archives.

To achieve its goal, the ARCOMEM project will pursue the following scientific and technological objectives.

1. **Social Web analysis and Web mining**: effective methods for the analysis of Social Web content, analysis of community structures, discovery of evidence for content appraisal, analysis of trust and provenance, and scalability of analysis methods;

2. **Event detection and consolidation**: information extraction technologies for detection of events and related entities; methods for consolidating event, entity and topic information within and between archives; models for events, covering different levels of granularity, and their relations;

3. **Perspective, opinion, and sentiment detection**: scalable methods for detecting and analyzing opinions,

perspectives taken, and sentiments expressed in the Web and especially Social Web content;

4. **Concise content purging**: detection of duplicates and near-duplicates and an adequate reflection of content diversity with respect to textual content, images, and opinions.

5. **Intelligent adaptive decision support**: methods for combining and reasoning about input from Social Web analysis, diversity and coverage, extracted information, domain knowledge, and heuristics, etc.; methods for adapting the decision strategies to inputs received;

6. **Advanced Web crawling**: the integration of event-centric and entity-centric strategies, the use of Social Web clues in crawling decisions and methods for crawling by example and integrating descriptive crawling specifications into crawling strategies;

7. **Approaches for "semantic preservation"**: methods for enabling long-term interpretability of the archive content; methods for preserving the original context of perception and discourse in a semantic way; methods for dealing with evolution on the semantic layer.

3.2 SCAPE

Based on the challenges confronting its stakeholders, the scientific and technical objectives of the SCAPE project are:

1. **Scalability**. SCAPE will address scalability in four dimensions: number of objects, size of objects, complexity of objects, and heterogeneity of collections. The project is concerned with extending repository software to store, manage, and manipulate larger objects (e.g., multi-gigabyte video streams) and a larger numbers of objects (hundreds of millions). SCAPE will also improve the ability of existing preservation tools to manage a variety of container objects and to recognize diverse object formats.

2. **Automation**. Automated workflows are state of the art; SCAPE aims to make these workflows scalable. SCAPE preservation workflows will be simple to design, making use of the well-known Taverna [19] workbench, and will be deployable and executable on large computational clusters. Automated workflows for quality assurance will be developed to accompany the preservation workflows. The project also intends to introduce automation and scalability to the areas of technology watch and preservation planning.

3. **Planning**. SCAPE will build on the award-winning preservation planning tool Plato in order to enable institutions to answer core preservation planning questions. For large heterogeneous collections, the planning tool should enable a curator to determine what tools and technologies are optimal for preservation within in a given context, defined by institutional policies. SCAPE will also advance the state of the art by delivering a catalogue of generic policy elements and a semantic representation of these elements in a machine-understandable form that can be leveraged by the planning and watch components, enabling automated policy-driven planning.

4. **Context**. In the area of research data sets, SCAPE aims to provide a methodology and tools for capturing contextual information across the entire digital object lifecycle. The advance proposed by SCAPE is to embed migration of scientific data as a preservation action in the workflow, whilst preserving the wider context in order to maintain the reusability of the data. Additional research will be dedicated towards the preservation of software. Software can be seen both as part of the representation information for the scientific data itself, but also requires preservation in its own right.

5. **Prototype**. An important goal of SCAPE is to produce a robust integrated preservation system prototype within the time-frame of the project. This prototype will be made available as open source software. SCAPE technologies are expected to be in productive use in partner institutions by the end of the project. SCAPE components should also be integrated in products offered by the project's commercial partners.

3.3 ENSURE

To meet the challenges that ENSURE addresses, four main scientific and technical objectives have been defined:

1. **Evaluate cost and value**. The value of data over time differs between different organizations and industries. While the design plans for a radio built with vacuum tubes from the 1940's may not have a high business value today, the design plans for the B52 aircraft from the same period, and still in service today, do. As the business value of data goes down, the investment that an organization is willing to make to preserve the data will similarly decrease. Defining the *quality* of preservation as inversely proportional to the risk of losing data, ENSURE will look at ways of balancing the quality of a preservation solution against its cost and the value of data over time. ENSURE will also examine how a configured solution should evolve as the cost of its underlying infrastructure changes.

2. **Preservation Lifecycle Management for different types of data**. Many organizations today manage their data with Information Lifecycle (ILM) tools. ENSURE will research the suitability of adapting today's lifecyle management tools to long term preservation. In particular, while nearly all of today's ILM tools are passive, being driven by other systems and decisions, ENSURE will create a Preservation Lifecycle Management engine which can dynamically react to triggers generated by events affecting the original preservation conditions, such as new regulations, format changes, economic changes, etc.

3. **Content-aware, long term data protection**. For a preservation solution to be acceptable, it must control access to sensitive data and prevent its leakage over time, even though the identities of users, the value of the information, the roles which can access the information, etc. may change. The definition of what constitutes Personally Identifiable Information (PII) may also evolve over time, causing previously valid assumptions of data anomymization to be violated. Additionally, a solution to these issues must scale with the

size of the preservation system, and work environments such as cloud based data storage.

4. **Scalability by leveraging wider ICT innovations**. Cloud Storage and standard virtualization technologies are promising technologies to meet the challenge of building a preservation environment which can expand over time, without having to make large capital expenditures, or encounter spiraling costs for operating expenses. However, today's storage clouds typically aim at providing low cost storage and give few guarantees to the reliability and security of the stored information. A major challenge for ENSURE is demonstrating how a preservation system can be based on such a platform.

3.4 TIMBUS

To support the continuity of business processes, TIMBUS has a number of objectives best viewed from its three stages of digital preservation effort:

1. **Expediency** of digital preservation effort - establishing the risk of not preserving and the feasibility of digitally preserving business processes. Fundamental to determining what should be preserved is analyzing the risk experienced by an organization. Analyzing risk is a complex process requiring many sources of information to be collated and reasoned over. TIMBUS will develop methods and tools that provide an itelligent enterprise risk management (iERM) approach that will support decisions relating to (1) when to preserve, (2) what to preserve and (3) how to maintain and test what has been preserved.

2. **Execution** of digital preservation process - performing the digital preservation of business processes. After the expediency has been established it is necessary to actually execute the digital preservation process. TIMBUS, will address legalities lifecycle management (LLM) and uncover the current legal issues around digitally preserving interdependent services comprising a business process.

   Today's services are deployed on multi-tier service platforms that are not engineered specifically with digital preservation in mind. TIMBUS will address re-engineering existing services for digital preservation (DP) and engineering new services for digital preservation. TIMBUS will also develop verification methods for the digitally preserved business processes which will prove the current preserved business process is valid (to some preservation guarantee level) and also provide some validation of the preserved business process in the (simulated) future. Appropriately, TIMBUS will develop processes for digital preservation of business processes which will be domain specific according to the use cases and processes that are generic for adaptation to new domains. These processes will be aligned with existing digital preservation standards and be the foundation for new standards specifically designed for digital preservation of business processes.

3. **Exhumation** of digitally preserved assets - re-running a digitally preserved business process. It must be possible to exhume and rerun the preserved business process. This issue will be dealt with by the visualization and storage innovations. However, it may still be the case that periodic business process exhumation will be required to provide ongoing guarantees of integrity. Obviously the future cannot be experienced now but TIMBUS must provide some level of assurance that a digitally preserved business process can be exhumed and re-run or exhumed and integrated into future business processes. TIMBUS will simulate technology changes to help indicate process exhumation and integration is feasible.

3.5 Comparison and Contrast

Out of the four projects examined here, three of them (ENSURE, SCAPE, TIMBUS) are organization-focused concerned with preserving in-house information, whereas ARCOMEM's domain is the web. It is therefore no surprise that the objectives for the first three projects tend to be more similar than those for ARCOMEM.

Central to all of the stated preservation projects is the ability to define what data needs to be preserved. ARCOMEM, concerned with preserving content found on the Web, will be looking for how to do this by attempting to analyze the information itself in the context of the Social Web. Amongst the other three projects, both SCAPE and TIMBUS will use tools to help the person responsible for preservation decide what needs to be preserved, whereas ENSURE assumes a set of supplied business rules will give this information. It is interesting to note that TIMBUS's evaluation of what to preserve is driven by the risk of *not* saving information, whereas in ENSURE, while abiding by regulatory constraints, attempts to balance cost versus. In all cases it is recognized that human intervention will be required to come up with the final decision on what to preserve.

Scalability is an issue of concentration in ENSURE, SCAPE and ARCOMEM, although the projects are emphasizing different aspects: ENSURE will tackle scalability in terms of infrastructure support,e.g., supplying a cloud based storage back-end that can support massive preservation; SCAPE focuses supporting a large number of different objects and object types; and ARCOMEM needs to analyze huge amounts of Web content for the content selection and appraisal.

The ability to rerun software after an extended period of time is a focus of the projects, and the use of virtualization technologies is a stated goal of ENSURE and TIMBUS.

The automation of the preservation lifecycle is being dealt with by all of the organization-focused projects. While SCAPE will be creating preservation lifecycles for deployment on large computational clusters, ENSURE and TIMBUS will examine extending existing lifecycle management tools to meet the additional requirements that digital preservation entails. TIMBUS will preserve processes encoded in a lifecycle management tool, while ENSURE, like SCAPE, focuses on the lifecycle management of the preservation process itself.

Additionally, all of the organization-focused projects are concerned with automatic verification of the quality of their runtime solutions. Quality will not only be monitored as part of the preservation lifecycle by all three, but also taken into consideration in the preservation planning stage.

4. APPROACH OF PROJECTS

4.1 ARCOMEM

The envisioned ARCOMEM system is built around two loops: content selection and content enrichment. The *content selection loop* aims at content filtering based on community reflection and appraisal. Social Web content will be analyses regarding the interlinking, the context and the popularity of web content, regarding events, topics and entities. These results are used for building the seed lists to be used by existing Web crawlers.

Within the *content enrichment loop*, newly crawled pages will be analyzed regarding topics, entities, events, perspectives, Social Web context and evolutionary aspects in order to link them to each other by the relationship between events as well as by the involved entities such as persons, organizations, locations and artifacts.

The implementation of the ARCOMEM system is structured into three main research areas. *Social Web-based content appraisal and archive contextualization* aims at the development of methods to analyze the Social Web for getting clues for content appraisal and for extracting information for the archive enrichment. Networks and media are part of a dynamic social process, rather than collections of documents; networks, contexts and meanings co-evolve. To achieve a better understanding of this process for preservation, we need to answer several questions, such as: how do we appraise and rank content in multiple forms and from multiple sources, taking into account the wealth of socially-generated information about the content itself; how is reputation built, who are the leaders and who the followers. etc.

Events, Perspectives, Topics, & their Dynamics aims at extracting information from crawled data in order to provide semantically rich metadata for organizing and contextualizing the archived collection, and for supporting intelligent and efficient crawling strategies. Content perception and memorization are typically focused on, and organized around, events, entities and/or topics. Therefore, these will also be the main ingredients for the semantic enrichment layer for transforming long-term archives into community memories.

Intelligent and Collaborative Content Acquisition Support will focus on intelligent, adaptive and collaborative methods for driving and prioritizing the content acquisition and curation process. The main outcome comprises a prioritized list of sources to be crawled. This decision is primarily based on the relevance, importance, coverage and diversity of the content. This is complemented with an adaptation process involving the archivist or other archive users, and support the collaborative creation and management of archives by communities of curators.

4.2 SCAPE

The approach of SCAPE is dictated by four research and development sub-projects: Testbeds, Preservation Components, Platform, and Planning and Watch.

The Testbeds sub-project is the primary driver of the rest of the project in that it determines the use case scenarios, defines the preservation workflows, and evaluates the platform. The main goal is to assess the large scale applicability of the SCAPE Preservation Platform and the preservation components developed within the project. Using these software components, it creates test environments for the different application scenarios and complex large scale preservation workflows. As part of the testbed evaluation methodology, the automated planning tool will be used to evaluate the strengths and weaknesses of the action components in several scenarios.

The Preservation Components sub-project should address three known limitations of the functional components of a digital preservation system namely scalability, functional coverage, and quality. This sub-project will improve and extend existing tools, develop new ones where necessary, and apply proven approaches like image and patterns analysis to the problem of ensuring quality in digital preservation. Building on the state of the art and focusing on formats and tools that are considered most important by the Testbed sub-project, SCAPE will investigate methods to parallelize and embed components in robust and scalable workflows. SCAPE will provide the ability to capture relevant provenance and contextual information and metadata, as well as the ability to provide usable outputs for automated policy-driven preservation. Finally, SCAPE will develop new methods to automatically detect quality defaults, based on conversion of objects into images to apply image analysis techniques to detect differences resulting from preservation actions.

The SCAPE Preservation Platform will provide an extensible infrastructure for the execution of digital preservation processes on large volumes of data. The Platform sub-project will provide a flexible mechanism for the integration of existing digital repository systems and provide a reference implementation. The Preservation Platform will also provide the underlying runtime environment for large-scale testing and evaluation performed within the Testbed and Planning and Watch sub-projects. The computational layer of the Preservation Platform system will make use of Hadoop, an open-source map/reduce engine, and the underlying distributed storage layer will be based on HBase, which provides high performance and scalable data storage on top of Hadoop's Distributed File System (HDFS) [6].

The sub-project Planning and Watch addresses the bottleneck of decision processes and processing information required for decision making. This sub-project will begin with a conceptual analysis based on extensive real-world application experience. It will also define and model a set of essential policy elements in order to create a policy catalog. In the implementation phase, the machine-understandable policy representation will feed into the first release of the automated planning component. Building on this, the core watch services will be delivered. These services will in part be based on the analysis of file-type trends in web harvests. In the final phase the policy-aware planning component will be fully integrated with the platform and repository operations.

4.3 ENSURE

ENSURE's architecture consists of:

- a set of plug-ins that provide specific functionality such as format management, regulatory compliance, integrity checks, access to specific storage clouds etc.
- a runtime SOS framework that allows composing an OAIS solution [30] from appropriate plug-ins to meet a user's requirements including economic considerations,
- a configurator and a cost/performance/quality analysis engine witch can evaluate a proposed preservation solution

The *ENSURE Configuration Layer* runs prior to the initial deployment of the solution and re-executes periodically or if there are major environmental changes. Based upon the external requirements and observations on changes to the environment, the configurator can propose several possible solutions. These solutions are composed by choosing a set of plug-ins for the ENSURE framework, which, when taken together, meet the requirements. These candidate solutions are then evaluated and optimized by cost and performance models and evaluated by the preservation planning layer determining the quality of the proposed solution. Based upon this analysis, an administrator can choose the appropriate solution to deploy.

The second major layer containing our innovations is the *ENSURE Preservation Runtime*. The runtime layer is the SOA infrastructure for executing the plug-ins selected by the configuration layer. This layer provides data management and archival storage as well as ingest and access. In addition, this layer interacts with external storage services which provide the physical space for storing the preserved object and which may provide mechanisms for offloading certain preservation-related computations to be "closer" to the objects.

The ENSURE Preservation Runtime layer has four components:

- *Preservation Digital Asset Lifecycle Management* that manages the workflow of the information being preserved, from the time it is handed over to the system until the time it can be deleted since it is no longer needed. This component provides the glue for invoking the other components in the system and provides search capabilities based on ontology evolution.

- *Content-Aware Long-Term Data Protection* is responsible for the long term protection of the digital information, managing changes in what it means to secure information over time. ENSURE will focus on long term access control, long term privacy via the use of appropriate de-identification mechanisms, and intellectual property protection.

- *Preservation Runtime Infrastructure* will support a range of approaches to future accessibility including both transformation and virtualization

- *Preservation-aware Storage Services* provides the interface and mechanisms that enable storing the digital resources managed by the preservation solution in external storage services, such as clouds, and implementing preservation actions, such as integrity checks, near the data.

4.4 TIMBUS

As stated previously, TIMBUS views the digital preservation of business processes as three stages:

1. ***Expediency*** *of digital preservation effort.*

   A crucial aspect of enterprise risk management with regard to digital preservation of business processes is a careful analysis of the service dependencies in a specific business process. The following are some of the common types of dependencies that need to be preserved:

   - A *needs* B — A can only be made available when B has previously been available. For A to be preserved, B must be preserved.
   - A *substitutes* B — A can be used as a replacement for B. A can be preserved instead of B.
   - A *mirrors B* — the behavior and data of A must maintain consistency with the behaviour and data of B. A load-balancing capability and availability property must be preserved.

   A service is preserved if and only if there is some assurance that its complete dependency graph can be reconstructed at any lifetime t, where 0 <t <= PG, and PG is the *Preservation Guarantee* provided.

2. ***Execution*** *of digital preservation process.* When preserving business processes comprised of many interconnected services the legal/regulatory issues become more difficult to maintain and evolve over a long period of time. Legalities Lifecycle Management (LLM) consists of four parts: (1) intellectual property management; (2) IT contracting; (3) data protection; (4) monitoring of legal obligations to preserve. TIMBUS will develop innovative legal/regulatory processes and tools that could be incorporated into commercial ILM products. The tools will be *aware* of legal issues and also changes to legal issues or the introduction of new regulatory standards.

   Digitally preserving a business process that may be comprised of hardware devices and multi–tier service platforms will be easiest if all services are specifically engineered for preservation. However, it is also vital to address the current situation, i.e., services not engineered for preservation. TIMBUS will approach both tasks by focusing on the interfaces and metadata produced by services and the producing/consuming mechanisms.

   Server and desktop virtualization is one of the more significant technologies to impact computing in the last few years. Using virtualisation technology, a business process of distributed inter-dependent services can operate as one "virtual" system. The convergence of affordable, powerful platforms and robust scalable virtualization solutions is spurring many technologists to examine the broad range of uses for virtualisation. For very long life cycles it may also be necessary to provide support for stacked virtualisation (when support for a virtualisation technology ends and the virtualised business process needs to be virtualised again).

   Storage of the digitally preserved business process will also be an issue. Should the business process be stored as one large object? Should it be stored as a set of virtualised inter-dependent services? Should it be stored by an independent storage provider? Can it be stored by a group and spread across different locations? TIMBUS will work with the use case partners to establish a set of business process storage models that are informed by legalities/regulations, security, integrity, and so forth.

3. ***Exhumation*** *of digitally preserved assets.* As previously noted, we cannot go into the future to perform the rerun/integration. However, we can begin

to provide some level of *simulated future*. Our objectives in TIMBUS are: (1) exhuming the business process with the underlying infrastructure hidden – end user perspective; (2) exhuming the business process with the underlying infrastructure exposed – verification perspective; (3) exposure of appropriate metadata regarding business process and supporting software/technology stack; (4) interfacing with other services via standardized information exchange formats specifically addressing digital preservation concerns; (5) a "future simulated" test bed providing guarantee of the preserved business process rerunning and integrating by simulating future changes such as new file formats, interface changes, OS changes, storage changes, database changes, etc.

4.5 Comparison and Contrast

All four projects intend to develop prototype software frameworks. SCAPE, ENSURE, and TIMBUS propose to implement platforms for the execution of preservation processes or workflows. Both SCAPE and ENSURE propose service-oriented architectures (SOA), although SCAPE intends to use SOA workflows as prototypes that should later be executed on a parallel processing architecture. TIMBUS is concentrating on the legal and IPR aspects of the digital lifecycle, while ENSURE is more concerned with economic cost/quality/performance trade-offs and how these are managed as part of information lifecycle management. ARCOMEM's two stage workflow (content selection, content enrichment) is, in contrast, highly specialized for the web archiving use case.

Both SCAPE and ARCOMEM hope to use the Internet itself as a guide for preservation practices. In the case of ARCOMEM the content of social media should guide the harvesting process; in the case of SCAPE, trends that can be observed from Internet harvesting (for example, the frequency distribution of file types) will be used as input for the automated preservation planning process. ENSURE foresees a configuration layer that manages preservation planning, again with specific emphasis on cost, performance and quality trade-offs. TIMBUS proposes a unique approach to planning through dependency and risk management.

Both ENSURE and TIMBUS explicitly plan to use virtualisation as a tool for preservation, although ENSURE appears to focus more on using virtualisation and emulation in order to access digital objects, whereas TIMBUS sees virtualisation as a means to preserve and recover entire business processes.

5. RELATED WORK

Because the projects we describe touch on so many aspects of digital preservation, there is a broad set of related work. Clearly these projects all build on prior major efforts such as CASPAR [9] and Planets [32] and standards such as OAIS [30]. And while there is overlap in the relevant prior art, each project pulls in its own specific related work. Given the breadth of areas touched, this description of related work only scratches the surface.

Service Oriented Architecture (SOA)/Service Oriented Computing is relevant to SCAPE, TIMBUS and ENSURE. Many prior preservation approaches, e.g., [32, 9], built on SOA. While SOAs have many positive aspects, based upon the Planets' experience, SCAPE concluded that there is a need for a more scalable approach to processing the vast amounts of data managed in a large scale digital preservation solution. One specific concern is the difficulty of defining, debugging and executing complex preservation flows in a SOA framework. Another concern is the overhead both on the network and computation in processing the text intensive SOA protocol.

Grid infrastructures address some of these concerns. Data grids, such as Integrated Rule-Oriented Data System (iRODS) [21] can manage huge amounts of scientific data dispersed over heterogeneous sites. As depicted in [18], it is conceptually possible to model simple workflows using the iRODS's rule declaration language. The relative complexity of the iRODS technical language, however, makes it inappropriate for use by workflow designers; on the other hand, it is possible to use a workflow engine like Taverna [19] on top of iRODS storage layer. Even if we use a tool such as Taverna to define the workflows, we still need to consider that data-grid approaches primarily focus on data access, replication, and bit-stream integrity rather than providing data-intensive execution capabilities.

One paradigm for executing operations in parallel is the cloud-derived MapReduce framework [13]. It provides an abstraction for a highly parallel data flow architecture where each processing step operates on some partition of the very large data set. Hadoop [4] is a publicly available MapReduce engine. SCAPE and ARCOMEM will build on efforts like Hadoop to address the research challenges outlined above, processing large numbers of objects in parallel. Initial experiments have already demonstrated the feasibility of this approach [36].

Related to this use of cloud-derived technologies for scale-out computation is ENSURE's use of storage clouds (public or private) for digital preservation. Storage clouds, with their pay-per-use model, are one of the most important new ICT trends. However, the immaturity of these offerings leads to questions on their appropriateness for digital preservation [34, 25]. In spite of these concerns, there have been initial efforts to use storage clouds as the infrastructure for digital preservation. Most notable is DuraCloud [14] which offers a service that can run on multiple cloud providers and which provides the first strong example of building preservation solutions on clouds, addressing issues such as using a cloud for a backup copy,working with multiple cloud providers and running compute jobs on the preserved content in the cloud. ENSURE will build on the concepts and approach of DuraCloud, examining ways to address the concerns that exist in using a storage cloud for preservation. In particular ENSURE will look at how to integrate into a preservation solution, concepts such as as proofs of retrievability/data possession [7, 12] and provenance tracking in the cloud [29]. To enable scalability, like SCAPE, ENSURE will examine how to move preservation computation closer to the data, building on CASPAR's Preservation DataStores (PDS) [33] and emerging paradigms for compute near storage such as the aforementioned MapReduce.

The preservation of service-based processes becomes a challenge of scale. Unlike static software components, for which preservation approaches exist, e.g., emulation and versioning solutions, service-based processes are characterized by being dynamic, frequent reconfigurations, replacement of single components, continuous release cycles, and dependency on informal contextual parameters. This makes it hard to always have a complete snapshot of an entire system/process to preserve. Learning and reasoning techniques have to be employed and handle changes.

With regard to TIMBUS, the EDOS [15] EU project provides techniques and tools for quality assurance and better dependency management of service based processes. The MANCOOSI [26] EU research project is also relevant to TIMBUS. It is encodes the relationships between software components such as dependencies and conflicts, and solves dependencies encoded as a multi-criterion optimization problem with different utility functions, e.g., cost of the software, time to setup, and human resources, etc.

One area of focus for TIMBUS is Intelligent Enterprise Risk Management. Understanding enterprise risk, not just financial risk, has been addressed from a business continuity management perspective. Approaches to model and analyse resource dependencies, failure propagation and recovery models will be used as a baseline and include Failure Mode, Effects, and Criticality Analysis (FMECA) [17, 31], Fault Tree Analysis (FTA) [11], Tropos Goal-Risk Framework [3], Risk Aware Process Evaluation (ROPE) [39].

Related to this analysis of risk in TIMBUS, ENSURE examines cost value trade-offs. ENSURE is not the first project to consider the cost of digital preservation. For instance, [16] and [37], among others, both describe approaches to modelling the cost of a preservation solution. There is also work to evaluate the quality of solutions, with tools such as [24, 20] which build on the emerging ISO standard for Audit and Certification of Trustworthy Digital Repositories. ENSURE goes beyond these efforts by adapting techniques such as benchmarking models [10] and extends these approaches with a view of whole life cycle cost to address obsolescence [35, 38] to allow cost/value tradeoffs.

Protecting data over the long term, which is one of the focus areas of ENSURE, has multiple aspects. One of the more significant is to prevent leakage of personally identifiable information. De-identification is a common approach to facilitate secondary use of personal data by sanitizing the data. Beyond basic techniques which remove or mask direct identifiers, more advanced techniques, e.g., [22], address more sophisticated re-identification attacks. None of these approaches, however, address the fact that what constitutes personally identifiable information changes over time.

Several projects have pursued Web archiving (e.g., [2, 1]). The Heritrix crawler [28], jointly developed by several Scandinavian national libraries and the Internet Archive through the International Internet Preservation Consortium (IIPC), is a mature and efficient tool for large-scale, archival-quality crawling. The IIPC has also developed or sponsored the development of additional open-source tools and an ISO standard for web archives (ISO 28500 WARC). On the operational side, the Internet Archive and its European sibling, the Internet Memory Foundation, have compiled a repository of more than 1 Petabyte of web content which is growing at 100 Terabytes per year. A large number of national libraries and national archives are now also actively archiving the Web as part of their heritage preservation mission.

The method of choice for memory institutions is client-side archiving based on crawling. This method is derived from search engine crawl, and has been evolved by the archiving community to achieve a better completeness of capture and to increase temporal coherence of crawls. These two requirements (completeness and temporal coherence) come from the fact that, for web archiving, crawlers are used to build collections and not only index [27]. These issues were addressed by LiWA (Living Web Archives) [23], which also develops new approaches for the capturing of rich and complex web content, data cleansing and filtering, and archive interpretability.

6. SUMMARY

We presented the four new large digital preservation projects funded by the EC that started in 2011: ACROMEM, SCAPE, ENSURE, and TIMBUS. The motivation for all projects is expanding the scope of long term digital preservation. However the use cases motivating the work vary from publicly available data on the web to data of commercial organizations. The data spans beyond documents to commercial, medical, and scientific data that needs to be interpreted by programs or workflows.

The objectives of the projects spans from methods to define what should be preserved to building the preservation environment. For deciding what to preserve different methods are planned, varying from use of social web, to risk and cost based approaches, and considerations of data protection. For preservation environments, scalability, reconfigurability, supporting different types of data, supporting preservation software, and handling the full lifecycle of preservation are among the areas addressed by the projects. All the four projects plan to develop prototype tools and to build on results of previous projects.

While the projects presented here differ in their objectives and approaches, together they try to cover a bigger part of the long term digital preservation problem by addressing wider range of organizations that need preservation, more types of data, and practical problems of tools and scalability. As the projects progress in the following years, interoperability between those projects and with other digital preservation efforts will be considered. Our hope is that our efforts will make digital preservation more accessible and will contribute to future usability of our digital information.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh FrameworkProgramme (FP7/2007-2013) under grant agreements No 270000, 270239, 270137, and 269940.

8. REFERENCES

[1] S. Abiteboul, G. Cobena, J. Masanes, and G. Sedrati. A First Experience in Archiving the French Web. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '02, pages 1–15, London, UK, UK, 2002. Springer-Verlag.

[2] A. Arvidson and F. Lettenström. The Kulturarw Project - The Swedish Royal Web Archive. *Electronic library*, 16(2), 1998.

[3] Y. Asnar and P. Giorgini. Modelling Risk and Identifying Countermeasure in Organizations. In J. Lopez, editor, *1st International Workshop on Critical Information Infrastructures Security*, volume 4347 of *Lecture Notes in Computer Science*, pages 55–66. Springer-Verlag, 2006.

[4] A. Bialecki, M. Cafarella, D. Cutting, and O. O'Malley. Hadoop: a framework for running applications on large clusters built of commodity hardware. *Wiki at http://lucene. apache. org/hadoop*, 2005.

[5] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Sustainable Economics for a Digital Planet, ensuring Long-Term Access to Digital Information, 2010. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

[6] D. Borthakur. *The Hadoop Distributed File System: Architecture and Design*. The Apache Software Foundation, 2007.

[7] K. Bowers, A. Juels, and A. Oprea. HAIL: A high-availability and integrity layer for cloud storage. *ACM CCS*, November 2009.

[8] BSI. BS 25999-1:2006 Business continuity management. Code of practice., 2006.

[9] CASPAR Digital Preservation User Community. `http://www.casparpreserves.eu/`, 2010.

[10] C. Chituc and S. Nof. The Join/Leave/Remain (JLR) decision in collaborative networked organizations. *Computers & Industrial Engineering*, 53(1):173–195, 2007.

[11] I. E. Commission. IEC 61025. Fault Tree Analysis, Ed. 2.0, 2006.

[12] R. Curtmola, O. Khan, and R. Burns. Robust remote data checking. In *StorageSS '08: Proceedings of the 4th ACM international workshop on Storage security and survivability*, pages 63–68, New York, NY, USA, 2008. ACM.

[13] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. *Commun. ACM*, 51:107–113, January 2008.

[14] DuraCloud. `http://www.duraspace.org/duracloud.php`.

[15] EDOS - Environment for the development and Distribution of Open Source software. `http://www.edos-project.org`.

[16] K. Fontaine, G. Hunolt, A. Booth, and M. Banks. Observations on cost modeling and performance measurement of long term archives. In *PV Conference*, 2007.

[17] Y. Haimes. *Risk Modeling, Assessment, and Management*. John Wiley & Sons, Inc, 2009.

[18] M. Hedges, T. Blanke, and A. Hasan. Rule-based curation and preservation of data: A data grid approach using iRODS. *Future Gener. Comput. Syst.*, 25:446–452, April 2009.

[19] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(Web Server issue):729–732, July 2006.

[20] P. Innocenti, S. Ross, E. Maceviciute, T. Wilson, J. Ludwig, and W. Pempe. Assessing digital preservation frameworks: the approach of the SHAMAN project. In *MEDES '09: Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, pages 412–416, New York, NY, USA, 2009.

[21] IRODS: Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems. `http://irods.sdsc.edu/index.php/Main\_Page`.

[22] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *ICDE*, pages 106–115. IEEE, 2007.

[23] LiWA – Living Web Archives. `http://www.liwa-project.eu/`, 2011.

[24] Long Term Digital Preservation Assessment Tool, IBM Haifa Research Lab,. `https://www.research.ibm.com/haifa/projects/storage/datastores/ltdp.html`.

[25] Long-term Preservation Storage: OCLC Digital Archive versus Amazon S3. `http://dltj.org/article/oclc-digital-archive-vs-amazon-s3/`.

[26] MANCOOSI - managing software complexity. `http://www.mancoosi.org`.

[27] J. Masanès. *Web archiving*. Springer, 2006.

[28] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. Introduction to heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, 2004.

[29] K.-K. Muniswamy-Reddy, P. Macko, and M. Seltzer. Provenance for the Cloud. In *8th USENIX Conference on File and Storage Technologies (FAST '10)*, San Jose, CA, USA, Feb. 2010. USENIX.

[30] *OAIS: Space data and information transfer systems – Open archival information system – Reference model*. ISO 14721:2003, 2003.

[31] I. A. Papazoglou, O. N. Aneziris, J. G. Post, and B. J. M. Ale. Technical modeling in integrated risk assessment of chemical installations. *Journal of Loss Prevention in the Process Industries*, 15(6):545 – 554, 2002.

[32] PLANETS home. `http://www.planets-project.eu/`, 2010.

[33] S. Rabinovici-Cohen, M. Factor, D. Naor, L. Ramati, P. Reshef, S. Ronen, J. Satran, and D. L. Giaretta. Preservation DataStores: new storage paradigm for preservation environments. *IBM Journal of Research and Development*, 52(4):389–399, 2008.

[34] D. Rosenthal. Preservation in the Cloud. In *Preservation in the Cloud*. Library of Congress, September 2009.

[35] P. Sandborn and G. Plunkett. The other half of the DMSMS problem-software obsolescence. *DMSMS Knowledge Sharing Portal Newsletter*, 4(4):3, 2006.

[36] R. Schmidt, C. Sadilek, and R. King. Workflow System for Data Processing on Virtual Resources. *International Journal on Advances in Software*, 2(2–3):234–244, 2009.

[37] J. Slats and R. Verdegem. SCost Model for Digital Preservation. `http://dlmforum.typepad.com/Paper\_RemcoVerdegem\_and\_JS\_CostModelfordigitalpreservation.pdf`, 2010.

[38] S. Strodl, C. Becker, R. Neumayer, and A. Rauber. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, page 38. ACM, 2007.

[39] S. Tjoa, S. Jakoubi, and G. Quirchmayr. Enhancing Business Impact Analysis and Risk Assessment Applying a Risk-Aware Business Process Modeling and Simulation Methodology. In *ARES*, pages 179–186, 2008.