# ENSURE: Long term digital preservation of Health Care, Clinical Trial and Financial data

Maïté Braud
TESSELLA
Abingdon, UK
maite.braud@tessella.com

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

Jochen Rauch
Fraunhofer (IBMT)
St. Ingbert, Germany
jochen.rauch@ibmt.fraunhofer.de

Simona Rabinovici-Cohen
Kenneth Nagin
John Marberg
IBM Research - Haifa
simona@il.ibm.com

David Voets
Custodix
Sint-Martens-Latem, Belgium
david.voets@custodix.com

Isaac Sanya
Mohamed Badawy
Essam Shehab
Cranfield University, UK
i.o.sanya@cranfield.ac.uk

Frode Randers
Luleå University of Technology
Luleå, Sweden
frode.randers@ltu.se

J.A. Droppert
Philips Digital Pathology Solutions
Best, The Netherlands
aad.droppert@philips.com

Marcin Klecha
Philips Digital Pathology Solutions
Best, The Netherlands
marcin.klecha@philips.com

## ABSTRACT

This paper presents the initial results of the ENSURE (**E**nabling k**N**owledge **S**ustainability, **U**sability and **R**ecovery for **E**conomic value) project, which focuses on the challenges associated with the long-term preservation of data produced by organisations in the health care, clinical trials and financial sectors. In particular the project has looked at the economic implications of long-term preservation for business, how to maintain the accessibility and confidentiality of sensitive information in a changing environment, and how to detect and respond to such environmental changes. The project has developed a prototype system, which is based around a lifecycle manager and makes use of ontologies to identify and trigger necessary transformations of the data objects in order to ensure their long-term usability. It also uses cloud technology for its flexibility, expansibility, and low start-up costs. This paper presents one of the use cases: the health care as a way to illustrate some of the challenges addressed by the ENSURE system.

## 1. INTRODUCTION

Ensuring long-term usability for the spiralling amounts of data produced or controlled by organisations with commercial interests is quickly becoming a major problem. Drawing on motivation from use cases in health care, finance, and clinical trials, ENSURE [1] extends significantly the state of the art in digital preservation, which to date has focused on relatively homogeneous cultural heritage data. ENSURE's use cases bring up a large number of issues, which have yet to be addressed fully, such as:

- How to leverage a scalable, pay-as-you-go infrastructure for digital preservation.
- How to get businesses to understand the economic implications of long-term preservation.
- How to create an archiving workflow that conforms to the regulatory, contractual and legal requirements of the health care, finance or clinical trials domains.
- How to maintain over the long term the integrity and authenticity of highly personal data and material covered by intellectual property rights, while ensuring access controls are respected.
- How to create a digital preservation system using only off-the-shelf IT technology.

Building on prior work, ENSURE addresses these issues with innovative approaches and tools to create a flexible, self-configuring software stack. Based on the business requirements the user enters, 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 over time of the preserved data. The main innovation areas of ENSURE are:

- *Assessment of Cost, Value, and Quality.* Ensure is creating cost, value, and quality models to help build the best preservation solution, in terms of price and performance that adheres to businesses' requirements.
- *Automation of Preservation Lifecycle Management.* Ensure uses workflow management tools to manage the execution of preservation workflows over time, thus ensuring regulatory compliance, allowing changes in the environment to be reflected in changes to the preservation approach, addressing the evolution of ontologies and managing the quality of the digital objects over time.
- *Expansion of Standard ITC Use.* Ensure is investigating using emerging technologies, such as Cloud Computing and virtualisation, to create scalable and financially-viable solutions for long- term digital preservation.

- *Creation of Content-Aware Long-Term Data Protection.* Ensure is researching how to secure data over the long-term, when that data is affected by new and evolving regulations, contains personally-identifiable information, and needs to be accessed by a changing user community with differing roles.

The ENSURE project started in February 2011 and has created a reference architecture already and demonstrated many innovations in its initial implementation.

Section 2 presents the overall architecture of the ENSURE system, section 3 and 4 describes the two main components of the ENSURE system: the Configuration Layer and the Runtime System, Section 5 present a use case, and section 6 gives our conclusions..

## 2. ENSURE ARCHITECTURE

The ENSURE system's architecture consists of:

- A set of plug-ins that provide specific functionality such as format management, regulatory compliance, integrity checks, and access to specific storage clouds.
- A runtime Service-Oriented Architecture (SOA) framework that allows an OAIS [2] solution to be created from those plug-ins needed to meet a user's requirements, including any economic considerations (s)he has.
- A configurator and an optimiser which use cost/quality analysis engines to create and evaluate a proposed preservation solution.

A high-level view of the ENSURE architecture is given in Figure 1, which shows that there are two layers: the *Configuration Layer* and the *System Runtime*.

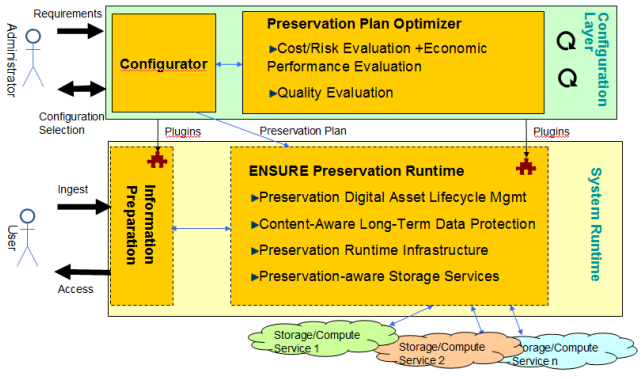


**Figure 1. ENSURE System Overall Architecture**

These two layers are described in the next two sections.

## 3. CONFIGURATION LAYER

The components in the *ENSURE Configuration Layer* are run prior to the initial deployment of the preservation solution and are re-run periodically; in particular they need to be re-run if there are major environmental changes. These components create the preservation plan used by the preservation solution in the first place and update it when the environmental or business needs change.

The architecture of the Configuration Layer is given in Figure 2.

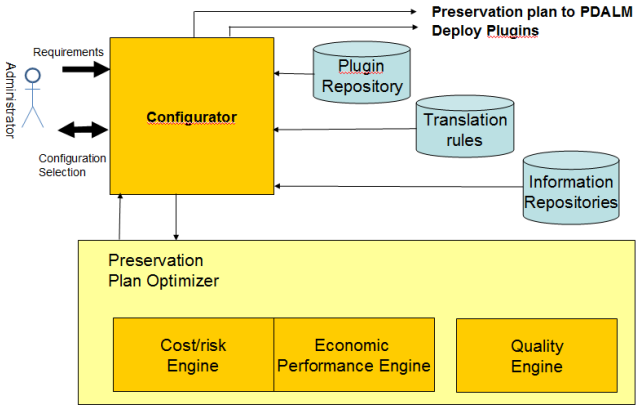


**Figure 2. Configuration Layer Architecture**

The flow of operation in the Configuration layer is as follows:

- The administrator is presented with a form and enters the business' requirements and preferences for the preservation system. When the system is being reconfigured, the previous configuration is shown to the administrator for reference.
- Using a Rule Engine and a set of rules, the Configurator constructs a parameterised global preservation plan (GPP) consisting of a preservation plan and its associated configuration. A global preservation plan (GPP) defines where and how data will be preserved; this includes encryption of data, fixity checks and storage provider. A parameterized GPP describes a collection of potential plans by means of parameters that take values from well-defined ranges. For example, one parameter could define a collection of possible encryption algorithms, and another parameter could define a collection of storage providers.
- The Preservation Plan Optimiser (PPO) explores the collection of potential plans, returning to the Configurator a small number of plans that are optimised with respect to cost and quality. The PPO uses the Quality Engine and the Cost Engine to provide evaluations of potential plans. These evaluations drive the optimization.
- The Configurator presents the top three preservation plans to the administrator together with their evaluations. Either the administrator can select the solution to deploy, or (s)he can request a modified configuration, which will restart the process.
- When a preservation solution is chosen, the Configurator deploys it by:

  1. Deploying the selected plug-ins in the runtime infrastructure and activating the associated services in the appropriate environment.
  2. Activating the Preservation Digital Assets Lifecycle Management component and passing it the preservation plan.
  3. Storing the selected configuration and its evaluation in the ENSURE system in order to preserve it.

## 3.1 Preservation Plan Optimiser

Finding preservation plans that are optimised with respect to cost and quality is a multi-objective optimisation problem. Typically the objectives are conflicting and there is not a single best solution. Evolutionary algorithms are widely used to find solutions that are Pareto optimal [3]. The PPO uses the evolutionary algorithm NSGA-II [4] to explore the collection of potential plans and find optimal solutions. It defines a genotype that encodes the parameters of the parameterised GPP. For example, there can be a gene representing a choice for an encryption algorithm. The evolutionary algorithm selects actual values for the genes of the genotype, thus generating candidate plans that PPO then sends to the engines for evaluation of quality and cost. The quality and cost values thus obtained act as objective values that are maximised or minimised in the optimisation performed by the evolutionary algorithm.

Several software frameworks exist that provide implementations of evolutionary algorithms. The Opt4J optimisation framework [6] has been selected for the PPO.

In order to take account of user preferences, the ENSURE project uses *a priori* preference articulation i.e. the user expresses preferences before the optimisation is performed. The PPO defines a weight function on the objective space to represent the user's stated rating of the importance of the different objectives. The Opt4J implementation of NSGA-II has been extended to use such weightings, as described in [5]. The selection performed by the evolutionary algorithm thus favours solutions that score well on the objectives that the user considers important.

## 3.2 Cost Modelling for Long-Term Digital Preservation

Assessing the cost and economic value of preserving digital information is important for organisations performing preservation activities. Therefore, one of the aims of ENSURE is to develop a cost model and a cost engine to predict the 'whole life-cycle cost' of LTDP in the cloud. The developed cost model will focus mainly on three business sectors: healthcare, financial and clinical trials.

The core activities involved in the design and development of the cost engine for ENSURE include:

1. Identification of the work break down structure (WBS) and cost break down structure (CBS) of digital preservation activities, as identified in Figure 3.
2. Identification of cost drivers, risks/uncertainties factors, and obsolescence issues in LTDP activities
3. Development of the cost model, including implementable cost equations and rules.
4. Implementation of the cost engine as a web service and its integration into the rest of the ENSURE architecture.

The activity based costing (ABC) methodology has been employed to develop the cost model. The ABC approach enabled the development of a generic cost model that is applicable to and relevant for not only the ENSURE use-case organisations but also other industries. The cost model is translated into a set of equations and rules to enable the accurate estimation of cost for LTDP activities.

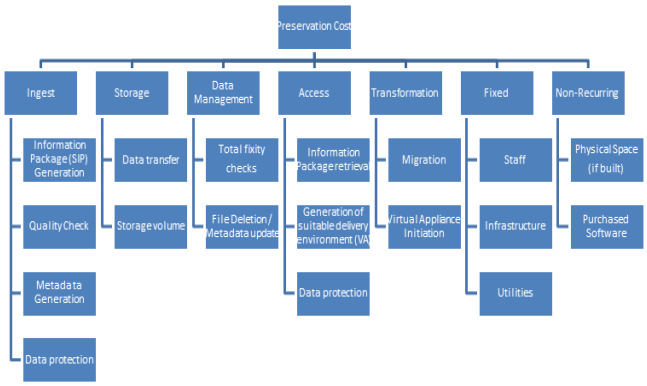


**Figure 3. ENSURE Cost Breakdown Structure of Digital Preservation Activities**

### *3.2.1 Challenges in Cost Estimation for LTDP*

For ENSURE, the main challenges of estimating the cost of LTDP are as follows:

- Long-term digital preservation is being applied to three new business sectors. Most previous work in digital preservation has focused on the science and cultural heritage sectors.
- There is no established definition of uncertainty for LTDP.
- No research on the impact of uncertainty on cost has been undertaken and information about this topic is scarce.
- Limited work has been done to investigate the cost of ameliorating obsolescence through LTDP.
- LDTP made use of cloud computing only recently, so cost data is scare. Cloud costs are split between cloud storage and cloud computing.
- Determining the cheapest configuration is made harder by the number of parameters that can be optimised.

### *3.2.2 Cost Engine Architecture*

The cost engine system architecture comprises several communicating components that implement the overall ENSURE cost evaluation and optimisation system. Figure 4 illustrates the architecture of the cost engine and how its modules interact with the rest of the ENSURE system. The GPP describes aggregation-specific (see section 4.5.2) (e.g. encryption, fixity, etc.) and copy-specific (e.g. storage, computing) preservation actions and the preservation configuration. The preservation configuration describes the physical architecture, software, and plug-ins employed for digital preservation activities. The cost engine results include initial investment cost, year one cost, ingest cost, data management cost, storage cost, access cost and reconfiguration cost for the data retention period (given in years) in the configurator.

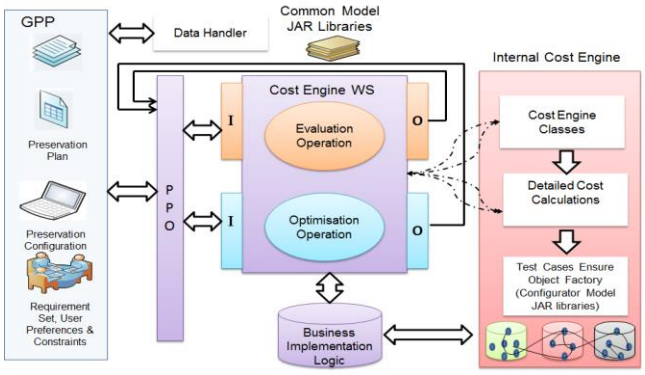


**Figure 4. ENSURE Cost Engine Architecture**

### *3.2.3 Validation*

The cost engine has been validated qualitatively via expert opinion in the digital preservation community. The first phase of validation was the cost break down structure, followed by the equations and rules that have been implemented. There are plans to validate the cost engine quantitatively with real cost values to ensure its generalisability, applicability and validity.

## 4. RUNTIME SYSTEM

The ENSURE System Runtime is the SOA infrastructure for executing the plug-ins selected by the Configuration layer. This layer provides data management and archival storage services, as well as ingest and access services. It interacts with external storage services which provide the physical space for storing the preserved data and potentially it interacts with the external compute service, which runs in the storage layer to minimise i/o overheads. In addition, this layer watches for environmental changes that may require the system to be reconfigured.

The architecture of the runtime system is given in Figure 5.

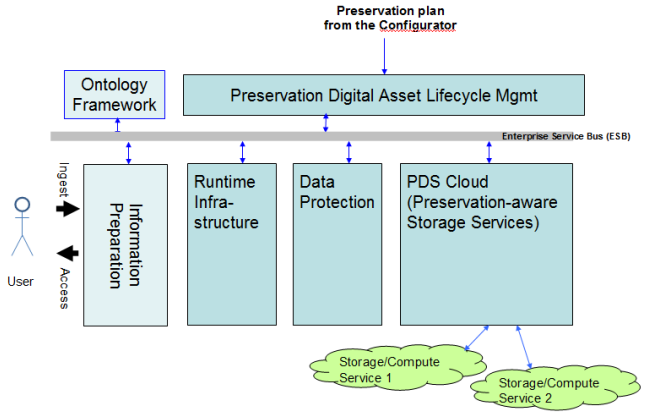


**Figure 5. Runtime System Architecture**

The components of the Runtime System are:

- Preservation Digital Asset Lifecycle Management: It manages the workflow of the information being preserved by executing the preservation plan built by the Configurator. In addition, it manages the system's log and all provenance information. Furthermore, it handles changing the workflow at reconfiguration, and it monitors internal and external events including watching the environment for events that need administrator attention or reconfiguration. Also, it handles sending out notifications and all interactions with the administrator.
- Information Preparation: It runs when the data is being ingested or accessed. Upon ingest, it prepares the information and metadata ready to be preserved, generates the search indexes, and packages the data. Upon access, it handles locating the data in the index and packaging it for the user. Its data protection functions are used to ensure access rights are observed.
- Ontology Framework: It manages the preservation Ontologies and search Index. It also supports the evolution of the Ontologies.
- Preservation Runtime Infrastructure: It evaluates the quality of the managed information, supports data transformations, and supports a range of approaches for future accessibility.
- Preservation-aware Storage Service: It stores the digital resources in external storage services using cloud storage, validates the bit-level integrity of the data, manages provenance at the storage level, and supports running computations in the cloud storage layer.
- Content-aware Long-Term Data Protection: It ensures that the use of sensitive information over the preservation life-cycle complies with the specified long-term access controls, privacy restrictions, IPR protection rules, and de-identification, and anonymisation requirements.

## 4.1 Preservation Digital Asset Lifecycle Management (PDALM)

ENSURE has researched the integration of existing approaches to Lifecycle Management with digital preservation and this research is encapsulated in the PDALM component. It orchestrates the management of an asset from ingest to disposal, by invoking components developed in other work packages. Objects are disposed of only according to the applicable rules and regulations of the relevant business sector, together with the relevant business objectives.

In essence the PDALM component is the "brain" of the ENSURE system and therefore is also responsible for controlling the system activities, and handling notifications to and interactions with the administrator.

### *4.1.1 Workflow Engine*

The PDALM component is principally a workflow engine, which is capable of running those workflows whose details are specified in the Preservation Plan created by the Configurator. It is capable of starting workflows manually or automatically (based on timers or pre-defined rules). The workflow types are consistent with the OAIS model: Ingest, Access, Preservation, and Data Management workflows are available.

The workflow steps themselves are not executed within the workflow engine, but it is responsible for sending web service requests to the other runtime components (Information Preparation, Data Protection, PDS Cloud) to execute the workflow steps.

The workflow engine is based on the open-source workflow engine jBPM (released by the JBoss Community under the ASL license). jBPM comes with a web-based console that allows the user to start workflows and control running workflows. This jBPM console forms the basis for the PDALM Graphical User Interface (GUI).

The PDALM workflow engine contains a component responsible for translating the Preservation Plan created by the Configurator into a series of Ingest, Access, Preservation and Data Management workflow definitions, which will be uploaded automatically into the workflow engine.

A key point to reacting to changes is the ability to reconfigure the system. In collaboration with the Configurator component, the ENSURE system is capable of reconfiguring a running instance of the workflow engine. There are two types of workflows available in the workflow engine: manual workflows which are started by the user, and scheduled workflows which are started automatically by the system. In addition, when the system is up and running there may be workflows waiting for a manual or timed trigger (e.g. transformation workflow) as well as running workflows. In order to reconfigure the live system, the PDALM component needs to be able to put the workflow engine in a dormant state, which is done in stages. When the Configurator notifies the PDALM component that a new preservation plan needs to be deployed, the PDALM component stops all the workflows that are either waiting or scheduled to run, before waiting for all the active, running workflows to complete; only then is the system in a dormant state. Once in this state, the workflow engine can be stopped safely and reconfigured based on the new preservation plan. If still required, the workflows that were scheduled to run or were in a waiting state can be restarted in the newly reconfigured system.

### *4.1.2 Event Engine*

The PDALM event engine is responsible for monitoring internal and external events relating to environmental changes that could affect the long-term preservation of the data preserved by the ENSURE system. Monitoring external events is a difficult task as the source of such events can be as diverse as the ways of monitoring them.

Initially effort was focussed on one of the pre-existing central repositories of information for long-term preservation: PRONOM. PRONOM is developed and maintained by the UK National Archives (TNA) and holds impartial and definitive information about the file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value.

One of the current limitations of PRONOM is that as its information is stored in a relational database, it is difficult to update or merge two instances of PRONOM. Also it makes it difficult to identify which information has changed when an instance of PRONOM is updated. To solve these problems, in 2011 TNA started to implement Linked Data Pronom with the plan to release the data held by PRONOM in a linked open data format in order to make it easier to reuse. Such a Linked Data registry makes it easier to compare two instances of the same registry and detect if and what changes have occurred.

The ENSURE project has started to extend this by adding additional Linked Data instances to complement the information held by Linked Data PRONOM including information in PRONOM but not yet by Linked Data Pronom and also information relevant to the ENSURE use cases, such as cost, hardware, and data protection. The resulting Linked Data network will consist of, for example, external data held in the Linked Data PRONOM instance maintained by TNA, data held in a Linked Data instance about other relevant technical information (e.g. local tool capabilities), and data held in a Linked Data instance about costs maintained by ENSURE. This Linked Data network will help to demonstrate how it is possible to get notification of external events and react to them by using Linked Data.

Apache Jena was chosen as the Framework to handle the Linked Data network as it provides the following functionality out of the box:

- an API for reading, processing and writing RDF data in XML, N-triples and Turtle formats,
- an ontology API for handling OWL and RDFS ontologies,
- a rule-based inference engine for reasoning with RDF and OWL data sources,
- stores to allow large numbers of RDF triples to be stored efficiently on disk,
- a query engine compliant with the latest SPARQL specification, and
- servers to allow RDF data to be published to other applications using a variety of protocols, including SPARQL.

The event engine contains functionality to query the Linked Data Pronom instance maintained by TNA; a scheduled BPMN workflow, running within the PDALM workflow engine, queries the distant Linked Data Pronom instance and compares it to a snapshot of stored locally in order to detect and identify any changes that might have occurred in since the last query. Then the impact of the change is calculated using the Linked Data network presented above and communicated to the administrator of the ENSURE system via email. Given this information, the administrator may choose to request that the Configurator calculates a new Preservation Plan.

This is illustrated in the following example: The date that the creator of a file format will withdraw support is updated in TNA's Linked Data PRONOM instance (or a copy of this instance) and this change is detected when the scheduled comparison workflow runs. This change is detected and triggers looking up a preservation action to perform as a consequence (e.g. format migration). Then, the event engine will calculate the financial impact of the change using the data stored in a further part of the Linked Data Network. In this case, therefore, the cost will be the cost of running the tool and the additional cost of storing the migrated data based off the chosen migration strategy triggered from the detection of the external event of obsolescence of the file format.

Much work in many different initiatives is being undertaken to unify technical registries and other repositories of digital preservation information: e.g. UDFR[24] and LDS$^3$[23] are focusing on using semantic web and Linked Data to enable the sharing of information. Therefore the Linked Data registry developed as part of the ENSURE project will not be limited to

linking to Linked Data Pronom only but will be capable of linking to other Linked Data registries as well, provided that their vocabulary specification is published and freely available.

## 4.2 Preservation Information Preparation

Information preparation plays an important role in any digital preservation system as it has to ensure that during ingest all the necessary information required for preservation, long-term accessibility and usability of the data objects to be preserved is gathered. The OAIS reference model reflects this both in the Ingest and Access components and in the different information packages of the OAIS information model that are produced or processed by the Ingest and Access component, namely the Submission Information Package (SIP), the Archival Information Package (AIP) and the Dissemination Information Package (DIP). The ENSURE project has demonstrated that it can ingest and retrieve simple and more complex data objects (e.g. DICOM image file sets) from its three target domains. Figure 6 illustrates the data workflow in the Information Preparation architecture. Ingest services were developed that select the right information to be preserved from the test data for each use case, extract the metadata relevant to the data objects' MIME types , package everything in an AIP and hand it over to the ENSURE Preservation Runtime for preservation. Furthermore, access services were developed to provide efficient search and retrieval of data objects in the form of DIPs. In particular, semantic web technologies were applied to model, collect and manage the metadata of the digital objects from the different domains effectively and to provide a powerful search and access mechanism for preserved data. By representing the Data Objects' metadata in terms of an integrated set of formal ontologies, the preservation knowledge and domain-specific object formats and concepts can be modelled in an application-oriented way. The ontologies contain concepts describing the general features of Data Objects (i.e., type, format, size, Preservation Description Information) as well as domain-specific information. The captured metadata of the Data Objects represent instances of the ontologies and are encoded as RDF triples and stored by the ENSURE Preservation Runtime in an index.

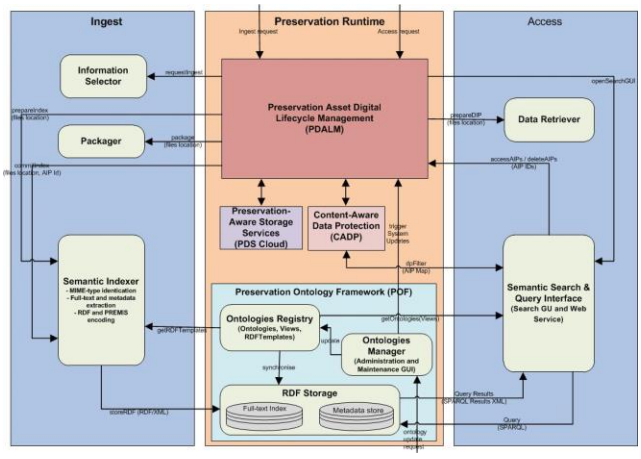


**Figure 6. Overview of the Information Preparation Architecture**

## 4.3 Ontologies Framework

The integrated Preservation Ontology Framework (POF) includes an Ontologies Registry and a set of ontologies related to preservation, which is provided as a subset of the Nepomuk Information Element Ontology (NIE) [7]. This provides the flexibility required to serve the unknown, future data retrieval needs of the user community. It provides the platform to research how the evolution of the ontologies over time can be managed in an archive, and can be exploited to identify and trigger necessary transformations of the data objects in order to ensure their long-term usability. Further, it enables investigations into how the knowledge coded in ontologies can be used to resolve other preservation-related problems, such as the protection of sensitive healthcare data under changing regulations. To do so, a management component, the Ontologies Manager, was implemented to enable the user to maintain different versions of ontologies through a GUI (see Figure 7). In addition to managing the update of ontologies, the Ontologies Manager executes any system adaptations necessitated by the creation of a new version of an ontology, such as re-indexing archived AIPs in order to keep the entire system consistent. The COnto-Diff algorithm [8] is used to calculate the differences between sequential versions of ontologies and provides both the information required to execute the necessary system adaptations and an estimate of the required effort.

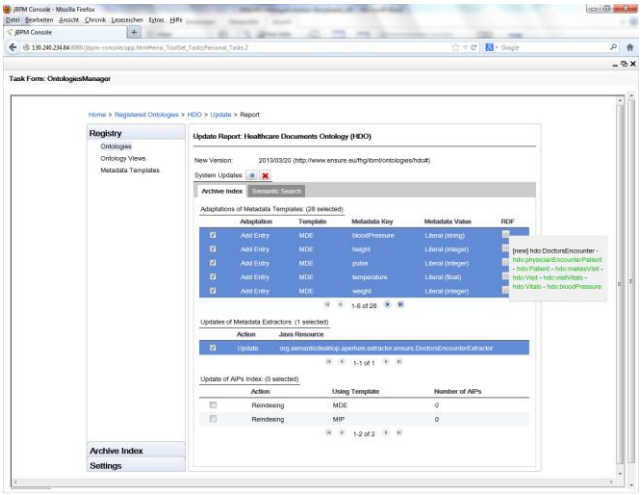


**Figure 7. Screenshot of the Ontologies Manager**

## 4.4 Preservation Runtime Infrastructure

The Preservation Runtime Infrastructure which is part of the Preservation Runtime supports a range of approaches to future accessibility including both transformation and emulation/virtualisation. This component is responsible for providing the transformations of formats, for evaluating the usability of the information after a transformation, and for periodically evaluating the quality of the managed information. This section describes how the quality of the information can be assessed.

In part a system for long-term digital preservation of information can be viewed as a communication system, sending information from a producer to a consumer through a channel (the preservation system). Unlike the channels encountered in standard communication systems, this channel has an extreme intrinsic time delay – possibly measured in decades – that makes any type of feedback-loop impracticable, if not impossible. Under very specific circumstances it is possible to use information theory to analyse the effect of a specific use of an information

system [12] (such as a communication system), but since it is not possible to represent mathematically the types of uncertainties that encountered in digital preservation, it is not possible to use information theory to study the effect of digital preservation systems in generalised use [13]. As the information transfer in digital preservation is determined not only by input and output symbol alphabets and their conditional probabilities, but also depends to a great extent on pre-knowledge and qualitative factors, the authors are forced to conclude that it is not possible to model the digital preservation ”channel” using traditional information theory [14].

What makes the digital preservation domain so elusive and hard to capture in strictly technical terms is the extent to which qualitative factors, such as trust and authenticity, influence the perceived quality of the transferred information. It could be argued that the rendering an image in an obsolete format using emulated viewing software does not differ from migrating that image and then viewing it using contemporary software, but they do differ in terms of the amount of trust you need to have: trust in the chain of migration software used to keep the image up to date and trust in the organisation managing the process [9][10] versus trust in the emulator and the process used to select it.

The ENSURE system aims to empower the preservation services customer (the producer in OAIS terms [2]) to choose an appropriate preservation plan for two reasons: (1) the choice affects the cost and thus should be taken by the customer, and (2) the customer is best equipped to assess the qualitative impact of the proposed preservation plans.

Digital preservation is not only a set of technical problems related to technology, formats and algorithms, but also a problem that concerns the interface between technological systems and humans. Most of all it is a problem concerning the mutual understanding between humans separated in time – and thus by culture. In the ENSURE setting, the preservation organisation aims to help the producer to understand the effects of the chosen preservation plan on the predicted needs of the future consumer and how to best fulfil these needs within the available budget.

There is a fundamental conflict in demanding a decision from the producer regarding a proposed preservation plan because at least two conflicting concerns govern the actions of the producer; minimising the cost and maximising the quality of the transferred information. It is inevitable that the producer and the consumer (and in the ENSURE case the consumer is the producer at a later point in time) will have different views of the information and the use of that information [11] (see Figure 8).

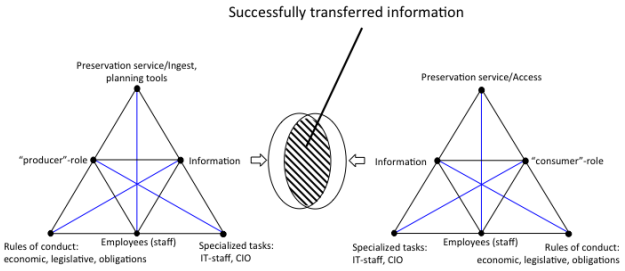


**Figure 8. The Producer and Consumer Perceive the Information Differently, c.f. [4]**

In order to help the producer make an informed decision, ENSURE equips the producer with a tool that supplements the cost of choosing a specific preservation plan with the consequences of choosing that plan. The consequences are provided as (1) the monetised cost of risks based on calculations of economic performance, (2) a metric assessment of perceived quality from a predicted consumer viewpoint, and (3) a set of qualitative statements of the failure to exhibit specific characteristics of quality [17]. These consequences are predicted by attempting to extrapolate the current usage of the information into the future based on the assumed purpose of use of the information together with the purpose of preserving the information.

Empirical data gathered through a series of interviews with the use case owners in ENSURE emphasises the differing concerns of businesses needing preservation and the organisations providing preservation services. Businesses often lack knowledge of digital preservation, while preservation service providers often struggle to understand the specific needs of those businesses requiring their services. As these two organisations are effectively working together to predict the future needs of the business organisation, it is essential that they communicate effectively. This communication has to be based on a shared mental model that is expressive enough to capture the immediate needs of both organisations [16][15].

## 4.5 Preservation-Aware Storage Service

Preservation Data Stores in the Cloud (PDS Cloud) is an OAIS-based [2], preservation-aware, storage service in a multi-cloud environment. Unlike existing cloud storage systems, or even some traditional archival systems, PDS Cloud supports logical preservation; in addition, it converts logical preservation information objects into physical cloud storage objects. The idea behind PDS Cloud is that digital preservation systems will be more robust and will reduce the probability of data corruption or loss if preservation-related functionality is offloaded to the storage system.

The foundations of PDS Cloud were established in PDS [18], a preservation storage architecture using Object Storage Devices (OSD). For the ENSURE system the scope has been expanded and adapted for the cloud environment. The following cloud-specific goals and requirements have been added:

- Support access to multiple cloud storage and cloud computing platforms, and enable migration of data between different clouds. This includes using multiple clouds concurrently, while taking advantage of the special capabilities of each platform.
- Provide a flexible data model for a multi-tenant, multi-cloud environment, with easily configurable data management capabilities that can be tailored for diverse aggregations of digital assets having different preservation requirements that can change over time. A key feature is the ability to change the physical placement of objects in the cloud without affecting how the user accesses the data.
- Enhance the future understandability of content by supporting data access using cloud-based virtual appliances. Each virtual machine instance is created from a previously published image or from readily available components and provided with the desired preservation data content and the designated software needed to render the data.

- Offer advanced OAIS-based services, such as fixity (aka integrity) checks, provenance records and auditing services that complement the generic cloud's capabilities. Also, it must support complex, interrelated objects and manage their relationships and links while maintaining referential integrity.

While this section provides an overview of the architecture and data model of PDS Cloud, a more comprehensive presentation of the PDS Cloud system can be found in [19].

### *4.5.1 Architecture*

PDS Cloud is designed as an intermediate service layer, providing a broker that connects the OAIS entities with the multiple cloud systems; in addition it fulfils the role of the Archival Storage component in the OAIS functional model. PDS Cloud exposes a set of OAIS-based services, such as ingest, access, deletion and preservation actions [2], to the client and uses heterogeneous storage and computing cloud platforms from different vendors. AIPs may be replicated to multiple cloud storage systems to exploit different cloud storage capabilities and pricing structures, and to increase data survival.

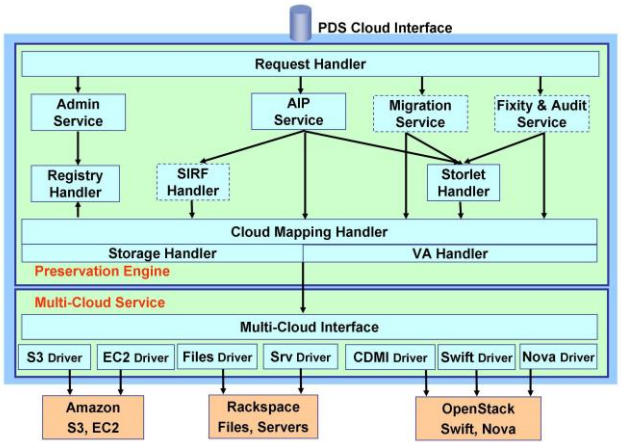


**Figure 9. PDS Cloud High-level Architecture**

As shown in Figure 9, PDS Cloud is divided into two main layers: a Multi-Cloud Service, and a Preservation Engine.

- *Multi-Cloud Service:* This handles access to a heterogeneous set of cloud storage and computation platforms. Its role is to encapsulate the specific interfaces and capabilities exposed by each different cloud platform. It is agnostic to preservation and is implemented using jclouds [20], an open source cloud interface library that comprises a unified interface (multi-cloud interface component) and a set of drivers that implement the interactions with the individual storage and computation clouds underneath.
- *Preservation Engine:* This provides the preservation functionality for AIPs. It receives requests from PDS Cloud clients and services them using various functional handlers organised in several levels. At the top level is the Request Handler, which is the server side of the HTTP interface. When it receives an HTTP request, it validates it, before handing it over to the appropriate handler for processing. At the lowest level is the Cloud Mapping Handler, which maps AIPs to the cloud object model, and interacts with the Multi-Cloud Service layer to perform operations in the cloud.

This architecture, with its separation of concerns, is designed to support the deployment of multiple clouds from different vendors. Providing such heterogeneity allows the user to experiment with diverse technologies and to determine whether appropriate actions have been taken to ensure continued access to the AIPs despite the diversity of current technologies; this is analogous to ensuring continued access to AIPs despite the change in technologies over time, i.e. ensuring their preservation.

### *4.5.2 Data Model*

Users should be able to access their data without needing to know the details of how or where it is stored. PDS cloud hides the complexity of a dynamically configured, multi-cloud, multi-tenant environment behind a simple facade that uses a uniform, hierarchical resource naming path for entities and an abstract data model that allows for multiple implementations to take advantage of the different capabilities of the cloud storage platforms being used.

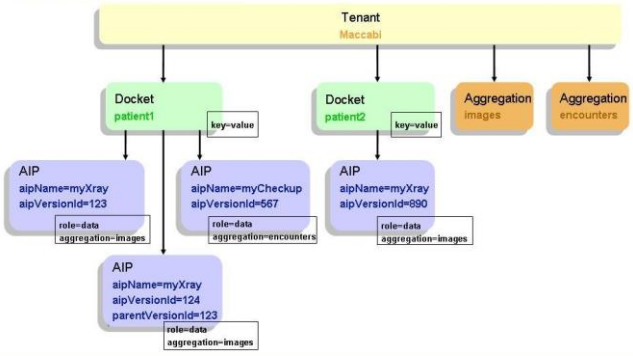


**Figure 10. PDS Cloud Data Model**

The data model, which is illustrated in Figure 10, comprises four types of entities: tenant, aggregation, docket and AIP.

- *Tenant:* This entity is an enterprise or organisation that engages in storing data in the cloud. Each tenant constitutes an independent information domain, which has separate administrative ownership, policies and users. Data assets belonging to different tenants are isolated logically from each other.
- *Aggregation:* This entity is a configuration profile, which defines the policies and capabilities for managing the data in storage. It specifies the details of one or more cloud platforms (address, credentials, etc.) that are being used for physical storage. It also designates various characteristics for maintaining and accessing data, such as integrity checking procedures or rendering properties that are relevant for the specific use case.
- *Docket:* This entity is a grouping of preserved entities; it is analogous to a directory in a file system.
- *AIP:* This entity is the fundamental preservation entity in OAIS. An AIP has a name (aipName), as specified in the hierarchical path, a logical identifier (aipLogicalId), and a version identifier (aipVersionId). Multiple versions of the same AIP are distinguished by their aipVersionId, while the

aipLogicalId is common to all versions of the same AIP. The combination of aipLogicalId and aipVersionId is globally unique. When an AIP is moved to a different docket, its Ids remain the same, this enabling continued access. Each AIP is associated with a specific aggregation, and is replicated to all the storage clouds configured in the aggregation.

Aggregations are configured based on the needs of the tenant. The AIPs in a given aggregation can be viewed as a collection of information assets that share the same characteristics and are managed together and in the same fashion. In that sense, aggregations can be considered as part of the service layer.

Users can access their data without needing to know the configuration details held in the aggregation. It is the responsibility of the storage service layer, i.e. PDS Cloud, to interpret the aggregation's configuration profile in order to access the specific cloud platform(s) and map the logical dockets and objects to the physical name space of each specific cloud. So although changes to an aggregation's configuration over time affect how it is handled by the storage service layer, they do not affect the user application interface.

## 4.6 Content-Aware Long-Term Data Protection

The preservation of data over long timeframes poses a series of unique problems for the protection of sensitive content. This includes both commercially valuable data and that covered by current and future data protection legislation. Initially ENSURE developed a set of scenarios focussing on the effects on data protection of changes in the legal, social, political, and technological landscape over long periods of time and then considered how to address these threats. For example, in order to deal with technological issues such as the cracking of an encryption algorithm, the system is designed with an automatic re-encryption method to preserve confidentiality.

A key challenge is in the design of a future-proof access control mechanism. The mechanism must be flexible enough to support different types of access control models (such as role-based access control (RBAC) or lattice-based access control (LBAC)) and must be able to deal with syntactic and semantic changes (e.g. changes in file formats, application domain concepts, user roles, etc.). Given these constraints, an access control engine that implements the OASIS XACML v2 specification, including hierarchical resource profiles [21] and multiple resource profiles [22] was chosen. By supporting these profiles, access control policies for hierarchical resources can be combined into a single authorisation decision, thus simplifying access control policy management, and content filtering can be applied to DIPs so that only those parts of the original AIP to which the requesting user has access are delivered to him/her. The engine was extended to support hierarchical subject attributes (i.e. hierarchical roles) and to support concepts of purpose of access. The latter are not supported by the standard XACML specification but have been identified as necessary to support the access control policies identified in the use cases.

In order to make writing access control and privacy policies as simple as possible, the RDF Storage component and ontology framework described in Section 4.3 is relied upon extensively. Which policies apply for a given authorisation request can be determined by exploiting the metadata produced by the Information Preparation component. Furthermore, policy rules can refer to attributes that are not specified as part of the access control request (e.g. attributes related to resource content, such as which patient a medical record pertains to). Policies may also rely on the ontology framework for classification of security and privacy-related concepts and to deal with potential changes in domain-specific concepts that might otherwise require access control policies to be rewritten.

A plug-in based obligation handler framework, which is integrated into the authorisation engine, is relied upon to deal with encryption, de-identification and other security and privacy obligations specific to the application domain.

The governing access control policies are ingested in the same way as regular AIPs and can be interchanged or updated at run-time, thereby changing the policies that are in effect and making the data protection system future-proof.

# 5. HEALTH CARE USE CASE EXAMPLE: DIGITAL PATHOLOGY

Long-term data preservation is vital for the Healthcare domain, just as it is for the sub-domain of Digital Pathology. The term Digital Pathology is used to describe the current trend amongst pathology departments to digitise pathology glass slides.

As the pathology glass slides are scanned at high resolution, a large amount of data is generated, up to 2TB per day, and this data must be preserved. As well as storing the digitised glass slides (aka Whole Slide Images (WSI)), the preservation system must store other objects, such as documents (e.g. reports, order forms), the results of image analysis applications, as well as the corresponding case, patient, and slide metadata. These different objects will be stored in the Digital Images and Communications in Medicine (DICOM) format.

As multiple sites may work and store data for the same customer, it must be possible to access the central preservation system from all these sites. Typically, newly created images and objects are likely to be accessed more frequently (e.g. for QA review) than older ones. Pathologists may retrieve images and objects via a digital pathology application, or patient histories, which could be relevant to the diagnosis of a recent case, directly via the preservation system's web interface. While researchers interested in developing image analysis applications or carrying out data analyses for scientific purposes also need access.

All users require real-time search and browsing of the preservation system, but not all users have the same access rights to all the objects. Therefore the system supports user authentication and authorisation with role- or profile- based access to the objects, in order to protect patient privacy.

As part of the image life-cycle management, it should be possible to transcode, transform or process images without exporting the entire (large) image files from the preservation system, in order to, reduce the cost of storage or save bandwidth. The ENSURE system can store image analysis applications as Virtual Applications and therefore offer image processing close to the stored data. Transformations, whether implemented as automatic or manual workflow steps, may be triggered by external events that the preservation system is configured to watch for, such as changes in regulations that require either a longer or shorter mandatory image storage period. Therefore, the preservation

system needs to allow images and other objects to be updated in a controlled manner.

## 6. CONCLUSION

In this paper we have provided a general description of the ENSURE system which aims at helping organisations in the health care, clinical trials and financial sectors to prepare and evaluate cost effective preservation plans and build corresponding flexible archival systems based on a set of available plug-ins. The system stores the organisation's content in public or private clouds while maintaining protected access to sensitive data. In addition, it supports a set of preservation-related ontologies, which provide a flexible and future-proof way to search for archived data. Finally, its technical registry, which is based on Linked Data and connected to external registries, is capable of detecting environmental changes that might affect the archival system and require a change in the preservation plan. Such reconfiguration of a live system is also supported. A prototype system implementing the presented architecture has been developed during the two first years of the project and its evaluation by partners in the health care, clinical trials and financial sectors is planned for the last year of the project.

## 7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270000 - ENSURE.

## 8. REFERENCES

[1] ENSURE, DOI=http://ensure-fp7.eu/

[2] The Consultative Committee for Space Data Systems (CCSDA). 2012. Reference Model for an Open Archival Information System (OAIS) – Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2. Also available as ISO Standard 14721:2012. http://public.ccsds.org/publications/archive/650x0m2.pdf.

[3] Coello Coello, C.A., Lamont G.B., Van Veldhuizen, D.A. 2007. Evolutionary Algorithms for Solving Multi-Objective Problems. Springer, New York, USA, second edition, September 2007. ISBN 978-0-387-33254-3.

[4] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. 2002. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. In *IEEE Transactions on Evolutionary Computation*, Volume 6 Issue 2, April 2002. IEEE Press, Piscataway, NJ, USA.

[5] Friedrich, T., Kroeger, T., and Neumann, F. 2011. Weighted Preferences in Evolutionary Multi-Objective Optimization. In *AI'11: Proceedings of the 24th international conference on Advances in Artificial Intelligence.* Springer. LNCS 7106.

[6] Lukasiewycz, M., Glaß, M., Reimann F., and Teich J. 2011. Opt4J – A Modular Framework for Meta-heuristic Optimization. In *GECCO'11: Proceedings of the 13th annual conference on Genetic and Evolutionary Computing* (Dublin, Ireland, July 2011). ACM, New York, NY, USA.

[7] NEPOMUK – Social Semantic Desktop, funded under FP6-ICT Programme, DOI= http://nepomuk.semanticdesktop.org/nepomuk/.

[8] Hartung, M. and Gross, A. and Rahm, E.. 2013. *COnto-*Diff: Generation of Complex Evolution Mappings for Life Science Ontologies. *Journal of Biomedical Informatics.*

[9] Duranti, L. 1989. *Diplomatics: New uses for an old science, part I Archivaria*, 1(28).

[10] Duranti, L. 2002. *The concept of electronic record. In Preservation of the integrity of electronic records*, pages 9–22. Springer.

[11] Engeström, Y. 1987. Learning by expanding. An activity-theoretical approach to developmental research.

[12] Hill, G. 2004. An information-theoretic model of customer information quality. *Proceedings of the Decision Support Systems Conference*, Prato, Italy, July, pages 1–3. Citeseer.

[13] Klir, G. J. 2003. An update on generalized information theory. *Proceedings of ISIPTA 03*, pages 321–334.

[14] Shannon, C. E., and Weaver W. 1949. *The mathematical theory of information*.

[15] Star, S. L. 2010. *This is not a boundary object: Reflections on the origin of a concept. Science, Technology & Human Values*, 35(5):601–617.

[16] Star, S. L., and Griesemer, J. R. 1989. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science*, 19(3):387–420.

[17] Wang, R.Y., and Strong, D.M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–34, Spring.

[18] Rabinovici-Cohen, S., Factor, M., Naor, D., Ramati, L., Reshef, P., Ronen, S., Satran, J., and Giaretta, D. July/September 2008. Preservation DataStores,: New storage paradigm for preservation environments. *IBM Journal of Research and Development*, *Special Issue on Storage Technologies and Systems*, 52(4/5):389–399.

[19] Rabinovici-Cohen, S., Marberg, J., Nagin, K., and Pease, D., March 2013. PDS Cloud: Long term digital preservation in the cloud. In IC2E 2013: *Proceedings of the IEEE International Conference on Cloud Engineering*, San Francisco, CA.

[20] Jclouds. http://jclouds.org.

[21] Anderson, A., 2005. Hierarchical resource profile of XACML v2.0. DOI=http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-hier-profile-spec-os.pdf.

[22] Anderson, A., 2005. Multiple resource profile of XACML v2.0. DOI=http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-mult-profile-spec-os.pdf.

[23] Tarrant, D., and Carr, L. 2012. LDS3: Applying Digital Preservation Principals to Linked Data Systems. *Ninth International Conference on Digital Preservation (iPres2012)*, Toronto, CA.

[24] Unified digital format registry (UDFR). DOI=http://udfr.cdlib.org/.