Preservation Services Planning: A Decision Support Framework

Ingemar Andersson, Göran Lindqvist, Frode Randers
Luleå university of technology, Luleå, Sweden
ingemar.andersson@ltu.se
goran.lindqvist@ldb-centrum.se
frode.randers@ltu.se

Abstract: Commercial organizations are experiencing a growing need to access business-critical data in the longer term of their operations. Governmental regulations as well as commercial interests influence this need. Organizations are willing to procure cost-effective services to this end - services that are increasingly based as public or private cloud solutions. With the advent of autonomous cloud services comes the possibility to assemble (mix and match) preservation services in a workflow-based service-oriented solution. Following the interaction with information managers in (three) commercial organizations operating in different markets and after a review of current literature, we have revealed a lack of comprehensive guidelines and decision support in service selection as part of preservation planning. Existing models and frameworks used for assessing the quality of preservation services either manage performance-based features that service provider's offer or the technical details of the preservation actions themselves. In this paper we present our preservation-planning framework (Preserv-Qual) that addresses the need for decision support in the selection of preservation services that explicitly acknowledge the differences among aspects of information use within an organization. We describe the outcome from an evaluation of the framework in three commercial organisations as a service quality assessment and decision support tool. This paper shows how our framework supports the use of existing and proven methods, models and principles for service assessment, digital preservation and decision support.

Keywords: Digital preservation**,** Preservation planning, Decision-making, Cloud computing

1. Introduction

The problem of ensuring future access to digital objects, referred to as Digital Preservation (DP), is a multi-faceted effort related to preserving a set of qualities of the digital object to the future users of the object. A typical quality to preserve could for instance be authenticity (Nilsson 2009; Duranti 2005). With an increase in both the need to store information as well as the growth of data volume, cost-effective preservation solutions have become more interesting. Issues associated with saving data over a longer period of time have traditionally been the focus of libraries, archives, government agencies, and academic institutions (Ross 2012; Anderson 2008). Recently, interest in this area also comes from commercial organizations. This interest is among other things driven by requirements for traceability of business transactions, legal obligations, etc. Being able to access data over a longer period of time has become an important part of their business model. This is especially true in businesses in the financial and medical care sectors, where evidence of the validity of relied upon data is crucial (Edelstein et al 2011). Until now, solutions for long-term digital preservation have been designed as monolithic systems located in a discrete environment. The development of scalable solutions designed as cloud services has become an alternative for long-term digital preservation (DuraSpace 2014; Preservica 2014). The ability to assemble (mix and match) autonomous preservation services, known as cloud broker solutions (Hogan 2011), has also become feasible. In keeping pace with this progress, there has been a rapid proliferation of vendors that complies with the new service model. This places new challenges on the preservation planning process in terms of how to compose preservation services.

Among the *challenges* that need to be addressed you will find; having a variety of service providers, how do you choose among them, selecting to whom you will entrust your data for a long period of time? Should you choose a single service provider or opt for a solution of configurations of services from different providers (the mix-and-match approach)? It is no longer just a question of the sustainability of the individual service provider infrastructure, nor of the issues surrounding the data itself (such as choice of file format, software, or even security). As the environment in which all organizations operate is subject to change over time, the requirements regarding what to preserve, how to preserve, and why to preserve may change as well. Because of these social and technical factors of change, we need to capture the purpose of preserving the information into the planning

process itself due to the predictive nature of the problem of ensuring future access to information. As a result of this, the act of planning for digital preservation using a set of services becomes more complex.

The objective of this research is to develop and demonstrate a framework (Preserv-Qual) as a foundation for design of decision support systems supporting the selection of composite preservation services. The decision support is based on the assessment of intersubjective factors such as staff turnover and economic stability of the individual service providers as well as technical factors such as choice of storage technology. The key motivation to our work is outlined above.

The development of the framework is based on use cases from three different business cases. The organizations are partners in a large project we were involved in (ENSURE 2014). Based on a review of existing literature, the framework is influenced by existing methods for assessing quality of service and recommended practices for assessing the trustworthiness of digital preservation solutions.

The paper is organized as follows; in the next section we describe background and prior work in digital preservation planning and different quality aspects related to digital preservation systems and cloud services. Thereafter, we describe an excerpt of the usage scenarios and presentation of findings from the project. This is followed by the presentation of our framework. The paper continues with a description of the application of the framework and a condensed evaluation. In the last section we discuss the implication of the framework and directions for future work.

2. Background and Related Work

2.1 Digital preservation planning: quality of service

Development of digital preservation systems is not new. In 1996 the Consultative Committee for Space Data Systems developed OAIS (CCSDS 2012), a high-level model for the operation of archives. Within the preservation area it has been important to establish *trust* by verifying whether a preservation system fulfills the OAIS standard. For this, various methods as DRAMBORA (McHugh et al. 2008) and auditing frameworks as Nestor (Nestor 2006) and Trusted Digital Repository (TDR) (TDR 2011) have been developed.

Alongside, methods for assessing the quality of services have been developed. An overwhelming majority of these have focused on benchmark tests and provides a catalogue of metrics and methods appropriate for performance-based assessments, such as CloudCmp (Li 2010), SMICloud (Garg 2013).

Figure 1 shows an illustration of the preservation planning landscape for a cloud broker preservation solution. At one end we have a client organization, according to the OAIS called "Producer" in need of preservation of a digital collection. The client organizations are also part of the entity called "Consumers" i.e. users of the solution. Other kind of consumers is authorities that verify regulatory compliance of data. Between these stakeholders, we have service providers (SP) that provides technical services required to make data available over time. These services are divided according to the OAIS (CCSDS 2012) in the entities of ingest, data management, archival storage, management, and access (see Figure 1). The Producer organization is responsible for the process of selecting the appropriate mix of services based on the characteristics and the purpose of use of data.

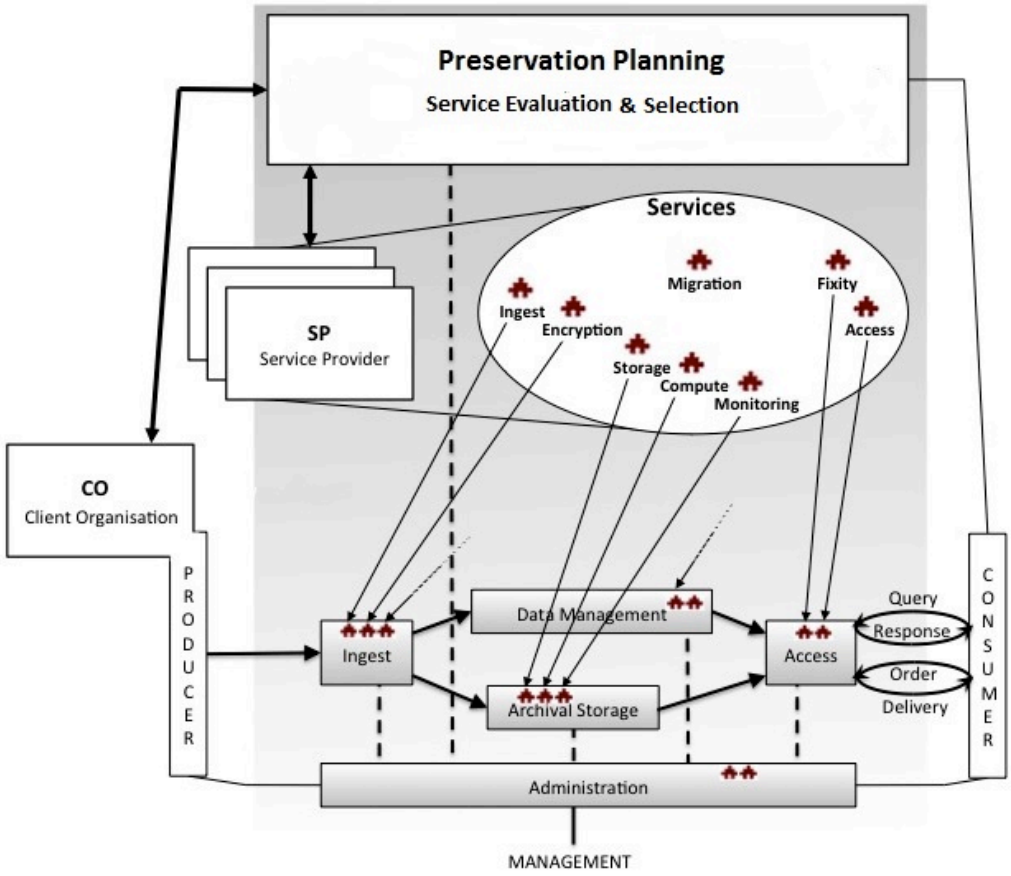

Figure 1: Preservation services planning landscape

Challenges in this preservation planning process do not only include the choice of a preservation strategy, such as the choice of an appropriate file format or the choice of preservation actions – such as migration tools (Becker et al. 2009; Farquhar & Hockx-Yu 2008). Additionally, you now also face the choice of storage and compute services, services for verification of integrity, services for encryption and decryption, services for monitoring the composite service, services for ensuring global security policies over all constituent sub-services, services for ingesting and disseminating information, etc. Since the choice of preservation services composed of a set of these sub-services has to be made in light of cost and quality aspects, risk mitigation has to be addressed. This is a risk management problem that should be supported by tools for decision-making (Ross 2012).

What is needed is a framework that takes a more holistic and comprehensive view of the preservation planning process – a framework that provides guidelines for client organizations in choosing an optimal preservation plan reflecting both trust and performance of the services offered by the individual service providers.

3. A decision support framework for preservation services selection

The development of our framework (Preserv-Qual) is part of the ENSURE cloud-broker digital preservation solution (ENSURE 2014) which is a solution composed of autonomous services where especially the distributed storage and compute services are key components of interest. Other component parts of the solution are services for integrity checking (fixity), transformation, and encryption offered by different service providers (SP). Our framework is part of a decision support component, a Configurator, as an aid for business organizations in the selection of the appropriate (mix and match) of services based on quality. The mission of the configurator is to be able to create an optimal solution for its commercial customer (i.e. the owner of the data) and as such to strike a balance between, on one hand, cost and economic performance and, on the other hand, quality of service. A condensed version of the ENSURE system process is: 1) capture user input as basic business requirements and data policies, 2) generate candidate services as output, 3) evaluate those outputs, 4) select the outputs for the user that is suitable for decision-making, 5) presentation of

candidate solutions to the user for selection, 6) install the selected services corresponding to the chosen configuration in the preservation runtime system.

Development of Preserv-Qual is based on three different use-cases from clinical-trial, healthcare, and financial services obtained through discussions and interviews distilled from separate revisions of the interviews. Here is a condensed excerpt from two different use-cases (ENSURE scenarios 2014) that influenced the development of our framework.

Scenario 1: clinical trials are conducted to bring new drugs to market. Data from medical records are the basis of these experiments and can be used for future reviews. Clinical trial data must be kept for at least 15 years for regulatory reasons in a way that ensures its authenticity, viability and security that ensures authenticity and validity of the stored data. Beyond the regulatory requirements, there is also a desire to ensure the usability, convenient access to the data, so that audits or inspections may be performed in a cost effective manner.

Scenario 2: the financial sector is characterized by ever increasing volumes of high frequency market and transaction data. During the past decade, a particular focus has also been placed on research and development in the financial business performance improvements of IT infrastructure. The financial sector is characterized by a variety of rules and obligations at the national and at the EU level – to act in accordance to rules and legal norms – which has gained increasing importance. Regulations also encompass the archiving of obligations regarding trade and customer information in investment advice, among other things as a countermeasure to fraud and money laundering. In particular, any information received from – or provided to – the client must be maintained for the entire duration of the contractual relationship beyond the legal minimum retention period of five years. Compliance with all such rules and obligations are monitored and reviewed on an annual basis by the regulatory authorities responsible for these controls and their implementation.

In summary, a number of observations can be made from usage scenarios. An important observation is that there are different motives for using preserved information. This leads to that different quality requirements must be met to varying degrees depending on the purpose of use. Based on an analysis of scenarios, related work and literature review, we identified three key dimensions that have influenced our framework (Andersson et al. 2014). Table 1 captures these dimensions.

Table 1 Summarizes the conceptual dimensions that have influenced our framework.

Dimension	**Description**
Trustworthiness	Determines the confidence of services in a preservation perspective. Defined by quality factors: authenticity, viability, and security.
Quality of Service	Estimation of performance-based factors and the ability to move data. Defined by quality factors: accessibility and portability
Purpose of Use (PoU)	This determines the use of preserved data. Defined by different purposes: evidential, historic, and business.

Figure 2 presents a holistic view of the Preserv-Qual framework in its context; supporting a business organization in the process of selecting an optimal cloud broker DP solution for its purpose of use based on quality. The figure shows how the different quality dimensions are used in the framework. The framework is described by the splitting into the layers of Cloud Service, Audit, Configuration, Run-time, and Cloud Consumer Layer.

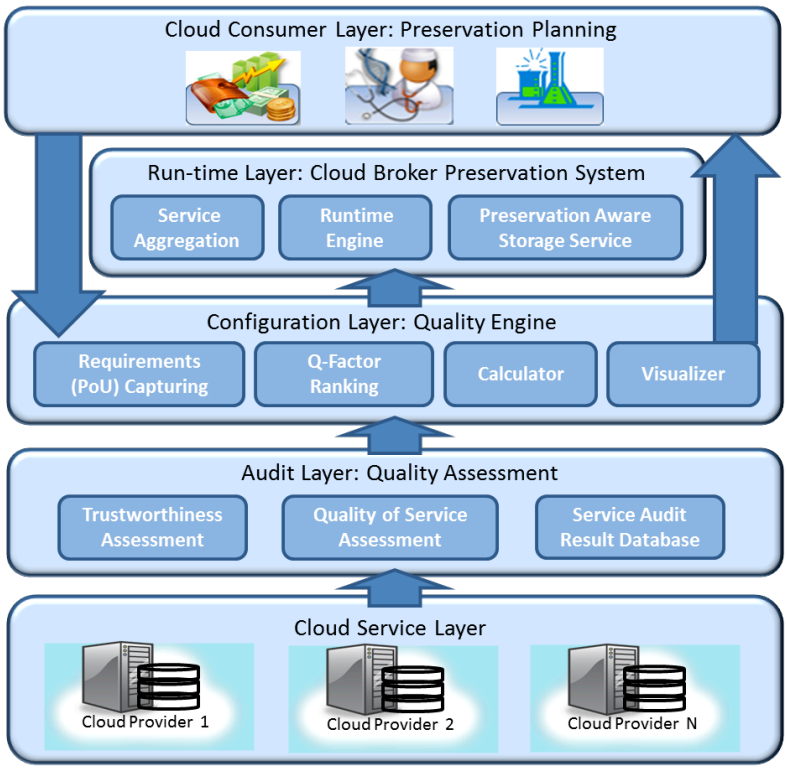

Figure 2: The Preserv-Qual framework

3.1 Cloud service layer

The services that will be included in the service-brokered DP solution are storage and compute services, migration services, fixity services to check accidental corruption of content or bit rot, and various encryption services to protect against unauthorized access and malevolent corruption of content. Different individual service providers may offer these services. The storage and compute services are the cornerstones in the solution and will support other services, e.g. a migration service will use the computation service to get work done. The services are selected as candidate services and registered in the audit result database (QE database).

3.2 Audit layer – quality assessment

This layer is responsible for carrying out the quality assessment and keeping track of the candidate service quality measurement result. Different types of services require different metrics and assessment techniques. It is necessary to assess and compare the same types of services in the same way, using the same metrics and the same assessment scale. The framework supports a classification of quality factors (Q-factors) in two dimensions *Trustworthiness* and *Quality of Service*. Trustworthiness represents Q-factors and metrics related to the management of digital objects in a reliable manner and refers to mechanisms, procedures, staff competence and organizational viability. Related to the Quality of Service dimension are Q-factors and metrics that enable measuring of performance and portability. The results of the service assessments are registered in the QE database used as input to the Quality Engine in the configuration layer.

3.3 Configuration layer - quality engine

This is a central part of the framework that is responsible for capturing basic requirements, calculation of quality scores and to express the outcome of the quality assessment to the users charged with making decisions. Examples of basic requirements are data transformation policies, geographical restriction on placement of data and the grouping of usage needs by defining *Purpose of Use* (PoU). The configuration layer will then produce proposals on various parameterized preservation plans that contain the basic requirements and the available services which conform to the requirements. The outcomes from the Q-factor ranking component are each factor's weight of importance in relation to PoU. The Q-factor value is obtained by a pairwise comparison of each value and a mathematical calculation in accordance with the Analytic Hierarchy Process (AHP) (Saaty 2008). For each Q-factor adequate metrics are identified depending on the type of service that we refer to. The sum of metric values (audit data) obtained by the service measurement instrument defines the service fulfilment rate

for each Q-factor. The calculator component computes a quality score based on service audit data and Q-factor ranking. Input to this layer is a parameterized preservation plan proposal with service specification, triggering events, digital object specification, basic requirements, and defined PoU. A graphical user interface (GUI) presents detailed data from quality measurement results for each service that is part of the preservation plan proposal. The results are presented to preservation planners by scores, graphical charts and quality risk expressions and the quality results are presented along with the cost assessment. The preservation services plan that match the client organizational needs is selected as the preservation service execution plan as input to run-time layer. The objective of the configuration layer is to be able to create an optimal solution for the organization.

3.4 Cloud consumer layer - preservation planning

This layer is responsible for communicating with the users of the system. The main objective of the framework is to support the decision-making of users of the system in selecting the preservation plan that best fits the organizational needs. The client organization can specify the basic operational requirements and policies as how long each type of data must be preserved affecting horizon of configuration, data transforming policy (e.g. compression limitations that affect usability) and if there are geographical restrictions on placement of data. The requirements are part of a parameterized preservation plan as input to the configuration layer. Users responsible for preservation planning are presented with the quality assessment result of each proposed solution with access to details by the visualizer GUI. The users can either modify initial requirements and policies and execute the plan another loop in the configuration layer, or choose a solution from the proposed. If needed in the future, the organization is able to execute a reconfiguration process with a new set of requirements. The framework has been validated in use cases from financial, healthcare and clinical trials sector.

3.5 Run-time layer – a cloud broker preservation system

The run-time layer is responsible for instantiating a preservation system based on the selected services (plugins) available in the repository according to the selected preservation plan. The configurator provides a preservation plan that invokes a plugin manager in the service aggregation component to install the plugins requested by the preservation plan, and then passes control to the workflow engine in the runtime engine component. The storage service component is based on the Preservation DataStores in the Cloud (PDS Cloud) (Rabinovici-Cohen 2013). PDS Cloud is a preservation aware storage service infrastructure component that provides an abstraction over multiple cloud storage and compute providers.

4 Application and evaluation

The application of the framework Preserv-Qual is divided into two major phases. The first phase in preparation of quality assessments, providing data to the QE database that stores the results of the quality measurements of potential services. These services are considered to be candidate services, which may – or may not - be part of a future DP solution. The second phase uses the information in the QE database for calculating the quality score for the aggregated services in each DP solution proposal. Figures 3 and 4 show how the Preserv-Qual framework can be operationalized.

4.1 Audit layer - quality of service assessment

Each candidate services that can be included in a proposed DP solution must be assessed, where each class of service requires a specific type of quality assessment. In Figure 3 the main components of the first phase are presented. The “evaluate quality factor for purpose of use” is the Q-factor ranking component. Another activity is related to the measurement of a Storage and Compute (S&C) service in two different process activities. The first process relates to the measurement of the Trustworthiness dimension with related Q-factors of authenticity, viability, and security. The Trustworthy Digital Repository Checklist (TDR 2011) is a suitable instrument for performing this assessment. The TDR-checklist supports the assessment of service mechanisms that span from organizational staff competence and financial strength to technical infrastructure mechanisms. The second process of the S&C measurement is related to the *Quality of Service* dimension with related Q-factors of accessibility and portability. Suitable instruments for this measurement are cloud service benchmark tools as CloudCmp (Li at al. 2010) and CloudHarmony (CloudHarmony 2014) and service specifications. The last activity is related to the quality assessment for fixity, encryption, and transformation services. The evaluation of fixity and encryption services is based on existing ratings of algorithms. The evaluation of transformation services is done by an internal component that compares properties before and after migration of object. Software that could be used in this activity is the PLANETS testbed (Farquhar & Hockx-Yu 2008).

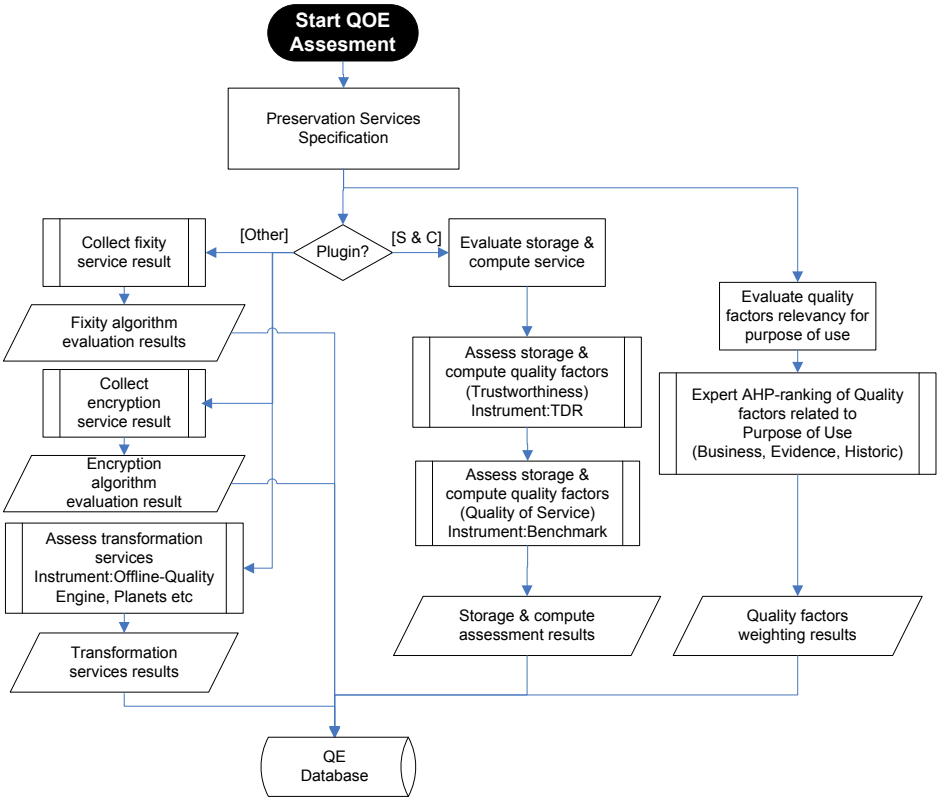

Figure 3: Preserv-Qual – Audit layer: quality of service assessment

4.2 Configuration layer - quality engine

This phase (figure 4) is triggered by the reception of the preservation plan (GPP). The GPP is an XML-based specification composed of candidate services, *purpose of use*, digital object, basic requirements, and preservation event specifications. A quality score for each GPP is calculated based on the results of various plugin measurements and the result from the Q-factor ranking component obtained from the QE database. Output is a quality score and risk expressions.

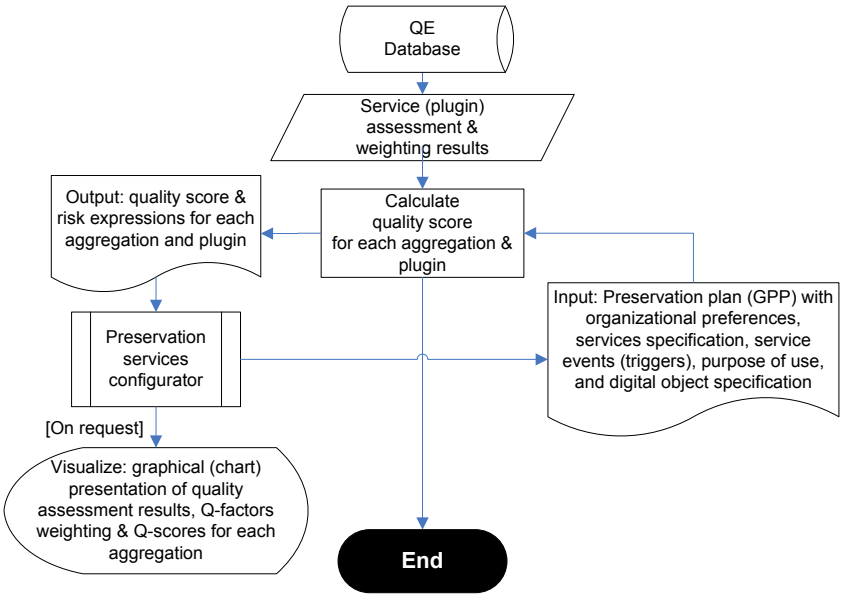

Figure 4: Preserv-Qual – Configuration layer: quality engine

4.3 Evaluation

The evaluation of the ENSURE system (ENSURE evaluation 2014) showed that the use workflows were logical, intuitive and clear. The system was demonstrated for each of the users, followed by a session where end users were able to perform a complete walk through of the system. The applied evaluation method was more qualitative than quantitative. Individual user feedback from the in-house tests was collected. Users were asked to give their subjective assessment of system usability (utility). The Q-factors used in the test of the configurator, which Preserv-Qual is part of, was demonstrated to be relevant to users in this decision-making context. The result of the evaluation phase showed that the system proved to be reliable and trustworthy, and designed to meet the auditability criteria as part of its quality assessment. All necessary Q-factors - authenticity, viability, and security issues were addressed, fulfilling the precondition of trustworthiness.

5 Discussion and Concluding Remarks

Our framework builds upon previous work in the area of digital preservation by supporting existing methods for measuring degree of trust in a preservation system (TDR 2011). The framework has also been influenced by, and provides support for - parts of existing methods for measuring the quality of cloud services (Garg 2013; Li 2010). Preserv-Qual can act as a foundation for design of decision support systems to be used in the preservation planning process. This is the core contribution of our work.

The key motivation for our work is the growing volume of data that has to be saved for a long period in time. This has increased a demand for cost-effective preservation solutions. With this backdrop, cloud-broker solutions composed of autonomous preservation services has become a realistic option. This provides an opportunity for organizations to assemble a preservation solution adapted to needs. This expands the view of the preservation planning process (CCSDS 2012). This is in *contrast to* previous preservation planning approaches that focused on the selection of appropriate file formats and migration services (Becker et al. 2009; Farquhar & Hockx-Yu 2008). We argue that this process has to be supported by a new kind of decision aid – a system that supports the organization in the selection of preservation services based on quality aspects adapted to their individual purpose of use.

A practical application of our framework has been tested in the ENSURE project (ENSURE evaluation 2014) with satisfactory results, but there is still room for further development, such as continued research in identification of appropriate quality factors and improved decision aid in the interpretation of the results. These are the main incentives for further research and development of our framework.

5.1 Acknowledgements

The research leading to these results has received funding from the European Community´s Seventh Framework Programme under the objective "Digital Preservation" (GA 600826)

References

Anderson, M. and Mandelbaum, J. (2008), Planning for the "long term" ... in library time, Digital Archive Preservation and Sustainability (DAPS) Workshop, MSST2008 25th IEEE Symposium on Massive Storage Systems and Technologies, Baltimore, MD, September 22, Available at:<http://www.digitalpreservation.gov/documents/anderson_mandelbaum_daps2008.pdf> [Accessed 24 June 2014]

Andersson, I., Randers, F., & Sein, M. K. (2014) "A Conceptual Framework for Preservation Planning", *International Journal of Digital Curation*, in review process.

Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., & Hofman, H. (2009) "Systematic planning for digital preservation: evaluating potential strategies and building preservation plans", *International Journal on Digital Libraries*, 10(4), 133-157.

CCSDS. (2012) Reference model for an Open Archival Information System (OAIS) Magenta Book CCSDS 650.0-M-2. [online] Available at: <http://public.ccsds.org/publications/archive/650x0m2.pdf>[Accessed 22 May 2014]

CloudHarmony, [online] Available at:<http://cloudharmony.com/>[Accessed 22 May 2014]

Duranti, L. (2005) "The long-term preservation of accurate and authentic digital data: the INTERPARES project", *Data Science Journal*, *4*(25), 106-118.

DuraSpace. DuraCloud. [online] Available at:<http://www.duraspace.org/about_duracloud> [Accessed 22 May 2014]

Edelstein, O., Factor, M., King, R., Risse, T., Salant, E., & Taylor, P. (2011). "Evolving domains, problems and solutions for long term digital preservation". Proceedings of iPRES, 2011.

ENSURE. Enabling kNowledge Sustainability Usability and Recovery for Economic value. [online] Available at: <http://ensure-fp7-plone.fe.up.pt/site/>[Accessed 22 May 2014]
ENSURE evaluation, Activity V Evaluation and conclusion, [online] Available at:<http://ensure-fp7-plone.fe.up.pt/site/deliverables/year-3/activity-v-evaluation-and-conclusion/at_download/file>[Accessed 22 May 2014]
ENSURE scenarios, Activity V Scenario definitions, [online] Available at:<http://ensure-fp7-plone.fe.up.pt/site/deliverables/year-3/activity-v-scenario-definitions/at_download/file>[Accessed 22 May 2014]
Farquhar, A., and Hockx-Yu, H. (2008) ”Planets: Integrated services for digital preservation. Serials”, *The Journal for the Serials Community*, 21(2), 140-145.
Garg, S. K., Versteeg, S., & Buyya, R. (2013) “A framework for ranking of cloud computing services”, *Future Generation Computer Systems*, 29(4), 1012-1023.
Hogan, M., Liu, F., Sokol, A., & Tong, J. (2011) ”Nist cloud computing standards roadmap”, *NIST Special Publication*, 35.
Li, A., Yang, X., Kandula, S., & Zhang, M. (2010) ”CloudCmp: comparing public cloud providers”, In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pp. 1-14.
McHugh, A., Ross, S., Innocenti, P., Ruusalepp, R., & Hofman, H. (2008), “Bringing self assessment home: repository profiling and key lines of enquiry within DRAMBORA”, In Archiving Conference (Vol. 2008, No. 1, pp. 13-19). Society for Imaging Science and Technology.
Nestor. (2006) Catalogue of Criteria for Trusted Digital Repositories, [online] Available at: <http://files.d-nb.de/nestor/materialien/nestor_mat_08-eng.pdf>[Accessed 22 May 2014]
Nilsson, Jörgen (2008) “Preserving useful digital objects for the future” Dissertation, Luleå tekniska universitet.
Preservica. World leading digital preservation technology in the cloud. [online] Available at: <http://preservica.com/>[Accessed 22 May 2014]
Rabinovici-Cohen, S., Marberg, J., Nagin, K., & Pease, D. (2013) “PDS Cloud: Long term digital preservation in the cloud”, In Cloud Engineering (IC2E), 2013 IEEE International Conference, pp. 38-45.
Ross, Seamus. (2012) “Digital preservation, archival science and methodological foundations for digital libraries”, *New Review of Information Networking*, 17(1), 43-68.
Saaty, T. L. (2008) “Decision making with the analytic hierarchy process”, *International journal of services sciences*, 1(1), 83-98.
TDR, Trustworthy Digital Repository Checklist (2011) [online] Available at: <http://public.ccsds.org/publications/archive/652x0m1.pdf>[Accessed 22 May 2014]