FP7-ICT-2007-3-231161

Deliverable ID3.4.1

Service Level Agreements for Preservation Services

Stephen C PhillipsStephen C Phillips (University of Southampton IT Innovation Centre)

2010-02-23

Document administrative table

Document Identifier	PP_WP3_ID3.4.1_SLASpec_R0 Release 0
Filename	PP_WP3_ID3.4.1_SLASpec_R0_v1.00.doc
Workpackage and Task(s)	WP3 Data Management and Processing for Media Preservation WP3T4 – Infrastructures for AV content Storage and Processing
Authors (company)	Stephen C Phillips (ITInnov)
Contributors (company)	Marcel Mattheijer (B&G), Francesco Gallo (Eurix), Matthew Addis, Nena Schädler (ITInnov), Roberto Borgotallo (RAI), Nick Ceton (Technicolor),
Internal Reviewers (company)	Richard Wright (BBC), Giorgio Diminio (RAI)
Date	2010-02-23
Status	Final
Type	Part of Deliverable
Deliverable Nature	R - Report
Dissemination Level	PU - Public
Planned Deliv. Date	M12 - 31/12/2009
Actual Deliv. Date	M14 - 23/02/2010
This IsPartOf	D3.1
This HasPart	
Abstract	This document outlines a framework for monitoring and managing services with a service level agreements (SLA) and proposes a complete list of terms suitable for inclusion in the SLA for a preservation service provider. Modelling techniques for service capacity management are discussed and the results of a survey investigating trust issues is presented. Standards and reference models for computer readable SLAs such as WS-Agreement are compared and the real world experience of managing the relationship between a preservation service provider and their client is documented.

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
V0.01	2009-06-02	First Draft	Outline	Confidential
V0.02	2009-07-16	Added SLA term templates	Outline	Confidential
V0.03	2009-07-22	Structural changes following WP2/3 meeting	Outline	Confidential
V0.04	2009-12-01	Major updates	Draft	Confidential
V0.05	2009-12-18	Major updates	Draft	Confidential
V0.06	2009-12-21	Completed trusted repository chapter	Draft	Confidential
V0.07	2010-01-06	Completed remaining chapters	Final draft	Confidential
V0.08	2010-01-15	Update following review	Final draft	Confidential
V0.09	2010-02-05	Added scope and executive summary, moved and updated questionnaire chapter, other minor corrections	Final draft	Confidential
V0.10	2010-02-09	Final small changes, glossary	Final	Confidential
V1.00	2010-02-23	Finalised - Delivered	Release	Public

Table of contents

Scope

The PrestoPRIME project is researching and developing practical solutions for the long-term preservation of audiovisual digital media objects, programmes and collections, and finding ways to increase access by integrating the media archives with European on-line digital libraries in a digital preservation framework. This will result in a range of tools and services, delivered through a networked Competence Centre.

The preservation of digital audiovisual assets can be performed by a “service provider”, whether this service provider is the same organisation as the producer and consumer, an out-sourced operation but on the same premises, completely out-sourced or even stand-alone. In this context, the interactions of the preservation service with producers and consumers can be defined and managed through service level agreements (SLAs).

Service level agreements may be written into a contract and monitored by a manual process, but a preservation service for digital audiovisual assets necessitates the use of complex IT systems and provides the opportunity for automated monitoring and management of the service in accordance with the SLAs.

This document consists of four main topics:

1. An investigation into what aspects are important when determining whether or not to trust a service provider, and how a customer can be convinced that their assets will be correctly preserved (regardless of the terms in an SLA).

2. How SLAs fit into a managed IT system, how these systems can be automatically monitored and managed, a proposal for suitable SLA terms for a preservation service provider and an overview of how an IT system’s capacity can be managed to successfully meet SLAs.

3. A comparison of many standards and reference models for documenting a service level agreement in a machine readable format.

4. A discussion of the experience gained from the real-world example of Sound and Vision out-sourcing their audiovisual archive to Technicolor.

This results in a detailed proposal for both a vocabulary to describe SLA terms and suitable terms for a preservation service. This and the monitoring and management framework described should be taken into account when designing such a system.

A strong conclusion results from the investigation into trust issues: that an independent expert audit of a service is of value and can cover the vast majority of important factors. The comparison of standards is less conclusive though: WS-Agreement appears to be a good framework but additional work is required to make a judgement.

Finally, the experience from Sound and Vision and Technicolor shows that it is not just SLAs and audits that are important. To maintain a good relationship communication and providing control to the user where possible are essential.

Executive summary

How can digital audiovisual material be preserved in the long-term? Answering this question is the focus of the PrestoPRIME project. The project will research and develop practical solutions for the long-term preservation of digital media objects, programmes and collections, and find ways to increase access by integrating the media archives with European on-line digital libraries in a digital preservation framework. This will result in a range of tools and services, delivered through a networked Competence Centre.

Digital audiovisual material can be stored on a variety of hardware types, from spinning discs in a server to tapes on shelves. Regardless of the hardware format, some company or department has the task of storing, preserving and providing access to the data. The related document PrestoPRIME D2.3.1[1] contains a wealth of information on storage aspects: the variety of architectures and value networks, characteristics of data and how to get it to and from the storage, security aspects and interfaces. It also discusses how a preservation system can be built upon a storage service. This document looks at preservation services, whether out-sourced or in-house and the considerations that are important when defining the service level agreement (SLA) with the service provider.

Defining the service level agreement is just one task of many that must be carried out when defining the relationship between a producer or consumer and an archive. The processes necessary for preservation including those related to service level agreements are expressed in detail in PrestoPRIME D2.2.1.[2]

This document contains two major new pieces of work:

1. A survey of audiovisual service providers, investigating what they consider the important considerations are when determining whether to trust a service provider.

2. A detailed proposal of terms to be included in a service level agreement with a preservation service provider.

Recognising that a service level agreement cannot be sufficient in itself to persuade a content producer or consumer to trust anyone with some data, we conducted a survey amongst professionals from the audiovisual service provider community to see what they thought was most important for convincing someone to use a preservation service.

The top six criteria for that were considered to be “important” or “very important” in this survey were:

- The preservation service has a clear preservation plan (e.g. when to migrate).

- The physical security of the audiovisual (AV) material held in the preservation service.

- The preservation service has a suitable disaster plan including at least one off-site complete copy of all preserved data.

- The staff of the preservation service have well defined and delineated roles and authorisations (e.g. to ensure that only senior staff can make critical changes to the data or system).

- The preservation service has the appropriate number of staff and a professional development plan.
- The preservation service has a succession plan (what should happen if the service ceases to exist).

All these aspects would be investigated by a professional auditor using the TRAC certification scheme. We found that 39% of the respondents were aware of TRAC and that, regardless of prior knowledge, 89% would take into account such an audit certificate when determining whether or not to trust a service provider. We therefore conclude that an audit of a preservation service provider by an independent expert auditor using a scheme such as TRAC is a practical solution to the problem of how to convince someone to use the service.

The majority of this document deals with a proposal of suitable terms to be used in a service level agreement with a preservation service provider. The fact that we are dealing with the preservation of digital objects necessitates the use of IT systems and provides the opportunity for automated, detailed monitoring and management of those systems to provide a well defined quality of service.

Firstly a vocabulary is defined to express the necessary concepts and a discussion of monitoring and management techniques and architectures follows. Fundamentally, there is no point measuring and reporting something that you cannot or do not want to do anything about. Any term appearing in the service level agreement must be monitored but monitoring an IT system is not a trivial task and this document goes into some detail about the different types of monitoring reports and their aggregation and interpretation.

A framework for gathering appropriate SLA terms is presented and has been used in making the proposal of SLA terms presented here in one chapter and an extensive appendix. In total there are:

- 21 capabilities

  Such as “ingestion”, “delivery”, “validation”, “demux”, “fast preview”.

- 12 features of interest and 15 metrics

  Such as “availability of services”, “storage occupation”, “SIP ingestion time”, “DIP conformance”.

- 12 quality of service terms

  Such as a set threshold on the SIP ingestion time.

- 4 constraints

  Such as a maximum number of simultaneous users.

- 6 pricing terms

  Such as a yearly subscription charge and a data movement charge.

- 7 penalty terms

For instance, payable when file integrity is lost.

To support a system that maintains the required quality of service, the SLAs and the monitoring data must be used by the service provider in the capacity management process. Capacity management systems range from “we’ve got another customer: buy some more tapes”, through back of the envelope estimations, spreadsheets and semi-automated models to automatic decision support services. A variety of techniques are discussed in this document with appropriate references for further reading.

Automatic monitoring, reporting and capacity management in complex IT systems can only be achieved if the service level agreement is understood by the system itself. For this reason a review of many web-service and grid standards for encoding SLAs in machine readable form and some related standards is included. There is no clear conclusion from this review. WS-Agreement is a popular choice but it leaves the schema for describing the SLA terms (such as those mentioned above) undefined. This may be seen as a good thing as it does not constrain an implementation but it cannot lead to full cross-domain interoperability or leverage the combined effort of many contributors.

Finally, this document is grounded in the reality of the existing relationship between the Dutch organisations Sound and Vision and Technicolor. Sound and Vision provides access to 700,000 hours of Dutch television, radio, music and film and uses the services of Technicolor to store and preserve this content. The two organisations have an existing service level agreement, some details of which are presented here. More important though is the information on the problems that have been encountered and potential solutions. Many of these solutions rely on giving the user control where possible, such as the choice of fast or slow lanes for download and sharing information: communicating the service’s status so as to set correct expectations.

The commentary on this service provider/client relationship informs us that whereas some business relationship may be quite detached, in the case of a preservation service provider the client must be invited in to understand commercially sensitive information that would not normally be shared. This provides the final reassurance and confidence to use a service provider as well as the necessary information to draw up an exit strategy to guarantee the long-term safekeeping of the material.

1. Introduction

A “preservation service” is a catch-all term for many functions including ingest, storage and access, supporting systems for administration and maintenance of integrity and AV specific services: transcoding for instance. This document explores the issues of what sort of service level agreements are required for a preservation service and what supporting documents would be necessary to establish a trust relationship. It should be noted up front that SLAs and their associated terms are not only relevant to bipartite inter-organisation agreements but are also of importance when negotiating services within a single organisation.

Service level agreements (SLAs) for IT systems of any reasonable complexity need to be managed by other IT systems. Even a basic SLA that states a “99.99% uptime” for a service raises the questions “how is the uptime measured?”, “who measures it?”, “what actions might the service provider take to maintain the uptime?” and “what happens if the guarantee is breached?”

1.1. Document Structure

It is important to realise that an SLA, no matter how detailed, will not provide sufficient assurance for a customer to trust a service provider with their data. The approach of auditing a service provider and using an audit certificate as evidence of trustworthiness is discussed in Chapter 2 and a survey of the AV preservation community’s attitude to trust is described and the results presented.

The key relationships between the service, resources, service level agreements (SLAs) and customer interactions are outlined in Chapter 3. Chapter 4 then provides the necessary background to understanding how SLAs can be managed, discussing the importance of not including objectives or constraints in an SLA that cannot be measured, proposing how to capture the necessary terms using a generic framework and outlining architectural options for measurement and management systems.

Chapter 5 then goes on to use the structure given in Chapter 4 to propose terms that might be used in an SLA with a preservation service provider. These terms are structured around capabilities, features of interest, quality of service terms, constraints, pricing terms and penalty terms. A quite detailed model of an SLA results from this analysis and further information is presented in the Appendix.

For SLAs to be effective the service provider must be able to answer some important questions such as “What resources do I need now and in the future to meet the terms of a proposed SLA?” and “What effect will a change in software or hardware have on the system performance?” Chapter 6 introduces the concept of “capacity management”, showing (from a high level) the effect of certain proposed SLA terms on the required resources. Pointers to existing approaches to this difficult problem are also given.

To effectively manage complex IT systems and provide the service levels set down in SLAs, automatic monitoring and management systems are required. For this reason it is desirable to have the terms of the SLA encoded in a machine readable format. Various standards have been proposed in recent years and Chapter 7 reviews these proposals.

Finally, we note that SLAs for preservation services do exist today and Chapter 8 presents some information about the relationship between Sound and Vision (who maintain the Dutch television archive) and Technicolor who provide the preservation service. Some examples of the terms used in their SLA are shown along with a commentary on the difficult issues faced by these companies in monitoring and managing the quality of service.

1.2. Overview

The focus of this document is on the SLA terms that are essential for the consumer and service provider to understand what is expected of each party in a business relationship (though please note that this includes a relationship between consumer and service provider within a single organisation).

The following services could be imagined from a Preservation Service Provider (PSP):

- **Ingest** - the ability to upload content to the service provider for subsequent storage and access. It should be possible for items uploaded during one or more sessions to be correlated with each other as part of collections.

- **Access** - the ability to get a copy of content already uploaded to the service provider. This includes access to individual items or access to items aggregated into collections.

- **Decoding/Encoding/Transcoding** - the ability to change the format of the content, e.g. as a preservation action such as moving from DV to lossless JPEG2000, or to create access copies needed for users of the content.

- **Wrapping/Transwrapping** - the ability to change the wrapper/container format for the content but leave the essence the same (no decoding and re-encoding of the audio or video). For example, moving from AVI or QuickTime to MXF (or perhaps in the future AXF).

- **Integrity checking/validation** - the ability to independently verify the integrity of the content held by the service provider, e.g. by running content corruption detection tests or by using digital signatures or checksums.

- **Annotation** - the ability to add metadata to the content that is needed by the users of the content or for the service provider to perform preservation actions on the content. Preservation theory might say that when the content is first ingest into the service that all possibly relevant information should also be captured at the same time on what the content is and how to preserve it, in practice it is likely that the content owner may want to add/augment this information during the lifecycle of the content.

- **Access Control** - the ability of the owner of the content to set rules on who can access the content, when and in what form.

- **Transferral** - the ability to package up and deliver one or more items of content in a particular format and send it to some destination, which could be another archive or another service provider. Once the content has been unequivocally verified as successfully transferred, the copy in the archive is deleted.

- **Removal** - the ability to delete content from the service

- **Audit/Report** - the ability to request an audit (technical, financial, process) of the contents held by the service provider on behalf of the content owner. Technical = what is there, what format is it in etc. Financial = how much is the service costing, what are the charges. Process = what actions have been performed on the content whilst at the service provider.

Some of the services listed above include the ability to manipulate or transform the content, e.g. to transcode it. If the software to do this is not available from the service provider, then the service provider might offer the ability to 'upload' the software needed (subject to software licensing constraints and a suitable platform being available at the service provider for executing the software).

In general, nothing should go in/out of the archive or be manipulated within the archive without going through a service interface and that is subject to service management, i.e. it can be monitored and controlled.

An SLA is part of a negotiated agreement between two parties: the customer of the Service and the provider of the Service. The Producer Archive Interface Methodology Abstract Standard from CCSDS[3] uses the terms “Submission Agreement” and “Order Agreement” to cover the complete contract between the parties. The Submission Agreement includes textual descriptions for the following items:

- Information to be transferred (e.g. SIP contents, SIP packaging, data models, Designated Community, legal and contractual aspects);

- Transfer definition (e.g. specification of the Data Submission Sessions);

- Validation definition;

- Change management (e.g. conditions for modification of the agreement, for breaking the agreement);

- Schedule (submission timetable).

PrestoPRIME document D2.2.1[4] has more detail on these processes and artifacts. The SLA terms discussed here in this document can be thought of as terms that codify aspects of the Submission Agreement. For instance, the preservation service is tasked with keeping the content safe: the SLA therefore includes terms that monitor storage failures and corruption.

An SLA may cover one or more Services, e.g. there could be one SLA that covers all the services listed above (it would be long and complicated) or there could be a SLA for each individual service. An SLA should cover the following areas:

- The function performed by the service, i.e. what it does

- How to interact with the service. i.e. how to use it

- Obligations on both the provider and consumer of the service

- Agreed bounds of performance (QoS) for the service

- How to measure the delivery of the service, i.e. what metrics apply
- How deviations are handled (exceptions), i.e. what happens when things go wrong
- Penalties or similar clauses if the SLA is breached by either side.

There are several things to note.

- The SLA contains obligations on both sides. The obligations of the service provider are usually obvious, e.g. to provide a service with appropriate functionality and required levels of availability and performance. However, obligations can also exist on the consumer, e.g. not to submit more than an agreed amount of content to the provider in a given month, submitting content in an agreed format, not to attempt to circumvent security measures etc.
- The SLA defines what happens when things go wrong and what compensation/penalties may apply. No service provider is ever perfect, so it should not be assumed that the service will always be delivered according to the SLA. Indeed, there may be cases where the service provider deliberately chooses to breach an SLA (e.g. it is economically more viable to meet 95% of commitments and pay a few penalties than it is to meet 100% of commitments and have to invest in more expensive or large scale infrastructure to resource the service).

All actions on the Preservation Services (ingest, access, processing) by a customer will cause a workload on a service provider. This workload needs to be measured and managed by the service provider not only to ensure the services remain with the SLAs where possible, but also to ensure any internal activities (backup, integrity checking, migration etc.) are not compromised.

From the perspective of the consumer of the service, the benefits of specifying performance and constraints in SLAs include:

- **Agreed and controlled ingest of content.** Defining exactly what can be ingested into the Preservation Service Provider and when can help the consumer plan and manage the submission process. This helps avoids queues and back-logs at the consumer side, both of which have the potential to put content put at risk. Priority might also be set for different types of content so the Preservation Service Provider can process submissions from the customer in the right order.
- **Agreed and controlled rates of access to content.** This is important as it helps avoids delays, conflicts, and unpredictable QoS when the consumer wants to access their content (there can be a large number of users with differing and competing needs that go beyond the capacity of the Service and hence contention needs to be managed).
- **Defined QoS for preservation actions performed at the Preservation Service Provider.** Many of the Services listed above allow a consumer to perform actions on their content when it is stored at the PSP, e.g. transcoding and transwrapping. With large volumes of content at stake, e.g. 100,000+ hours, the performance of these services is important so there is certainty that they can be applied and will complete within necessary timescales.

From a Preservation Service Provider perspective, SLAs will typically need to include:

- **Limits and priorities for different types of user and/or the content they supply or access.** This is about limiting the total workload on the service or helping to ensure that the workload is manageable, i.e. it doesn't have huge peaks or periods when the service is idle. This allows the PSP to make planned and managed use of their resources and hence have confidence in meeting user needs. It may also be needed to protect or 'ring fence' any specific resources needed for preservation actions – e.g. to ensure that enough capacity is maintained to do a format migration whilst consumers continue to submit or access content. Only if there is a way for the PSP to control workload can they manage priorities for safety, availability and accessibility of the content they hold on behalf of their customers.

2. Trusted Service Provider

Before we get into any discussion of SLA terms themselves, it is important to recognise that even the most detailed SLA containing promises of uptime and minimal data loss may not convince someone to actually use a service provider to store and preserve their precious data. Customers commonly want some evidence that the service provider will still be around in ten years and that they actually do a "good job". How to determine these less-quantifiable aspects is not easy though. By analogy, if you were choosing a kindergarten to look after your child, you would not be satisfied to be given an SLA that said "there is a 99.9% chance your child will be returned to you unharmed at the end of the day", you would also want some independent verification that the kindergarten had the correct child-safety policies in place and that the staff were correctly trained and vetted. For kindergartens this verification would commonly come from an independent government audit but what is the equivalent for a digital repository?

An answer to this question is provided by the result of the international collaboration between the Digital Curation Centre (DCC), the Online Computer Library Center (OCLC), the US National Archives and Records Administration (NARA), Nestor and the US Center for Research Libraries: TRAC (Trusted Repositories Audit & Certification). TRAC has it roots in a joint task force created to develop criteria enabling the identification of digital repositories capable of reliably storing, migrating, and providing access to digital collections, originally sponsored by RLG and NARA only[5]. The basic influences for TRAC came from the OAIS Reference Model and the Report on Trusted Digital Repositories by RLG[6]. Where OAIS lays out fundamental requirements for preservation the RLG report focussed on requirements for the body undertaking the preservation activities[7].

TRAC provides tools for the audit, assessment, and potential certification of digital repositories, establishes the documentation requirements required for audit, delineates a process for certification, and establishes appropriate methodologies for determining the soundness and sustainability of digital repositories. The publication "Trusted Repositories Audit & Certification: Criteria and Checklist"[8] incorporates the sum of knowledge and experience, new ideas, techniques, and tools that resulted from cross-fertilisation between the U.S. and European efforts.

Besides TRAC, another working group is currently advancing the establishment of an ISO standard on which a full audit and certification of digital repositories can be based. Lead by the CCSDS (the same body that gave us OAIS) the venture aims to combine the efforts of TRAC, DRAMBORA[9], Nestor[10] and ISO/IEC 27001:2005[11].

and to standardise the results in the same way as the OAIS Reference Model (ISO 14721)[12]. The working group is expected to publish two documents, one containing metrics to audit and certificate digital repositories and one specifying the requirements for the bodies that actually provide these audit and certification. Consequently the consistency of the audit and certification process is ensured additionally by investigating the expertise and qualification of the auditors as well[13].

The criteria of the TRAC checklist are divided into the following sections:

A. Organisational Infrastructure,

B. Digital Object Management and

C. Technologies, Technical Infrastructure, & Security.

Section A “Organisational Infrastructure” includes characteristics of the repository organisation that affect performance, accountability, and sustainability. These are supposed to be indicators of a digital repository’s comprehensive planning, readiness, ability to address its responsibilities, and trustworthiness. The criteria are organised in the following groups:

A1. Governance and organisational viability,

A2. Organisational structure and staffing,

A3. Procedural accountability and policy framework,

A4. Financial sustainability and

A5. Contracts, licenses, and liabilities.

The section “Digital Object Management” includes both some organisational and technical aspects related to these responsibilities, such as repository functions, processes, and procedures needed to ingest, manage, and provide access to digital objects for the long term. Requirements for these functions are categorized into six groups based on archive functionality:

B1. Ingest: acquisition of content,

B2. Ingest: creation of the archivable package,

B3. Preservation planning,

B4. Archival storage & preservation/maintenance of AIPs,

B5. Information management and

B6. Access management.

Section C describes best practices for data management and security. In total, the criteria measure the adequacy of the repository’s technical infrastructure and its ability to meet object management and security demands of the repository and its digital objects. The requirements are grouped into three layers:

C1. System infrastructure,

C2. Appropriate technologies and

C3. Security.

The requirements throughout the TRAC checklist refer to several documents (policies, procedures, plans, etc.) that a repository should keep current. The following lists contain the minimum required documents that should be provided by the repository.

Documents for section A “Organisational Infrastructure”:

- Contingency plans, succession plans, escrow arrangements (as appropriate),
- Definition of designated community(ies), and policy relating to service levels,
- Policies relating to legal permissions,
- Policies and procedures relating to feedback,

- Financial procedures and
- Policies/procedures relating to challenges to rights.

Documents for section B “Digital Object Management”:

- Procedures related to ingest,
- Process for testing understandability,
- Preservation strategies,
- Storage/migration strategies,
- Policy for recording access actions and
- Policy for access.

Documents for section C “Technologies, Technical Infrastructure, & Security”:

- Processes for media change,
- Change management process,
- Critical change test process,
- Security update process,
- Process to monitor required changes to hardware,
- Process to monitor required changes to software and
- Disaster plans.

The intention is therefore that a company holding a TRAC certificate (or whatever follows the ISO standardisation process) would be able to present it to a potential customer and show with just that one document that a whole host of policies, plans and strategies are in place to protect their customers’ data.

Related Standards

As well as TRAC and the forthcoming ISO specification based on TRAC, DRAMBORA, Nestor and ISO/IEC 27001:2005 there are some other general quality standards and recommendations in the field of IT:

ITIL: Information Technology Infrastructure Library (ITIL) from the United Kingdom's Office of Government Commerce (OGC)[14] which claims to be the most widely adopted set of principles for IT service management worldwide.

COBIT: Control Objectives for Information and Related Technology[15] is a globally recognised and adopted controls-based, value and risk management framework used to support overall IT governance.

ISO 9000: This is a family of standards for quality management systems. It includes ISO 9001:2008 which amongst its requirements has for example: monitoring processes to ensure they are effective, keeping adequate records and facilitating continual improvement.

ISO/IEC 20000:2005: This standard is based on the ITIL service management processes and promotes the adoption of an integrated process approach for service delivery.

SAS 70: Statement on Auditing Standards number 70[16] is a standard audit report, not a checklist, developed by the American Institute of Certified Public Accountants (AICPA). It is mentioned here because of the online storage service providers reported upon in PrestoPRIME D2.3.11, Nirvanix[17] used this as evidence of their trustworthiness.

2.1. Digital Preservation Services Questionnaire

An important aspect in the selection of a service provider for the preservation of digital audiovisual assets is the trustworthiness of the service providers' digital repository. Efforts like TRAC from the Digital Curation Centre have been carried out to identify common audit and certification criteria which can be applied in order to assess how trustworthy a digital preservation repository is.

In this context it is interesting to see how the perception and acceptance of such evaluation methods are within the community of audiovisual preservation service users and providers, and how important the applied criteria are considered to be.

Analysis of the survey results

A survey to answer exactly these questions was conducted amongst the service providers registered on the PrestoPRIME interest group (approximately 200 in number). The survey was open from 2009-12-03 until 2009-12-16 and during that time 36 complete responses were received. The respondents were mainly professionals from the audiovisual digital preservation community. Figure 1 shows which different job positions were indicated.

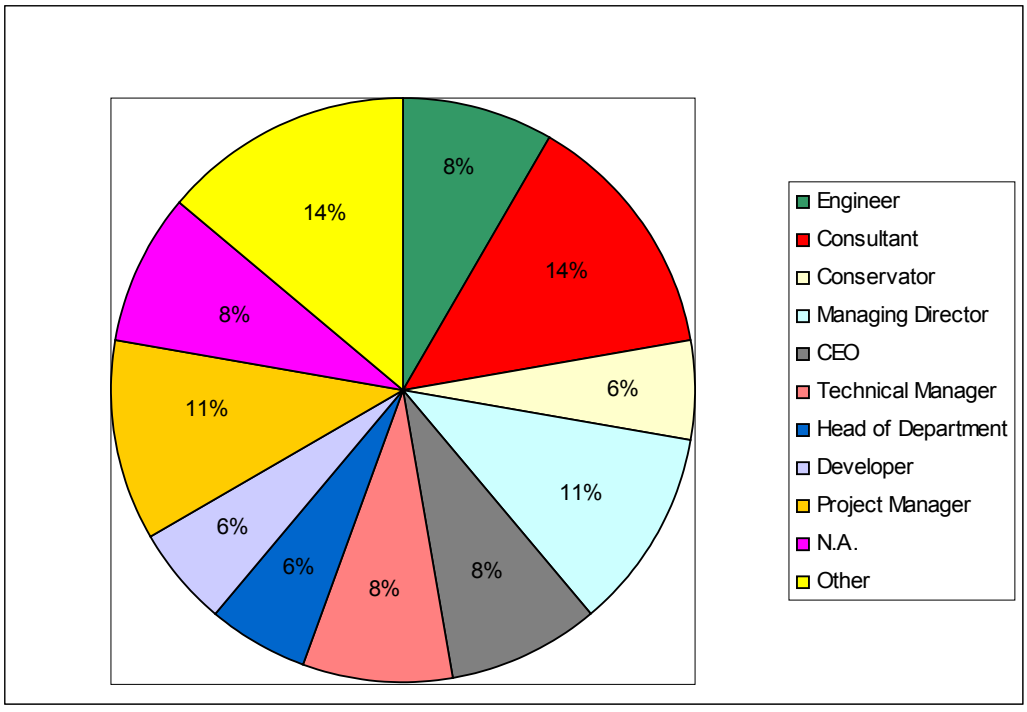


Figure 1: Job position range of respondents

In the course of the survey the respondents were first asked some personal questions about their role in digital audiovisual preservation and if they consider outsourcing of preservation services as an option for their organisation. Figure 2 shows that 22% of the respondents are users and 89% are providers in the preservation process (with some respondents describing themselves as both provider and user). This is to be expected

from the target of the questionnaire. The questionnaire will be re-run in the near future to harvest users' opinions.

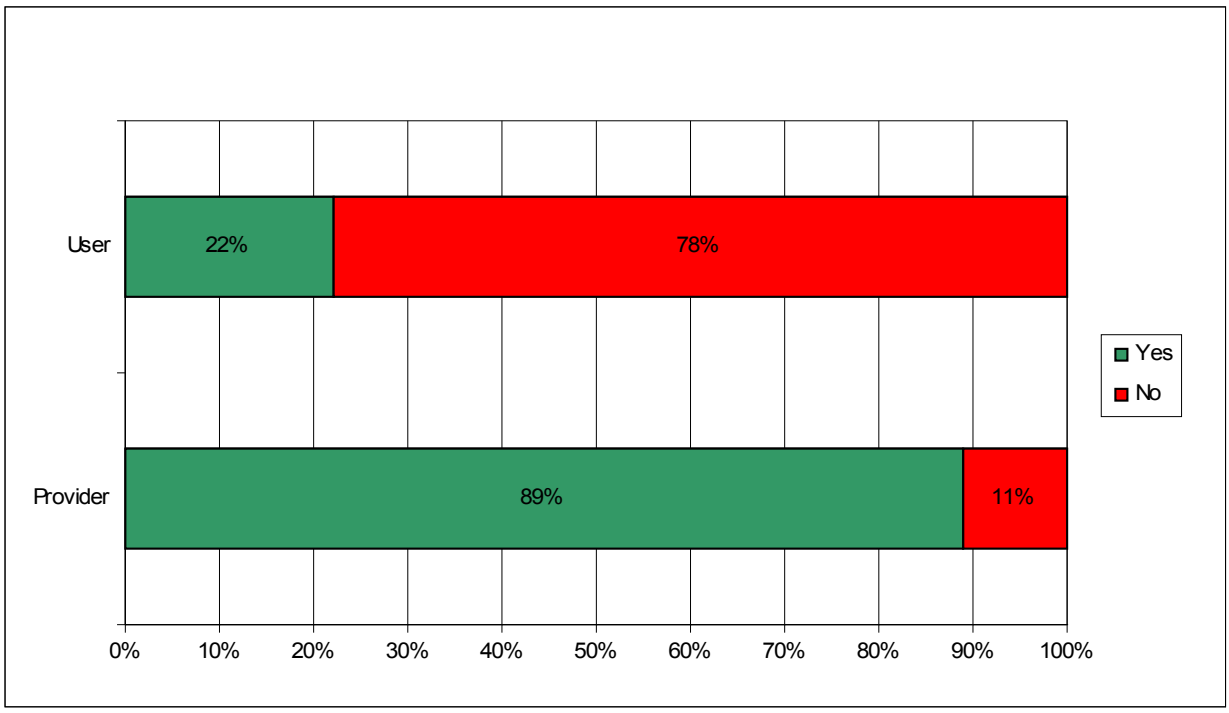


Figure 2: Roles of the respondents in the digital preservation process

In answer to the question "Are you considering out-sourcing some or all of your archiving services" *39% were in favour of out-sourcing*.

The respondents were then asked to rate the importance of different criteria taken from the TRAC checklist. These criteria where similar to TRAC allocated to the different categories. These are "Governance", "Audiovisual Material Management" and "Security".

Figure 3 presents the criteria in the category "Governance" and how their importance has been rated by the respondents. It can be clearly seen that nearly all criteria in this category are considered to be very important for trusting a digital preservation repository. Especially the criteria that are pointing towards the future and enable the repository to maintain its sustainability, such as financial and technological planning are favoured. However, it is perhaps surprising that one third of respondents were not very concerned about tracking and management of intellectual property rights (IPR).

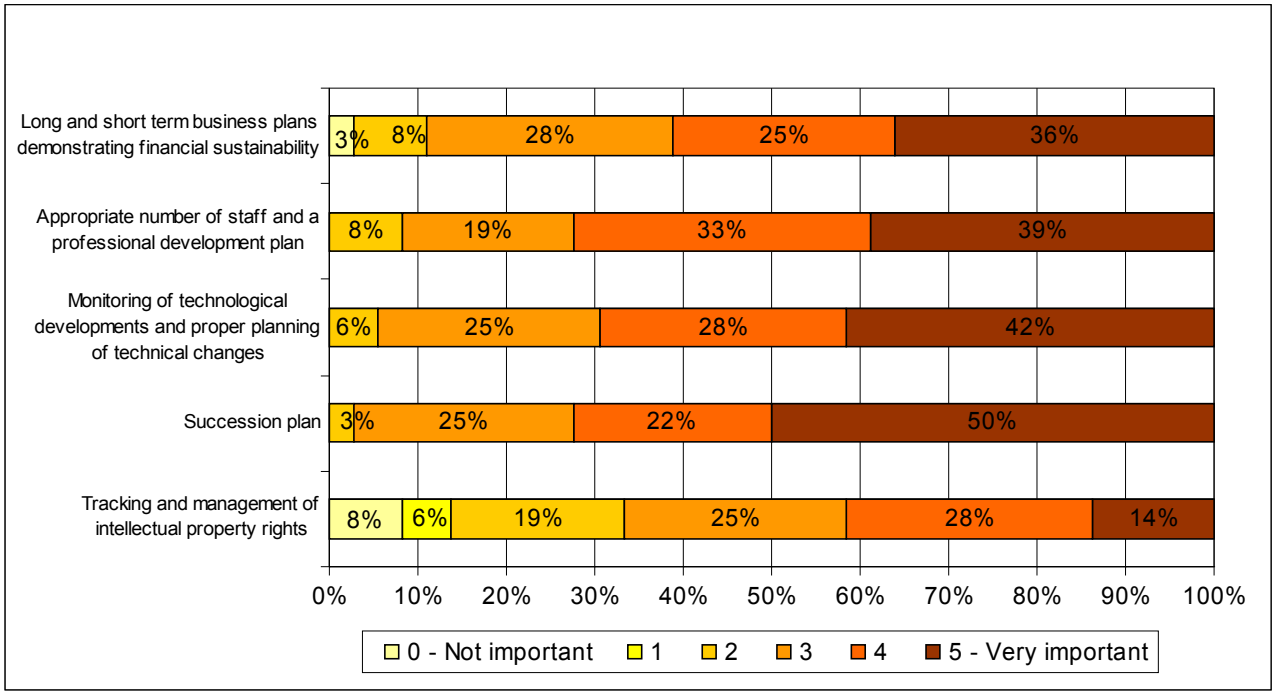


Figure 3: Rated importance of the criteria in the category "Governance"

The percentages for "Audiovisual Material Management" are illustrated in Figure 4. Accordingly, clear preservation planning, active monitoring and logging are most important activities of a digital repository in this category. There is a certain lack of concern about authenticating the source of ingested material though which fits with the previous response relating to IPR.

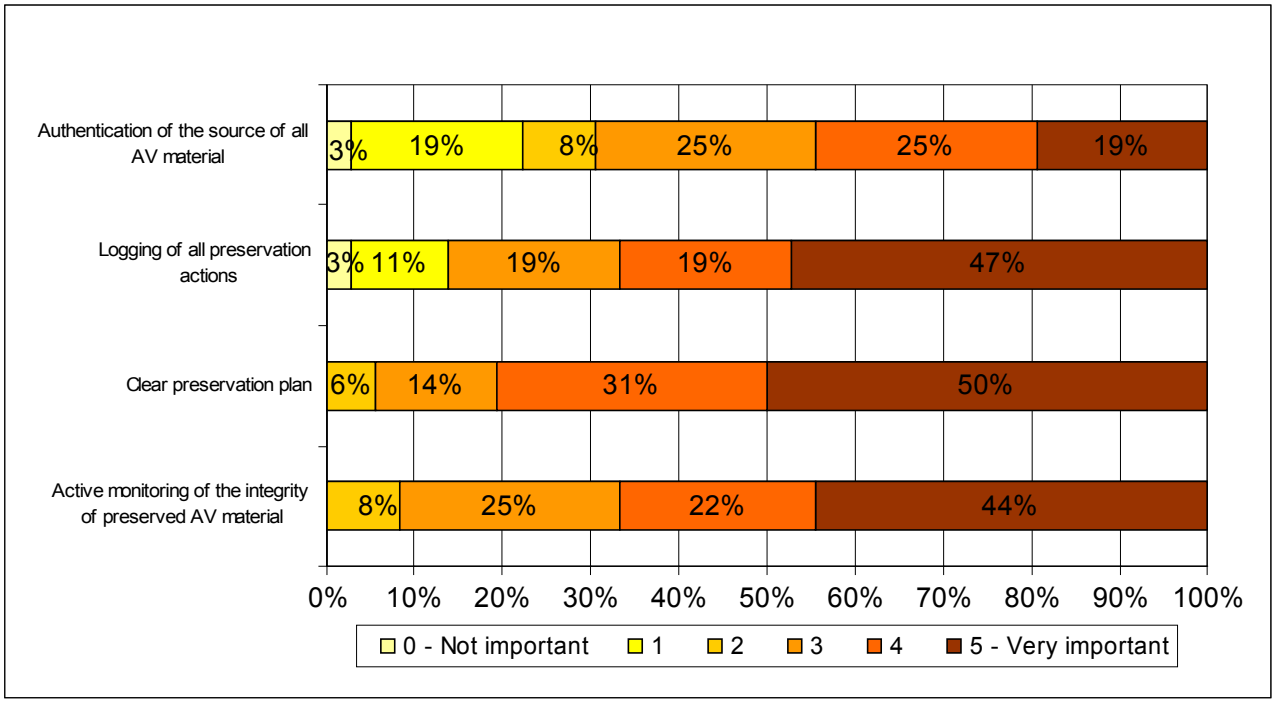


Figure 4: Rated importance of the criteria in the category "Audiovisual Material Management"

Figure 5 shows the rating of different security aspects. Not surprisingly all aspects have a similar level of importance for the trustworthiness, however disaster planning and physical security are the leaders.

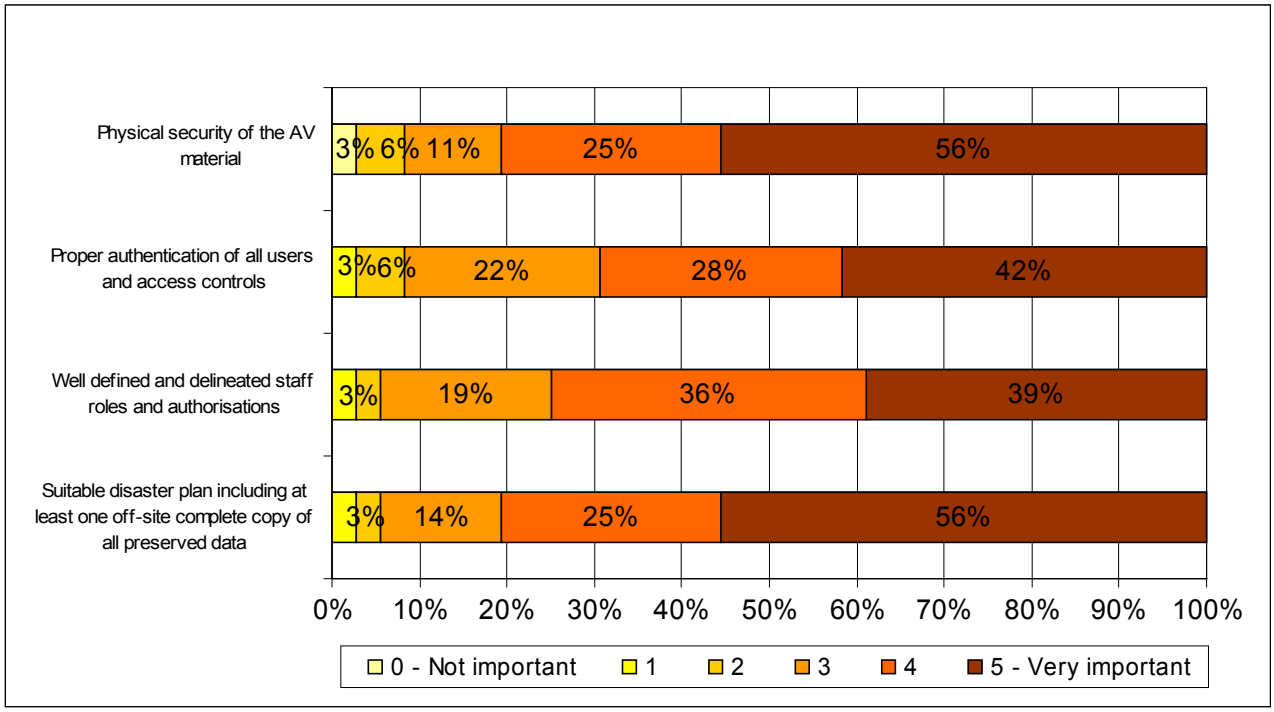


Figure 5: Rated importance of the criteria in the category “Security”

After these questions, the respondents were told that the aspects appearing in the questions above were examples of what would be checked were a TRAC audit carried out on a service. When asked, it was found that *39% were already aware of TRAC*.

We re-analysed the responses for all questions into responses given by those people who were aware of TRAC and those who weren’t. The sample size is small and the answer given by the two groups to most questions was very similar. The analysis of the former question “Do you consider out-sourcing for some or all of your archive services?” showed the most prominent difference: 50% of TRAC-aware respondents would consider outsourcing and only 32% of unaware respondents would (see Figure 6). It is hard to tell whether this is a causal relationship though: is TRAC giving people confidence in out-sourcing or is it that people considering out-sourcing have done some research and come across TRAC?

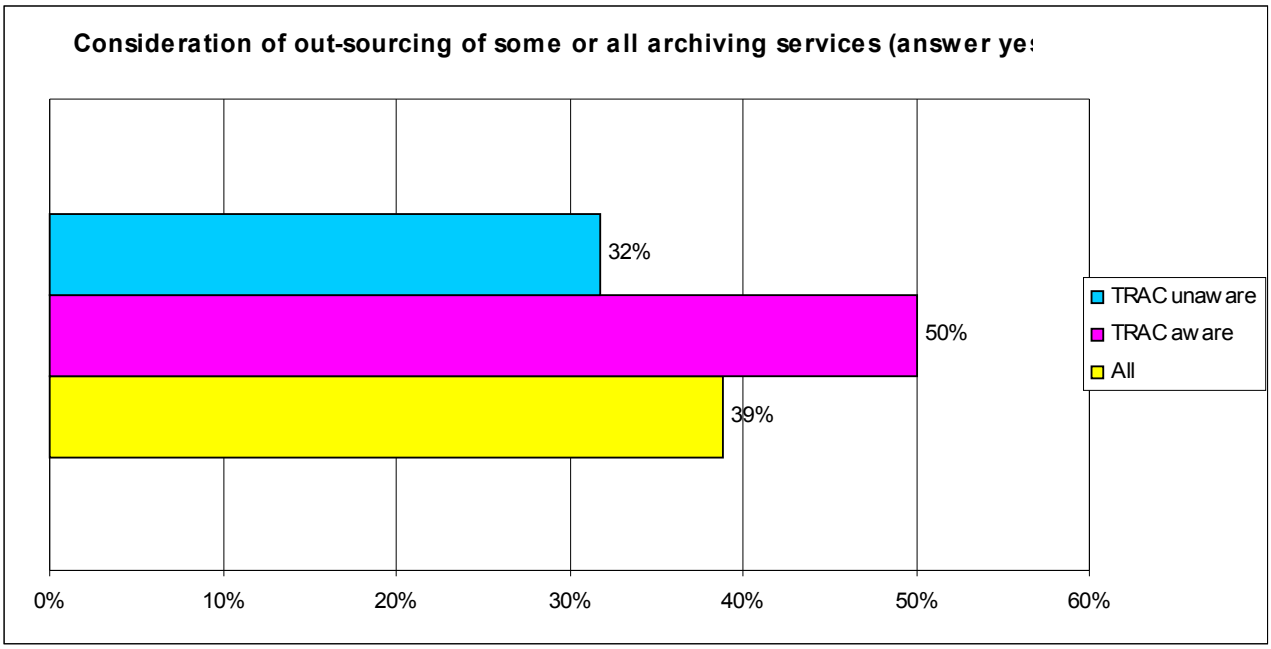


Figure 6: Consideration of out-sourcing among TRAC aware and unaware people

Moreover the respondents have been asked whether they would consider it helpful for determining the trustworthiness if a preservation service had a certificate from a competent third party auditor who followed a recognised scheme such as TRAC. The responses can be seen in Figure 7. Only 8% the respondents would completely trust the judgement of the auditor. Approximately 11% of all respondents would refuse to accept such a certificate; however, with 81%, the vast majority would regard the certificate as helpful but would want to do some checks themselves.

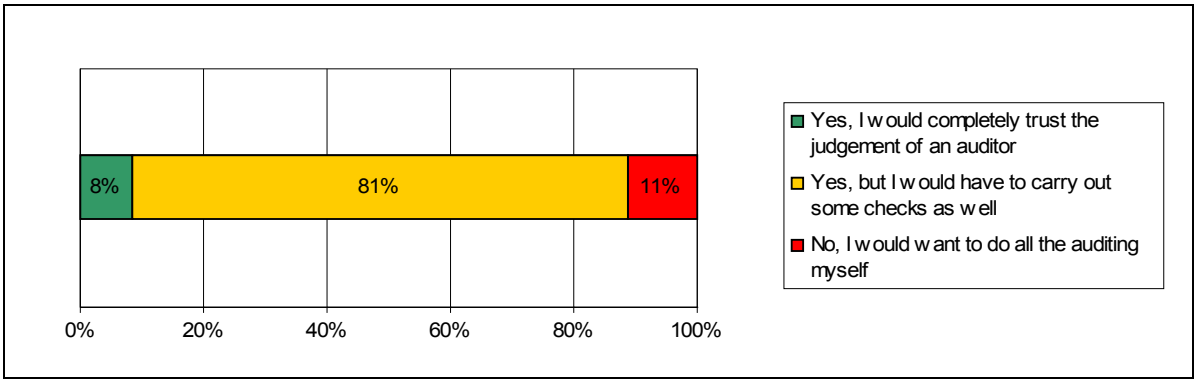


Figure 7: Would an audit certificate following for example TRAC help to trust the digital preservation service?

Beyond that, further aspects of trustworthiness that are not directly auditable in terms of a certificate have been listed in the questionnaire to be rated by the respondents. The results are shown in Figure 8. The permission to inspect the facilities, personal contact with the staff and recommendation of various sorts all appear highly rated. Marketing of the service, however, is not considered to be important.

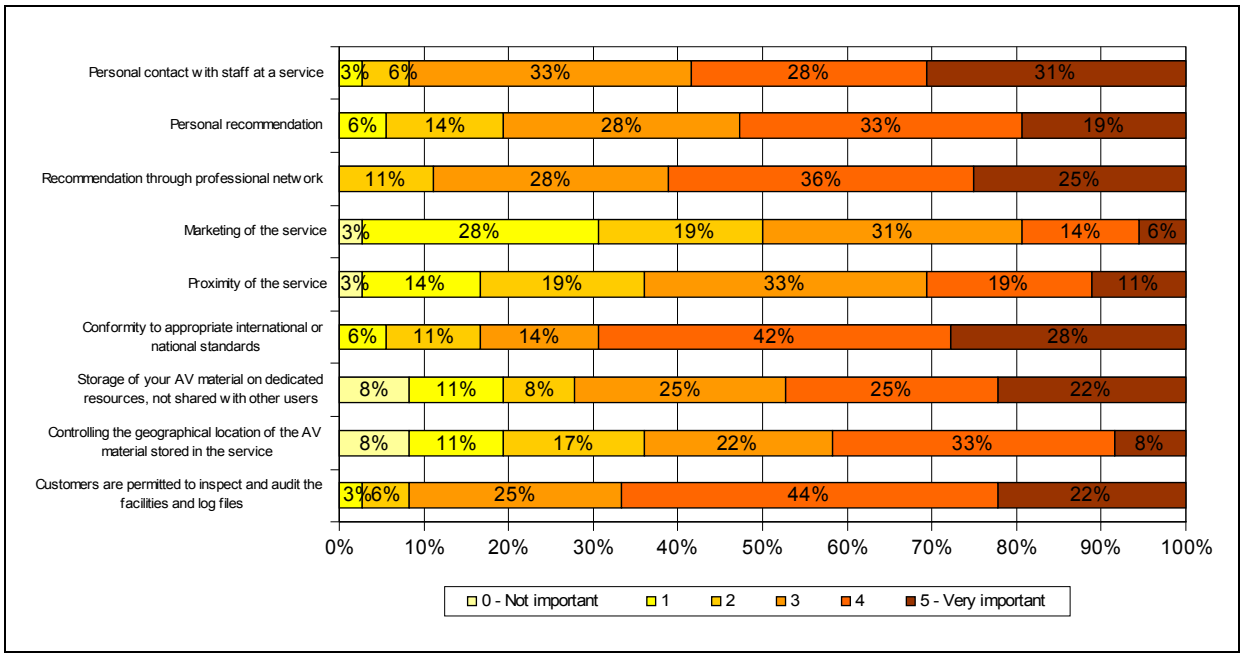


Figure 8: Rated importance of un-auditable criteria

Finally, respondents were asked if there were any other criteria they considered important when deciding of a repository was trustworthy. Some answers covered aspects that would also be checked for by TRAC and some related to SLA terms, but apart from those, answers included:

- Premises located in a minimal risk area.
- Longevity of the company itself and length of service of staff working on the project.
- Having a community around the repository to safeguard its sustainability.
- The business of the provider not conflicting with the interests of the customer.
- In-depth knowledge of OAIS models, metadata structuring; ability to handle range of A/V formats and evolution of file formats.

2.2. Conclusions

The top six criteria for trusting a service provider that are considered to be “important” or “very important” in this survey were:

With 81%:

The preservation service has a clear preservation plan (e.g. when to migrate).

The physical security of the AV material held in the preservation service.

The preservation service has a suitable disaster plan including at least one off-site complete copy of all preserved data.

With 75%:

> The staff of the preservation service have well defined and delineated roles and authorisations (e.g. to ensure that only senior staff can make critical changes to the data or system).

With 72%:

> The preservation service has the appropriate number of staff and a professional development plan.
>
> The preservation service has a succession plan (what should happen if the service ceases to exist).

The analysis of the survey results has shown that certification schemes for digital preservation services such as TRAC have already reached a reasonable awareness in the digital audiovisual preservation service provider community (39%). It has also been shown that the criteria covered by TRAC are in fact considered to be important by the community for determining the trustworthiness of a preservation service. Most (89%) of the respondents would actually take the results of such an audit into account when they evaluate a repository. The respondents gave a lack of emphasis to the importance of IPR issues in general. Finally it has been shown that other criteria, which are not necessarily part of an audit, play also an important role for the people involved.

3. Lifecycles

The lifecycles of content, rights, metadata, organisations, contracts, SLAs and service are discussed in PrestoPRIME D2.3.1. This chapter focuses on the relationships between the lifecycles of SLAs, resources, service offers and the service itself.

Figure 9 is a simplified representation of the states that a service, a service offer (or SLA template), the SLA and the underlying resource go through adapted from work done by IT Innovation in the SERSCIS project[18]. It consists of four state diagrams which start at the solid circles and finish at the outlined circles. The solid arrows show transitions between states and the dashed arrows indicate a prerequisite from a different state diagram that must be fulfilled before a state is entered. For instance, a service must be defined before resources can sensibly be allocated to it. The special case of the “service accessible” state shows that before a service can be used, the service must have been deployed and have an SLA in force.

The four state models represented here are:

- The service: its definition, resourcing, deployment and decommissioning,
- The resources: their acquisition by the service provider and allocation to the service.
- The service offer: the specification of the terms under which the service is made available to the customer.
- The customer interactions with the service: agreeing and then using an SLA.

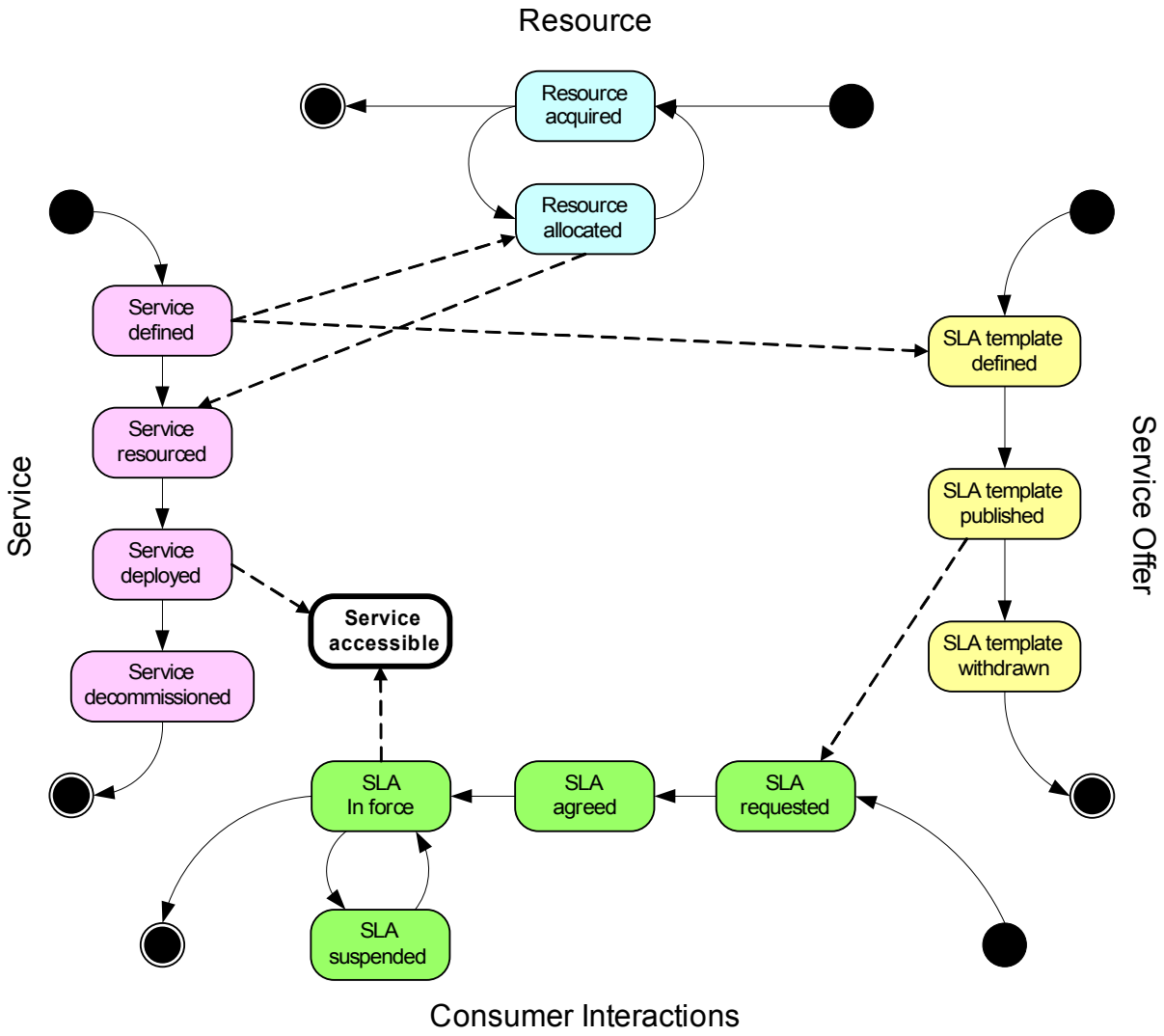


Figure 9 Interacting lifecycles of services, service offers, resources and SLAs.

The language of the service offer diagram may need some explanation. The phrase “SLA template” is used here to mean an offer being made by the service provider to the consumer. The template is first defined and then published, that is, made visible to the consumer in some way. Once the template is published, the consumer can look at the offer and if they are satisfied with it, request an SLA based on the template.

The diagram makes no attempt to show any negotiation processes but these are not excluded from this scheme. If a service provider is intending to offer a service to many customers, each one relatively small then the most economic approach is to define one or more fixed SLA templates and let the customer choose. For large (or single) contracts the possibility always exists of the service provider negotiating with the customer the precise terms that the service will be offered on.

Finally, the diagram does not attempt to take into account any of the multiplicities. For instance, a customer could have more than one SLA. There could be more than one customer per service. Many services could exist and an SLA could refer to several of them.

4. Measurement and Management

Before getting into the specifics of SLAs for online storage services, we must consider the more general question of "what can you measure and what do you want to do about it?"

4.1. SLA Design Process

To create a system of application services managed by SLAs requires understanding what you want to:

- monitor (e.g. "I'm interested in seeing how much data is transferred"),
- constrain (e.g. "a client should not be allowed to store more than 10GB"),
- promise (e.g. QoS: "the response time should never exceed 50ms") and
- charge for things (e.g. "£3 per job, 1p per CPU.second").

Knowing this tells you what *metrics* are important. A metric is loosely defined as "something measurable" such as the number of CPUs, amount of data or number of sessions, see the "Terms" section below for a little more detail.

The application service must then be instrumented somehow to provide the data necessary to perform the four functions listed above. This leads us to considering what can be measured at the application service itself. This is what the application developer should be able to tell you. It is important to use precise language when discussing these measurements. If someone says "we can measure *xyz*" they might mean *xyz* is sampled every hour or every second, they might be talking about the measurement now or the mean measurement over a 5 minute sliding window, etc. The units of the measurement are obviously crucial.

So, it is unlikely to be a simple case of working out what should be monitored, constrained, promised or charged for and then getting an application developer to report the necessary data. It is more normally an iterative process of "I need this, what can you give me?" and then "if you can tell me that then I can do this". During the process everyone has to keep in mind the question "why do I want to be measuring this?" There is no point measuring and reporting data that is not going to be used for anything.

4.2. Terms

We must be careful about the use of terminology, and especially the distinctions between metrics, measurements and constraints. This section defines specific words to have specific meaning adapted from work in the SERSCIS18 project, which was based on earlier work in the Edutain@Grid[19] project. These words are highlighted in italics and their relationship to each other is illustrated below in Figure 10.

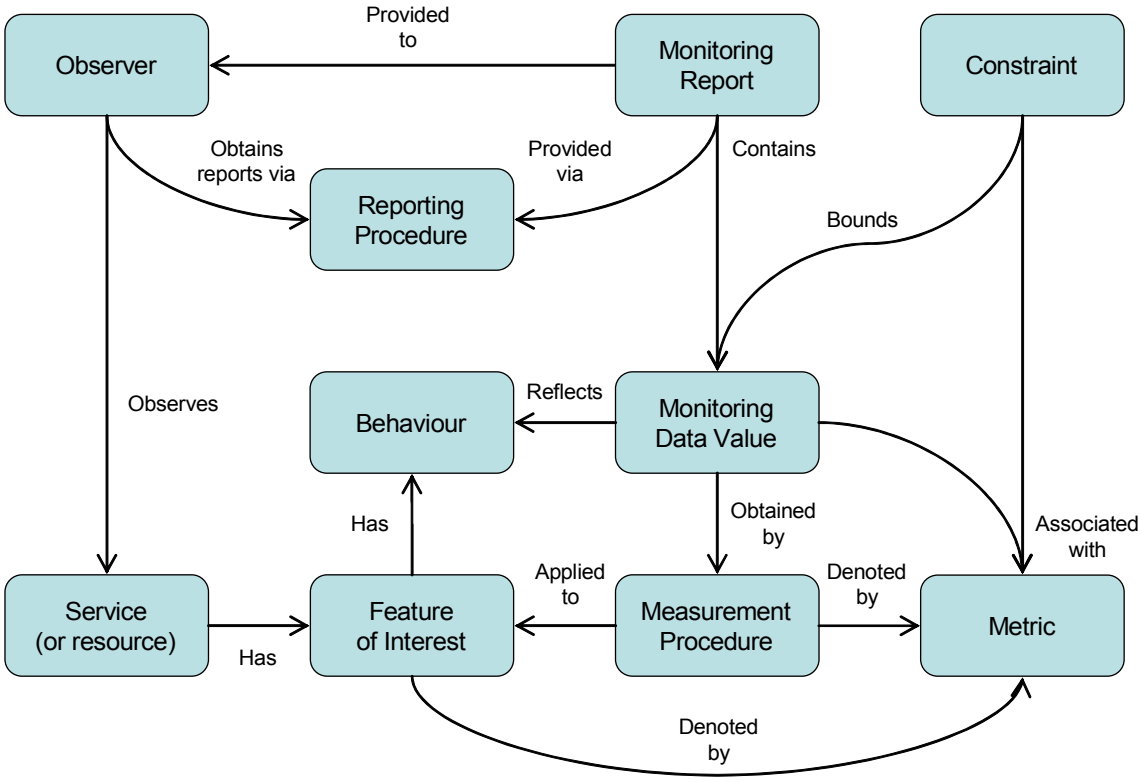


Figure 10 Metrics, measurements and constraints

Services (or sometimes the *resources* used to operate them) are monitored to provide information about some *feature of interest* associated with their operation. The *monitoring data* by some *measurement procedure* applied to the feature of interest at some time or during some time period. *Metrics* are labels associated with this data, denoting what feature of interest they refer to and (if appropriate) by which measurement procedure they were obtained. Finally, monitoring data is supplied to *observers* of the service at some time after it was measured via *monitoring reports*, which are generated and communicated to observers using a *reporting procedure*.

It is important to distinguish between monitoring data for a feature of interest, and its actual *behaviour*. In many situations, monitoring data provides only an approximation to the actual behaviour, either because the measurement procedure has limited accuracy or precision, or was only applied for specific times or time periods and so does not capture real-time changes in the feature of interest.

Constraints define bounds on the values that monitoring data should take, and also refer to metrics so it is clear to which data they pertain. Constraints are used in *management policies*, which define management actions to be taken by the service provider if the constraints are violated. They are also used in *SLA terms*, which define commitments between service providers and customers, and may specify actions to be taken if the constraints are violated. Note that management policies are not normally revealed outside the service provider, while SLA terms are communicated and agreed between the service provider and customer.

Constraints refer to the behaviour of services or resources, but of course they can only be tested by applying some *testing procedure* with an associated data model to the relevant monitoring data. The testing procedure will involve some mathematical manipulation to extract relevant aspects of the behaviour from the monitoring data. For instance, if the

monitoring data was a count of the number of recorded bit errors then an appropriate constraint might be on the number of bit errors per month. This would be obtained by the simple mathematical manipulation of subtracting the total number of errors one month ago from the current number of errors. It would not be appropriate (or useful) to integrate the number of bit errors over time.

4.3. Monitoring Information Flow and Processing

Just monitoring a metric is itself a potentially complex and subtle task. To understand what is possible and what must be defined by the SLA we must look at the flow of information from something being measured, reported, stored and queried:

1. We have a *feature of interest* represented by a *metric*, i.e. something that is being observed and measured (distinct from how it is measured/sampled etc).

2. There is some *measurement procedure* applied to the *feature of interest* which results in a *monitoring data value* which we can access in software. It can be:

    a. a “raw” value such as the amount of CPU or disc in use at time *t*, or

    b. an already processed value. For instance a network switch may only be able to tell you the error rate over the previous 5 minutes.

    Regardless of *raw* or *processed* the *monitoring data value* could potentially have an error associated with it.

3. We access this *data value* at the application service and can potentially report it directly to the *observer* (e.g. an SLA service) in a *monitoring report*, or we may want to process it at the application service first to perhaps summarise the data ourselves (e.g. if we can access the raw data every 10ms we might prefer to report the average over every 1 second instead of sending 100 separate reports).

    N.B. any processing done in (2) or optionally in (3) has an effect on the types of constraints, prices etc we can do at the *observer*.

4. The *observer* service receives usage reports and places them into a database.

5. Various mathematical manipulations can then be used on the data in order to test constraints, including things like computing the average, maximum value and cumulative value. Which manipulations can be used depend on the sort of data that went in. For instance if we were able to receive and store a histogram of packet processing time for each second then we could answer questions about "what proportion of packets took longer than 20ms to process between t_1 and t_2?" but we could not say with precision what the maximum value was - only a range.

It is important to understand precisely what can be or is being measured at an application service. For instance, in the case of measuring CPU usage, it is generally possible to find out from the operating system what the current *instantaneous* usage is for a process and also what the *cumulative* usage is from the start of the process, e.g. “process *xyz* is using 98% CPU at this moment in time and has so far used 987 CPU.seconds.”

Once the usage reports are in the SLA service we can use mathematical models to infer additional information. For instance, if we have reports of instantaneous bandwidth

measurements (the black dots in Figure 11) then we can approximate the total data transfer by assuming that the transfer rate remains constant between reports and performing the numerical integration represented by the hatched area in the figure. (Of course, there are easier ways to find the total data transferred, this is intended to demonstrate the concept of inferring one value of a metric from others.)

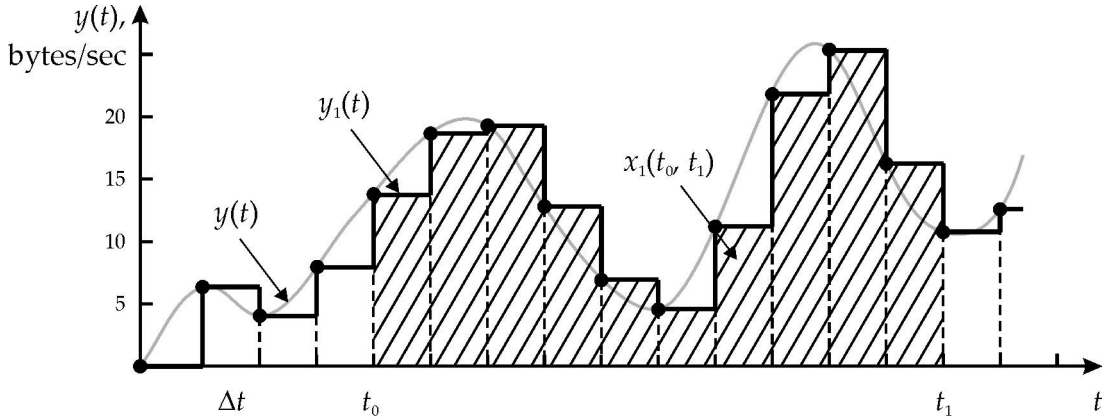


Figure 11 We can approximate the area under the curve by sampling the rate (black dots) and performing numerical integration (the hatched area).

4.4. Reporting and Management Mechanisms

First of all, what types of *measurement procedures* are there?

- analysis of log files, either periodically or in response to some event,
- querying of a software system (such as the OS), either periodically or in response to some event,
- measurement integrated directly into a software or hardware system.

For instance, if we wanted to measure the number of users logged in to a service we could:

1. have the service record login and logout events to a log file and (in a separate component) periodically read it in and analyse it,
2. have the service expose an interface to permit a query for the number of users of the service directly and use this interface from another component,
3. have the service respond to login and logout events itself by generating a *monitoring report* directly.

Option (1) is often the only viable method if the application being monitored cannot be modified. Option (3) is the tightest integration and ideal for a new application written specifically for the monitoring system. Option (2) holds the middle ground. Options (1) and (2) both require some other component to generate *monitoring reports* whereas in option (3) the reports are generated directly.

Once a report is generated, how is it transported to the system that takes care of testing the constraints and implementing management policies: the SLA manager? There are two main options:

1. the SLA manager periodically pulls reports from managed services,

2. the managed services periodically push reports to the SLA manager.

The disadvantage of (2) compared to (1) is that in (2) if there is a communications problem between the managed service(s) and the SLA manager (either a network problem or if the SLA manager is not running) then the managed service(s) have to deal with the SLA manager not being available and being unable to deliver their reports. In a poorly written system this could lead to the managed services failing. In addition, retrieving reports through a pull mechanism means that it is possible for a user to access reports even when separated from the service by a firewall (as is usually the case).

Which architecture is better also depends upon the summarisation strategy for the data. For instance, using the same simple example as before, if a managed service was reporting login and logout events once per hour then there is the choice of it reporting either a list of every event in the hour with its associated time or a summary saying e.g. “Under SLA *xyz* there were 5 login and 4 logout events”. Where to summarise the data in the architecture depends on what each component needs to know which is in part determined by the required management actions.

If the managed service is not summarising the data then there is an additional problem of the queue of reports needing to get to the SLA manager becoming very large and taking a long time to process.

IT Innovation has been considering issues of SLAs for many years now and added an SLA management system to their grid middleware stack, GRIA[20], in 2006. The interactions between the SLA Manager and the Application Service in GRIA 5 are shown in Figure 12. This is not the only possible architecture but it serves to illustrate the concepts discussed. The GRIA 5 architecture used a pull model and the managed services did not summarise the usage reports.

The GRIA 5 SLA Service has two distinct opportunities to manage an application service:

1. An application service can make explicit requests to start or continue an activity that will use significant resources. At this point, the SLA Service can refuse permission.

2. An application service can report how much of each metric has been used and what the rate of usage is *now*. This reporting is currently done asynchronously via a pull-point. Periodically, the SLA Service processes these reports, predicts current usage and compares usage to the SLAs. It may find that a management action is necessary and contact the application service to manage the activity.

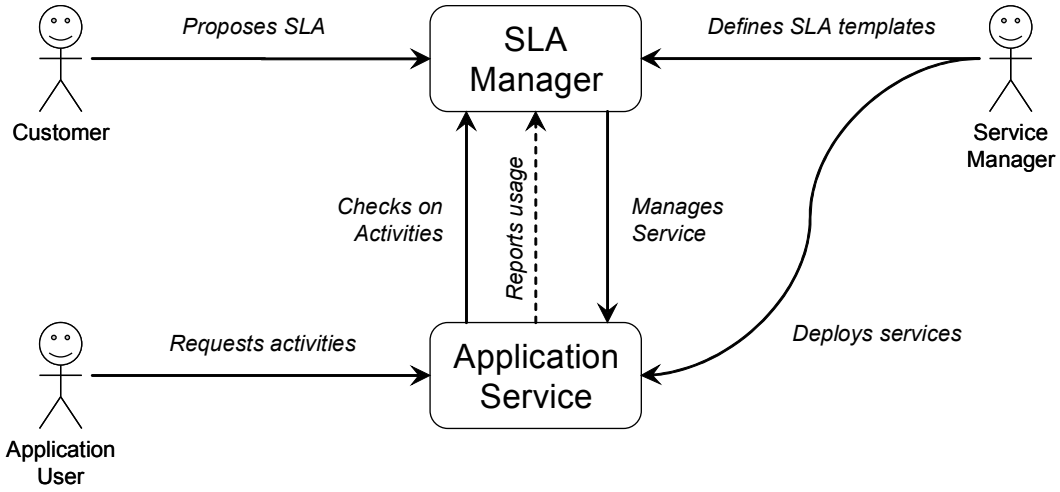


Figure 12. Interactions between the GRIA SLA manager and an application service

An example of (1) is if a user wanted to upload a file. The application service might ask the SLA Service "can the user increase their disc space usage by 10GB?" and the SLA Service would check if doing so violated any constraints (e.g. a disc quota) and accept or deny the request.

An example of (2) is if a user had a long-running computation. As the job progresses the SLA service is kept updated with how much CPU time has been used and what the current rate of usage is. At some point, the SLA Service may find that the user has reached their CPU time limit and will tell the job service to pause the job.

4.5. Management Actions

Or: what do you want to do about it?

There is no point measuring and reporting something that you cannot or do not want to do anything about.

There are a variety of clauses that might go into an SLA:

- Constraining the user of the service:
  - What can they do?
  - When can they do it?
  - How much of a resource can they use?
  - What happens if the client tries to use more than permitted?
- Constraining the provider of the service:
  - What resources must they provide and when?
  - What quality of service must they provide?
  - What happens if the resource or quality is not delivered?

- Concerning prices:
    - What is the tariff for the use of the service?
        - Flat rate?
        - Varying with usage?
        - Paid up-front or in arrears?
    - Are there penalties payable by either party?

Quality of service (QoS) guarantees/promises are often couched in terms of percentages, for instance the response time of a query might be guaranteed to be below 20ms 95% of the time and 20-25ms 5% of the time. For anything outside this range a penalty might be payable.

The definition of QoS must be done carefully (or perhaps care must be taken by the consumer). A promise of 99.9% uptime for a service sounds good but could mean the service is unavailable for 86.4 seconds every day at noon which may be exactly when you are using it.

What you can actually do obviously depends on your application service. For instance, a customer may require a guarantee about the latency of a connection. Often the network cannot adapt so the service provider must offer relatively low latency guarantees and pay a penalty if they do not meet the threshold. With specialised hardware the network can be controlled precisely so a latency guarantee can be met by issuing management instructions to the network hardware to improve a link.

There is an open question about where a management action should be defined. It could be argued that the SLA service, upon detecting a constraint breach should be able to determine exactly what actions should be taken, including instructing an application service to control a resource in some way. This implies that the SLA service must know in detail what the application service is capable of and how different management actions will affect the resource usage. An alternative is to have the SLA service to merely tell an application service that action must be taken to adjust some metric and leave it for the application service to determine the best course of action. There is an issue with getting the application service to choose the management action: it means that the SLA Service does not need to have application-specific management actions which reduces coupling, but it also means that the SLA Service does not have control over what management action is taken.

4.6. Templates

This section tries to pull together everything that needs to be considered when capturing information about measurements, QoS, constraints and pricing into four sets of questions. Once captured, this information can be used to consider for example whether, given a desired constraint, whether it is possible to obtain the necessary information from the monitoring.

Knowledge about the measurement process is important. For instance, if a service provider wanted to guarantee that the bandwidth of a connection would not generally drop

below 1 MB/s then we need to know how the bandwidth is measured. The customer will not be satisfied if the bandwidth is only sampled once every hour as the chances that the measurement would miss the bad spots are very high. If the bandwidth is sampled every millisecond then the customer would be satisfied but sending all that data to the SLA service would probably overwhelm it.

Here is a set of questions to answer about a measurement made by an application service:

Measurement <M-ID>:

- What is the feature of interest?
- Name of metric:
- How is it measured?
- How often is the measurement made?
- Is it a raw measurement or pre-processed?
- What are the units of the measurement?
- Is there any uncertainty in the measurement?

Coming from the business perspective rather than the technical perspective, we need to consider what QoS guarantees are required. QoS guarantees might refer directly to metric measurements such as “the measured response time must not exceed 5s” but may also potentially refer to functions of metrics or time periods such as “the average response time over each calendar day must not exceed 5s”. A QoS guarantee might also use terms that are better understood by the user rather than technical resource-level terms, so for instance it might talk about “*x* hours of SD video” rather than “*x* bytes of information”.

QoS guarantee <Q-ID>:

- Description:
- Relating to metric(s):
- Units:
- Is this a hard guarantee or something that should be met a certain percentage of the time, in which case, how much?
- How can conformance with this QoS guarantee be judged?
- What is this QoS guarantee conditional on? (e.g. any particular user behaviour?)
- What happens if the QoS guarantee is not met?
- What automatic management actions can be taken to ensure the QoS is maintained?

To have a manageable service certain limits must be set upon the users’ behaviour. Limits may apply to instantaneous measurements such as “never let the user store more than *x* GB of data” or might apply over time periods such as “each calendar day the user may upload *x* GB data” or “over a 1 hour moving window the user cannot access more than *x*

files". They might not apply directly to the measured metrics but may constrain functions of them such as "the sum of data uploaded and downloaded must not exceed *x* GB".

Constraint <C-ID>:

- Description:
- Relating to metric(s):
- Units:
- Is this a hard constraint? That is, something that the system should always ensure is met?
- If the constraint is breached, what should happen?

You might want to charge the user a variable rate depending on their usage of the service or perhaps a flat rate. If the charge is dependent on the usage then the charges must be calculable from the measured metrics.

Pricing term <Prc-ID>:

- Description:
- Relating to metric(s):
- Formula in words:

It may be that penalty terms are triggered if, for instance, a QoS term is not met. They are very similar to pricing terms but the flow of credit goes the other way.

Penalty term <Pty-ID>:

- Description:
- Relating to metric(s):

5. SLA Terms for Storage and Preservation

5.1. Introduction

With business SLAs we refer to the agreements made between a provider and a customer with a specific business point of view in mind and with respect of a given service. For a preservation service, we try here to take the role both of the customer and the provider and conjecture at business level which is the minimal set of features and performances that this sort of service should have and respect.

SLAs have very practical effects because they formalise on one side what the customer expects as features and behaviours and on the other side what the provider promises to its customers. SLAs have direct impact on the quality of service guarantees and last but not least determine the operating costs.

Under this collection of business issues exists the very practical need to impartially measure, monitor and evaluated a set of quantities. We know that measures outcomes are by definition approximate, moreover they can be derived, mediated over different time intervals, quantified with different measure units and bounded in different ways to assorted constraints.

In the light of these considerations, it is clear that a sharp division between business and technical SLAs is not possible. Even at high level, there is the necessity to state for example that the submission of a package of a certain size should not take more that a given time, thus is necessary to quantify, even if with measure units closer to a business way of thinking. The amount of multimedia could be measured in term of hours of play instead of gigabytes, the speed of transfer with hours per package submitted without mentioning bandwidth etc. In the following chapter we try to give a business view thus considering high level metrics we want to monitor to have a satisfying service.

5.2. Amazon Web Services

For comparison with the ideas presented for preservation SLAs below, here we include information about the SLAs provided by Amazon: the well-known storage (S3) and compute (EC2) web services.

The Amazon storage service, S3, has a simple SLA stating that Amazon commits itself to a “monthly uptime percentage” of 99.9% and that if this is not met then users will receive a refund of a proportion of what they pay.

Users of S3 pay:

- Storage: $0.150 per GB-month of storage used (for first 50 TB of data, $0.055 per GB-month for data over 5000 TB).
- Data Transfer: data transfer in is currently free, data transfer out ranges from $0.170 per GB for the first 10 TB per month to $0.100 per GB when going over 150 TB per month.
- Requests: $0.01 per 1,000 PUT, COPY, POST or LIST requests and $0.01 per 10,000 GET and all other requests (delete free).

This all implies metrics of uptime, total data stored, data transfer and requests.

The Amazon compute service, EC2, uses the same SLA term regarding uptime. The pricing terms are more complex than for storage as it depends on the virtual machine type, operating system and reservation technique (on demand, reserved or spot prices). For a Linux machine, on demand, the following prices apply:

Standard	***Price per hour***
Small	$0.085
Large	$0.34
Extra Large	$0.68
High Memory	
Double Extra Large	$1.20
Quadruple Extra Large	$2.40
High-CPU	
Medium	$0.17
Extra Large	$0.68

Table 1Selected Amazon EC2 prices

Finally, these Amazon services are now complimented by Amazon Cloud Watch and Auto Scaling with Elastic Load Balancing. Cloud Watch monitors Amazon EC2 instances and the Auto Scaling uses the monitoring data to trigger additional deployments with the load automatically being balanced amongst the instances.

Amazon Cloud Watch uses a small set of metrics for monitoring:

- CPU utilisation (percentage)
- Network traffic in and out (both in bytes)
- Disc read and write operations (both just a count)
- Disc read and write volumes (both in bytes)

These metrics are reported each minute.

5.3. Preservation Services

A typical preservation service is quite complex and includes a wide set of possible operations both generated from the provider (e.g. ingest and access) and from internal management subsystems that take care of the preservation of data/information (e.g. migration processes). Because of this complexity, for the sake of clarity and better

implementation a well-structured view is needed, where each function takes place in a specific sub-system. In the following sub-chapters we describe the major aspects of SLAs, according to sections 4.1 and 4.2

Capabilities

Below is a list of features or capabilities that the service should or could provide. Some of them are at the core of a preservation system and are stated to be mandatory, others could be desirable for some customer but not essential (e.g. advanced search by similarity). In Table 2 are shown some of the features deemed more interesting; for the complete list please refer to the appendix.

ID	Name	Requisite level	Description
CAP-01	Ingestion/Delivery	Mandatory	Capability to upload/download material and related metadata in the form of an agreed SIP, more protocols (e.g. ftp, http) and modalities (e.g. push, pull) should be provided
CAP-02	Validation	Mandatory	Identification and formal check of formats including wrappers, AV essence and metadata formats.
CAP-03	Partial extraction	Mandatory	It is essential to be able to ask for a specific portion of a stored object e.g. from frame 5000 to frame 51500 of a certain programme instance
CAP-07	Rights management	Mandatory	This is a complex aspect to be better investigated, it is only partly related with Access control. As a minimum feature the system has to be able to understand (validate) rights management metadata and forward them as part of the delivery package to the customer.
CAP-09	Package update	Highly Recommended	Capability to accept revisions of metadata, but eventually also some change regarding part of AV essence
CAP-10	Transcode /Transwrap	Highly Recommended	The capability to transcode between different audio-video formats and different containers (the same could be for metadata formats)
CAP-11	Migration	Highly Recommended	The capability to manage massive batch processes of migration from some audio-video (and eventually metadata) formats and wrappers to others in order to preserve the usability of the content. Migration should be mandatory if there does not exist an alternative mechanism like multivalent.
CAP-13	Demux	Recommended	The possibility to ask and obtain only a specific track of audio / video to be extracted from the multiplex
CAP-14	Upload/Download Resume	Recommended	The possibility to resume the download/upload operations when the process has been interrupted because of network or application troubles; the process could also have been paused by the customer
CAP-15	Search/Retrieve by meta	Recommended	Possibility to ask for a specific stored objects by specifying several kind of metadata

Table 2 Selected capabilities

Features of interest

Starting from capabilities, a set of related “features of interest” has been developed. They are the physical quantities we assume a customer would like to monitor and control, such as the elapsed time for a SIP ingestion. Also the provider has to monitor these quantities in

order to feel the pulse of its systems and possibly react in case of problems but also for providing evidence about the quality of service delivered.

No doubt the provider would also monitor other things at deeper technical level, for efficiency and system improvement, but such monitoring is not of interest to the customers and would not fall into the realm of SLAs.

First we tried to assume the role of a customer (e.g. a broadcaster who wants to preserve its archive material) and wondered what we would get. A set of requirements and questions arose like:

- I want the content to be safely preserved against corruption and obsolescence
- I would like the system to be “always” usable when needed
- How long doest it take to have the material I need?
- How precise is the search?
- I would like to have all the services like search, preview etc. sufficiently fast and precise.
- I would like to be able to manage rights.
- …

We organised and worked on these inputs obtaining a list of features to be monitored, from which emerges the principal aspects to be considered: the preservation of data/information and the performances in terms of elapsed time and precision of the basic operations (ingestion, search, access). Table 3 is an extraction from the full list reported in the appendix. The behaviour column tells something about the way this feature typically changes over time while the referenced metrics points to the associated measures.

ID	Name	Description	Behaviour	Ref. metric
F-01	Availability of services	A Boolean function of time (available: true or false?) representing the fact that the service is up and effectively usable. This feature should be detailed for each atomic service making up the entire preservation service, covering at least: - SIP ingestion (and update) service - Search service - Browse service (if present) - DIP delivery service	For many and varied reasons the system is unavailable; the time-course of the availability is not predictable but the customer can expect an account of overall avaibility (per hour, day, week or whatever is agreed), so the availability needs to be logged (again, as a time function).	ME-01
F-02	Content Information corruption	A function of time that represents the unrecoverable corruption level of the information package (ingested A/V and metadata contents). In the case that the system has been asked to keep original files the corruption check is made directly on bits and not on A/V quality.	Typically It can only increase, because of bit rot or data loss, affecting both essence data and metadata. Not all the bytes have the same importance e.g. the representation information is particularly important because in case of lost it could be impossible to exploit the content. In case of corruption it is actually possible in some cases to restore at least partially the audiovisual content	ME-02, ME-03

ID	Name	Description	Behaviour	Ref. metric
			using complex algorithms	
F-03	Storage occupation	The actual storage space occupied by the customer (storage allocation - actual remaining space)	It depends on the number and weight of ingested and eventually deleted SIPs	ME-04
F-04	SIP Ingestion time	The total elapsed time from the SIP submission to the confirmation from the system that everything has been correctly acquired. This includes: - the time necessary for the upload transfer of the package - the time necessary validate the SIP The ingestion can also be an update of a pre-existing object.	This time could be affected by the total system workload (e.g. the network could be congested), it is reasonable that the ingestion time is proportional to the dimension of the package.	ME-05
F-06	DIP Delivery time	The total elapsed time from the request of a specific SIP (e.g. discovered by a previous search) to the complete and correct reception of the package. This time includes: - the time necessary to extract and prepare materials with a coherent DIP wrapper - if necessary, the time for recovering a corrupted file - the time necessary for the download transfer	The time necessary for preparing the material is highly variable depending on required operations like transcode/transwrap, aggregation of separate content etc. It could be affected by the total system and network workload.	ME-07
F-07	Search time	The elapsed time from the query submission to the production of the result list	This time could be affected by the total system workload (e.g. multiple concurrent queries), also depends on the complexity of the query, the total amount of indexed material as well as on the actual implementation of the indexing engine.	ME-08

Table 3 Selected features of interest

Metrics

For each feature of interest, a metric is given with:

- its unit of measure;
- an indication of how to calculate or derive it;
- whether the value has to be mediated over a certain time slot defined in a certain way (e.g. calendar date, hour or sliding window with a certain width).

Table 4 is an extraction from the full list reported in the appendix and it reports as meaningful examples the metrics related to *features of interest* (Table 3).

ID	Name	Description	Unit of measure	How to calculate monitoring data value from measures
ME-01	Availability over time	Given a certain time slot that can be fixed (e.g. a specific month or year) or sliding windows (e.g. the last hour) is the percentage of time where the service was available (up and working correctly) over the total time. The assumed time slot and modalities are explicitly agreed between the parties.	Percentage (calculated over a certain time slot)	On a practical point of view measures are made with a sampling approach, e.g. by testing the availability every minute. The monitoring data values are thus simply calculated for each time slot making the ratio between successful and the total number of calls. Some basic measurements can be done without ad hoc tests but only considering customer calls
ME-02	Bit Integrity	Expresses the fact that the information originally submitted by the customer has been preserved from corruption. One way to state this is to count the number of corrupted bytes over the total amount of storage occupied and the total time of retention.	Incidence of byte corruption / GB*year (e.g. 10^ -3 byte/GB*year)	Calculated on fixed periods (e.g. each month or year), on this period sum the number of mistaken and corrupted bytes and divide this number by the mean storage occupation during the period. Could also be decided to calculate over a sliding window e.g. a window of one month wide calculated each day.
ME-03	File Integrity	Another way to state the integrity is to express a percentage of damaged files over the total amount of retained files, in a specific lapse of time.	Incidence of file corruption / numfiles*year (e.g. 10^-4 per year)	Calculate on fixed periods (e.g. each month or year), on this period sum the number of corrupted files and divide this number by the mean number of files during the period.
ME-04	Storage occupation	The ratio between the actually occupied storage and the total amount reserved for that customer	percentage	Absolute storage occupation can be queried directly on the system or a running total could be kept based on ingestions and deletions. A storage occupation history could be saved in order to calculate interesting QoS parameters like the incidence of corrupted bytes along the timeline
ME-05	SIP ingestion time	The time necessary for ingesting the submission package, usually the time is normalised against the size of the package because most of the time factors go linearly with the size (copy, transcode, transfer)	Hours or Hours / GB	It has to be measured directly, it can be probably obtained inspecting activity logs
ME-07	DIP delivery time	The time necessary for delivering the dissemination package, usually the time is normalised against the size of the package because most of the time factors go linearly with the size (copy, transcode, transfer)	Hours or Hours / GB	It has to be measured directly, it can be probably obtained inspecting activity logs
ME-08	Search time	The elapsed time from the submission of the query to the complete answer of the system search engine.	seconds	It has to be measured directly; it can be probably obtained by inspecting activity logs. If promises are given as statistical percentages, statistics have to be calculated.

Table 4 Selected metrics

Quality of Service

The quality of service expresses how the service is effective and fruitful. The promise about a minimum quality is what we are concerned with. The providers in fact guarantee an acceptable level of service to its customers.

Promises are typically expressed as bounding metrics and form one of the most important aspects of SLAs. For example an SLA clause could state that the probability of a file becoming lost or corrupted after a period of a year of retention has to be under 0.01%. The following table represents an extract of the global one reported in the appendix.

ID	Name	Description	Ref. metrics	Monitoring criterion (bounds)
QS-01	Availability	The guarantee that the service will be available (up and exploitable) at least as much as agreed. The measure of this quantity can be done as percentage of time (e.g. 99% of the time) or percentage on the number of usage (e.g. 98% of the times one tries, the service is usable)	ME-01	The availability should never go below a specific threshold, there could be more than a threshold (e.g. for business hours and night)
QS-02	Integrity	The guarantee that the ingested A/V and metadata contents has been preserved keeping an agreed quality level (assessed for example with PSNR). These probabilities have to be normalised over the amount of data and the retention time.	ME-02 ME-03	The integrity should never go below a specific threshold
QS-03	SIP Ingestion time	One of the most important parameter perceived by a user when submitting a new SIP (or even for updating) is the total elapsed time from the SIP submission to the confirmation from the system that everything has been correctly acquired. This includes: - the time necessary for the upload transfer of the package - the time necessary to extract, validate, index and transform the SIP into an internal representation (AIP)	ME-05	The SIP ingestion time should never go above a specific threshold, there could be more than one threshold (e.g. for business hours and night).It can be given as percentage, e.g. 90% of deliveries are done under a threshold 1 and the rest under threshold 2.
QS-05	DIP Delivery time	One of the most important parameter perceived by a user when asking for some material (media + metadata packaged in a DIP) is the total elapsed time from the request to the complete and correct reception of the package. This time includes: - the time necessary to extract and prepare materials with a coherent DIP wrapper - if necessary, the time for recovering a corrupted file - the time necessary for the download transfer The time necessary for preparing the material is highly variable depending on	ME-07	The delivery time should never go above a specific threshold, there could be more than one threshold (e.g. for business hours and night). It can be given as percentage, e.g. 90% of deliveries are done under a threshold 1 and the rest under threshold 2. Because the delivery time depends not only on dimensions but also on the kind of operations required, threshold should be given taking this into account

		required operations like transcode/transwrap, aggregation of separate content etc.		
QS-06	Search time	Search results should be obtained in a reasonable amount of time depending on the complexity of the query and on the amount of the indexed material	ME-08	One of these ways: - The search time should never go below a specific threshold - In terms of percentage of invocations and different thresholds (e.g. 90% of times time is under threshold1, the remaining under threshold2)

Table 5 Selected QoS promises

Constraints

Constraints are quite similar to QoS promises but are in the opposite direction i.e. they are promises of the customer towards the provider.

For example the customer accepts to submit only conformant packages, in case of infringement there should be a penalty. The list of constraints is quite short and is therefore included entirely in Table 6 but also reported in appendix for convenience.

ID	Name	Description	Measure unit	Ref. metrics	Monitoring criterion (bounds)
C-01	Authorised formats (wrapper, essence and metadata)	Provider and customer agree on a set of predefined formats accepted for essence and metadata e.g. MXFD10 and DV50 for essence and Mets as xml for metadata. Adhering to OAIS model also a proper definition of SIP and its validation has to be established.	N/A	N/A	A package that is not in the form of an authorised format; should be discovered by the validation phase that has to be applied for each complete or partial ingestion.
C-02	Maximum amount of storage	The maximum amount allowed to a specific customer by contract	GBytes	ME-04	The occupied storage should never cross a specific threshold, partial exceeding for a limited period of time could be tolerated
C-03	Maximum number of simultaneous users	Maximum number of users logged in at the same time	positive integer	ME-14	The actual number of logged users should never cross a specific threshold, partial exceeding for a limited period of time could be tolerated
C-04	Maximum number of simultaneous operations	Maximum number of simultaneous operations, e.g. No more than 3 simultaneous ingestions or 2 simultaneous transcoding	positive integer	ME-15	The actual number of simultaneous operations should never cross a specific threshold, partial exceeding for a limited period of time could be tolerated

Table 6 Constraints

Pricing terms

In order to establish the price of a complex service, different business models can be used. Quite surely there is a base cost that depends on agreed capabilities and magnitude of the service (e.g. it can transcode on demand and support up to 10 simultaneous users) and some variable components that depend on the effective usage of the service itself.

Some reasonable assumptions include operating costs related to the amount of handling of material (i.e. cost per GB fetched or ingested) and penalty clauses in case of infringement of the agreed level of quality.

In Table 7 there is a full description of the several cost components we suggest are reasonable.

ID	Name	Description	Units	References
PRC-01	Fixed price	This is the fixed component of price, it is determined by service characteristics, promised quality, constraints, variable cost model and market competition. Usually is in form of annual fee. Example: The service provides a storage capacity of 500 TB for a period of 5 years with the possibility to vary this amount ongoing. The maximum number of users is 50 (30 simultaneous). The service is 24h and Availability of each service (ingest, search etc.) is guaranteed for 99,9 % of time, integrity is assured with a byte corruption probability of 5*10-3 bytes/GB*year. Additional capabilities include upload-downolad resume, transcoding, demux.	Euros / year	All constraints All QoS guarantees
PRC-02	User charge	Usually included in fixed price with a maximum number of licenced and contemporary users. Some cost model can foresee a variable charge depending on number of licenced users or their connection time (login logout)	Euros / user Euros / hour	ME-14
PRC-03	Hit charge	Assign a charge for each specific service invocation e.g. 0.2 euro each ingestion, 0.5 euro for each export	Euros / service usage	
PRC-04	Data movement charge	A charge for uploaded and dowloaded quantities (e.g. 0.1 euro for each uploaded GB)	Euros / GB	ME-04
PRC-05	Storage usage	Even if in the fixed cost takes into account the maximum storage made available, the cost model could also consider the effective storage usage	Euros / GB for month of storage usage	ME-04
PRC-06	CPU usage	Some kind of operations like transcoding (e.g. for migration) are CPU intensive and it could make sense to have a variable price component based on it	Euros / (CPU * month)	ME-15

Table 7 Pricing terms

Penalty terms

Penalty terms are part of the pricing model. They are represented by the refunds given by the provider to the customer when an SLA is infringed in some way. The picture shown in Figure 13 summarises several concepts expressed in this document, including the

necessity to measure and monitor the feature of interest for taking into account possible penalties.

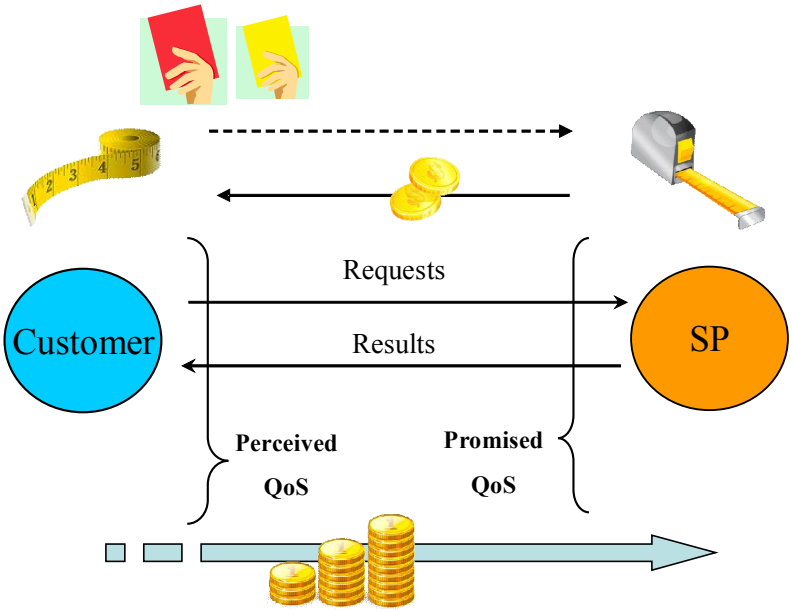


Figure 13 Abstract view of the pricing, monitoring and measurement processes related to quality of service.

In Table 8 is the proposed list of suitable penalties.

ID	Name	Description	References
PTY-01	Broken contract	The customer decide to stop the relationship with the given provider prior of the natural expire date of the contract	
PTY-02	Lack of availability	The provider does not respect the promise with respect of availability and gives a refund for this reason	ME-01
PTY-03	Data corruption	The provider does not respect the promise with respect of integrity guarantee and gives a refund for this reason	ME-02
PTY-04	Lack of band	The provider does not respect the promise with respect of upload and/or download performances	ME-05, ME-07, ME-11
PTY-05	Poor Search	The provider does not respect the promise with respect of search performances or recall/precision	ME-08, ME-09
PTY-06	Maximum storage exceeded	The customer continues to upload material even if the maximum agreed space has already been used	ME-04
PTY-07	Invalid SIP submitted	The submission package has some problems, e.g. some XML not well formed or invalid or unsupported or corrupted formats for essence	ME-10

Table 8 Penalty terms

6. Capacity Management

A business providing a service to a customer or customers needs to have sufficient resources to meet the customer requirements. More specifically, in the case of a business providing preservation services, sufficient hardware needs to be in place to meet the SLAs agreed with their customer(s). Knowledge of future trends is also necessary to correctly judge what additional hardware to purchase. This chapter looks at the relationship between the SLAs and the hardware requirement and various techniques for managing the capacity of the service.

The sorts of questions that must be addressed to successfully manage a service are:

- What resources do I need now and in the future to meet the terms of a proposed SLA?
    - How much will it cost me?
- What effect will a change in software or hardware have on the system performance?
    - Will the existing SLAs still be sufficiently resourced?

We can go through the example quality of service and constraint terms proposed in Chapter 5 to see how they have an impact on the underlying hardware requirements:

ID	Name	Discussion	Impact on
QS-01	Availability	A high availability for servers (e.g. for ingest or access) implies reliable hardware and/or redundant systems with automatic fallover. The servers must also have reliable access to the data so duplication of the data storage is also implied.	+ external servers + networking + storage
QS-02	Integrity	An integrity guarantee implies duplication of data storage in some way and processes in place to check the data integrity and repair where necessary which implies additional CPU and storage bandwidth to deal with these processes.	+ storage + CPU + networking
QS-03	SIP ingestion time	To rapidly process a SIP (extract, validate, index and transform) requires sufficient external bandwidth, fast servers and sufficient servers to ingest SIPs in parallel	+ external servers + CPU + external networking + internal networking + tier 1 storage
QS-05	DIP delivery time	Delivering a DIP requires retrieving the AIP and packaging it as a DIP. Sufficient bandwidth is also required to the customer.	+ external servers + CPU + external networking + internal networking + speed of storage
QS-06	Search time	The search time depends on the speed and capacity of the metadata database and the external search server(s)	+ external servers + database servers
C-01	Authorised formats	The constraint on which formats are authorised has no direct effect on the hardware. It is accounted for in the validation stage of QS-03	See QS-03
C-02	Maximum	This has a direct and obvious effect on the size	+ storage capacity

ID	Name	Discussion	Impact on
	amount of storage	of the storage system.	
C-04	Maximum number of simultaneous operations	More operations means more ingest, delivery or searching and is therefore related to QS-03, QS-05 and QS-06.	See QS-03, QS-05 and QS-06.

Table 9 Impact on hardware of various QoS and constraint terms

The resources referred to in Table 9 are not the whole list that must be considered. Servers, networking and storage systems of all sorts (spinning disc, tape robot, discs or tapes on shelves, etc) all need physical space (e.g. in a server rack, on a shelf or on the floor), cooling and power. In addition people are needed to actually run the system.

The PrestoPRIME document D2.1.1[21] has a lot of information on calculating the cost of storage for a given quantity of data so that will not be repeated here.

6.1. Processes

The task of managing the service's capacity is touched on by OAIS in the preservation planning functional entity, specifically the following sub-processes:

- Monitor Technology: monitoring existing technology for obsolescence and provide prototyping capability for evaluation of emerging technologies.
- Monitor Community: track changes in community requirements.
- Develop Preservation Strategies and Standards: recommend strategies; assess risks; advise on how to handle new requirements.

The administration functional entity is also relevant:

- Manage System Configuration: develops and implements plans for system evolution.
- Establish Standards and Policies: make appropriate decisions to minimise the risk of not fulfilling the archive's commitments.

OAIS does not have a great deal more to say about these processes. More informative are the processes documented in the Information Technology Infrastructure Library (ITIL) from the United Kingdom's Office of Government Commerce (OGC) which claims to be the most widely adopted set of principles for IT service management worldwide. The relevant parts are the "Capacity Management" process of the "Continual Service Improvement" the "Service Level Management" process discussed both in that volume and the "Service Design" volume.

The Service Level Management process has a mission statement: to *"Plan, coordinate, negotiate, report and manage the quality of IT services at acceptable cost."* The "process mission" should be achieved by implementing, amongst other things:

- Business-aligned IT services through a constant cycle of agreeing, monitoring and reporting

- IT Service Catalogue (setting out your services to your customers)
- Service Level Agreements for customers of IT services
- Operational Level Agreements and Underpinning Contracts with IT suppliers
- Reports on the quality of IT services on a regular basis

The ITIL book discusses these processes with respect to an IT department supporting the business of which it is a part, but the same principles surely apply to a business providing what is essentially an IT function. So, in a nutshell, define your service(s), agree SLAs and make sure they are underpinned by appropriate agreements with IT suppliers. Monitor and report on the quality of the service and repeat.

The capacity management process—*"To ensure that all current and future capacity and performance aspects of the IT infrastructure are provided to meet business requirements at acceptable cost."*—is most pertinent to this discussion. It is further broken down into three sub-processes:

- Business capacity management: to ensure that future business requirements are considered and understood, and that sufficient capacity to support the services is planned and implemented in the appropriate timescale.
- Service capacity management: to identify and understand the services, their resource usage, working patterns, peaks and troughs, as well as to ensure that services can and do meet their SLA targets.
- Component capacity management: to identify and understand the capacity and utilisation of each of the components of the IT infrastructure.

These sub-processes all share a common set of activities that are applied from different perspectives including modelling, service monitoring, optimisation and trend analysis. The processes take as input: performance monitoring, workload monitoring, application sizing, resource forecasting, demand forecasting and modelling. From these processes come the capacity plan itself, forecasts, tuning data and service level management guidelines.

ITIL also identifies workload management and demand management processes. Workload management can be defined as understanding which customers use what service, when they use the service and how using the service impacts the performance of a single or multiple systems and/or components that make up a service.

Demand management is the process of influencing users' behaviour in order to change the workload. Demand management can be an effective way of improving service performance without investing a lot of money. For instance, if the workload of a service can be smoothed out then infrastructure to support the average workload will suffice, but if the workload is very peaky then additional expensive hardware may be needed to support the peaks in demand. There are different ways to influence customer behaviour and some of these, such as charging more at peak times and providing information to customers are discussed in the example in Chapter 8.

6.2. Modelling

As already noted in this document, monitoring both the service and customer behaviour are very important in making sure both current service level agreements are kept to and understanding how additional customer demands can be met. Modelling goes hand in hand with monitoring. Monitoring data can be used to validate a system model, train a model and as input to a model that predicts future requirements.

Modelling itself varies from a domain expert making estimates based on experience to pilot studies and prototypes. Models can be used to predict the resources required for an SLA and to predict the affect of a change in the system (hardware or software).

There are a huge variety of modelling techniques available. The simplest models may just use trend analysis: taking the historical usage and extrapolating into the future. A domain expert can use this type of information and predict resource requirements for SLAs. Work along these lines was carried out in the SIMDAT project.[22]

Analytical models can be built to represent system behaviour using mathematical techniques such as queuing theory. Such models can be used to predict response time for instance. Data on expected customer and resource performance can be used to train models such as Bayesian belief networks and artificial neural networks. Stochastic models can then be built. The IRMOS project[23] is using models such as these as well as finite state machines to predict resource requirements for SLAs.

Finally, simulation modelling may be used to understand the effect of different customer workloads on a real or prototype system. Software can be used to simulate user behaviour (service requests etc), perhaps simulating high workloads not normally reached in day to day operation. In this way the behaviour of a system to workload can be accurately assessed.

7. SLA Standards and Reference Models

To implement the terms presented in Chapter 5 will require a complex monitoring and management system at the service provider. For instance, even a conceptually simple quality of service term such as “QS-03, SIP ingestion time” requires the different stages of the ingestion workflow to be automatically monitored and timed, reports sent to an SLA management system, combined and compared with the limit specified in the relevant SLA and perhaps resulting in a change in the allocation of resources to ingestion or information to be fed into a capacity management model. These monitoring and management processes are made significantly easier if the SLA is encoded in a machine-readable form which sets out the identifiers used for the various metrics and the limits that must be adhered to.

There has been a great deal of work in recent years in the field of machine-readable service level agreements and many specifications have been proposed. This chapter reviews the most prominent specifications and also some related work on data reporting specifications.

7.1. SLA Specification

The SLA specification languages reviewed here are: (i) HQML, (ii) Web Service Level Agreement (WSLA), (iii) SLAng, (iv) Web Service Management Language (WSML), (v) Web Service Offering Language (WSOL), (vi) W3C WS-Policy (WSP), (vii) WS-Agreement, (viii) WSDM, and (ix) WS-Management

HQML

The Hierarchical QoS Markup Language (HQML)[24] developed at the University of Illinois in 2002, is an XML based language to enhance distributed multimedia applications on the web with QoS capabilities. The design of *HQML* was based on two observations: (1) the absence of a systematic QoS specification language, that can be used by distributed multimedia applications on the WWW to utilize the state-of-the-art QoS management technology; and (2) the power and popularity of XML to deliver richly structured contents on the web.

HQML employs XML DTD as a schema model, which include tags such as <App>,<Configuration>,<Price>, and <PriceModel>. In order to use HQML, an associated visual QoS programming environment, called *QoSTalk*, has also been developed. The HQML schema is simple and it is more like a specification language for QoS management than a specification language for SLA. It is not closely tied to the use of web services. The proposed XML schema mixes the QoS metrics and price terms together. An example of using HQML for QoS specification is shown in Figure 14.

```
<App name = "Live Media Streaming" ServiceProvide
   <Configuration id = "100">
      <UserLevelQoS> high </UserLevelQoS>
   </Configuration>
   <Configuration id = "101">
      <Price unit = "$"> 5 </Price>
      <UserLevelQoS> Average </UerLevelQoS>
      <Price unit = "$"> 1 </Price>
      <PriceModel> flat rate </PriceModel>
      <PriceModlealt> rfate </PriceModel>
    </Configuration>
    <Configuration id = "200">
       <UserLevelQoS> high </UserLevelQoS>
       <Price unit = "$"> 2 </Price>
       <PriceModel> per hour increase <PriceModel>
     </Configuration>
</App>
```

Figure 14 An example of HQML

Web Service Level Agreement (WSLA)

The Web Service Level Agreement (WSLA)[25] is a specification language for service level agreements. It was proposed by IBM and version 1.0 was released in 2003. In WSLA, the structure of SLA can include: (i) Parties, (ii) Service definition, and (iii) Obligations.

- "Parties" define parties involved in the management of Web service such as customer, service provider, third parties, etc.
- "Service definition" describes service properties on which obligations are defined, which include (i) definition of the service, (ii) SLA parameters, and (iii) the way SLA parameters are measured and computed. In service definition, a term service object is used to describe what Web service operations SLA relate to. In WSLA, SLA parameters and metrics are distinguished: SLA parameters are defined by metrics. Metrics either define how a value is to be computed from other metrics or describe how it is measured.
- "Obligations" defines the service level that is guaranteed with respect to the SLA parameters, and promises to perform actions under particular conditions. It provides two kinds of guarantees: (i) Service Level Objective (SLO), and (ii) Action Guarantees. SLO expresses a commitment to maintain a particular state of the service in a given period, while Action Guarantees expresses a commitment to perform particular activity if a given precondition is met.

A top-level document structure using WSLA is shown in Figure 15.

```
<?xml version="1.0">
<wsla:SLA
xmlns:xsi="http://www.w3.org/2001/XMLSchema
xmlns:wsla="http://www.ibm.com/wsla
name="StockquoteServiceLevelAgreement12345"   >
  <Parties/>
  <ServiceDefinition>
  <Obligations/>
</wsla:SLA>
```

Figure 15 Overview of WSLA

One of unique feature of WSLA is that it supports logical expression. For example, the expression below states that the response time must be less than 0.5 unless the transaction rate is greater than 10000.

```
<Expression>
 <Or>
    <Expression>
       <Predicate xsi:type="Less">
          <SLAParameter>ResponseTimeThroughPutRati
          <Value>0.5</Value>
       </Predicate>
    </Expression>
    <Expression>
       <Predicate xsi:type="Greater">
          <SLAParameter>TransactionRate</SLAParame
          <Value>10000</Value>
       </Predicate>
    </Expression>
 </Or>
</Expression>
```

Figure 16 A WSLA logical expression

WSLA is fully documented and publicly available. The WSLA v1.0 specification clearly defines the structure of SLA, especially it distinguishes the SLA parameters and metrics. It provides a framework for specifying and monitoring Service Level Agreements (SLA) for Web Services. WSLA is also extensible. All these make WSLA promising as QoS/SLA specification language. However, one problem identified is that v1.0 of the specification was released in 2003, and there is little recent work. IBM have more recently been involved in the WS-Agreement specification (see below).

SLAng: A language for defining Service Level Agreements

SLAng as SLA language was developed by University of College London (UCL) under the TAPAS project (2002 - 2005). SLAng defines six different types of SLA, corresponding to service usages present in the model. These are divided into “vertical SLAs”, in which the service provides technical support for the client, and “horizontal SLAs” in which the client subcontracts part of the functionality of a service to a service of the same type. The hierarchical structure of SLAng’s syntax subdivides the SLA terms into SLA type specific groups. The terms are further subdivided into client, provider and mutual responsibility clauses.[26]

The Vertical SLAs are *Hosting* (between service provider and host), *Persistence* (between a host and storage service provider) and *Communication* (between application or host and Internet service providers). The Horizontal SLAs are *ASP* (between an application or service and ASP), *Container* (between container providers) and *Networking* (between network providers).

SLAng does not clearly describe the structure of the SLA. The classification of vertical SLAs and horizontal SLAs is easy to confuse people. The TAPAS project finished in 2005, and the further development of SLAng cannot be guaranteed.

Web Service Management Language (WSML)

Web Service Management Language (WSML) was developed in 2002 by HP Laboratories. It can be regarded as an extension of QoS Modeling Language (QML)[27] by allowing the definition of service level objectives, validity period and mathematical operations on measured data, etc. which were not supported in QML. According to[28], WSML can enable formal and unambiguous specification of information about when SLAs should be evaluated, which inputs should be considered for evaluation, where are the measurements

should occur, as well as what an how to evaluate. In addition, it is a flexible SLA formalisation, fully compatible with WSDL and WSFL (Web Services Flow Language).

However, WSML does not enable specification of management third parties. Further, WSML does not define the language for expressions to be evaluated. It is assumed that expressions will be written in some other mathematical languages, such as MathML. This means that the infrastructure for the evaluation of WSML constraints should also support these mathematical languages.

An example of WSML is shown in Figure 17.

```
…
<SLA id="stationaryShop1">
    <startDate>2001-01-01</startDate>
    <endDate>2002-01-01</endDate>
    <nextEvalDate>2001-02-01</nextEvalDate>
  <provider>Stationery.com</provider>
 <consumer>OfficeSupplies.com</consumer>

<SLO id="SLO1">
   <daytimeConstraints>
 <clause id="SLO1Clause1">
 <measuredItem>
     <item>
          <constructType>wsdl:operation</const
        <constructRef>osn:processEstimate</cc
     </item>
 </measuredItem>
 <avgResponseTime>5</avgResponseTime>
 </clause>
</SLO>
…
```

Figure 17 An example of WSML

Web Service Offerings Language (WSOL)

Web Service Offerings Language (WSOL)28 claims to be a language for the formal specification of various constraints, management statements, and classes of service for Web Services. It was developed in 2003 by Carleton University, Canada. WSOL is based on the following specification constructs: (i) constraints (Boolean expressions), (ii) statements (e.g. price, penalty, management responsibility), (iii) constraints groups, (iv) constraints group template, and (v) service offering. For example, constraints defined in WSOL include functional constraints, QoS and access rights. WSOL uses the service offering as a formal representation of a single class of service of one web service.

The development of WSOL has made much reference to the WSLA and WSML work discussed previously. One of the distinct features of WSOL is that it has defined an external ontology of QoS metrics and measurement units for the specification of QoS constraints. In their current implementation of WSOL, it is assumed that ontologies of QoS metrics are collections of names with information about appropriate data types and measurement units. Similarly, ontologies of measurement units are simple collections of names without any additional information.

WSOL claims that it contains formal representation of various constraints: functional (pre-, post-, and future conditions), Quality of Service (QoS, a.k.a. non-functional, extra-functional), and access rights. It also contains management statements, such as

statements about prices, monetary penalties, and management responsibilities. One Web Service can be associated with multiple service offerings.

An example WSOL service offering is shown in Figure 4.

```
<wsol:serviceOffering name="SO1" service="
buyStockService" accountSiUnPgPPLaIrEtRyW=S""W>SOL
   <wsol:instantiate CGTName="CGT2" resServv
        "buyStock:buyStockService" resPEoVrEtROYr"P
         resOperatio-nE=V"EWRSYO"L resCGName="CG5">
      <wsol:parmValue name="maxResTime">
         <wsol:numberWithUnitConstant>
         <wsol:value>300</wsol:value>
         <wosl:unit type="QoSMeasOntology:mill:
         </wsol:numberWithUnitConstant>
      </wsol:parmValue>
</wsol:instantiate>
...
</wsol:serviceOffering>
```

Figure 18 WSOL service offering

WS-Agreement

WS-Agreement is a popular standard for aggregation into web service architectures to support the management of non-functional requirements in web-services[29]. WS-Agreement, proposed by the Global Grid Forum (GGF), now the Open Grid Forum (OGF), describes a protocol for establishing an agreement on the usage of services between a service provider and a consumer. It defines a language and a protocol to represent the service providers, create agreements based on offers and monitor agreement compliance at runtime.

WS-Agreement the expressive power to describe service level objectives, which state the requirements and capabilities of each party with respect to the availability of resources and service qualities. An agreement consists of the agreement name, its context and the agreement terms. The context contains information about the involved parties and metadata such as the duration of the agreement. Agreement terms define the content of an agreement: Service Description Terms (SDTs) define the functionality that is delivered under an agreement. A SDT includes a domain-specific description of the offered or required functionality (the service itself). Guarantee Terms define assurance on service quality of the service described by the SDTs. They define Service Level Objectives (SLOs), which describe the quality of service aspects that have to be fulfilled by the provider. The structure of an agreement is shown in Figure 19. The context section contains the meta-data for the agreement including the names of the participants and the agreement lifetime. The terms section describes the agreement itself with at least one service definition term and zero or more guarantee terms. The precise language of the terms section is not defined by Ws-Agreement. In fact two of the aims of WS-Agreement are to permit domain-specific service terms and to allow the use of any condition specification language.

```
<wsag:Agreement AgreementId="xsd:string">
    <wsag:Name>
        xs:NCName
    </wsag:Name> ?
    <wsag:AgreementContext>
        wsag:AgreementContextType
    </wsag:AgreementContext>
    <wsag:Terms>
        wsag:TermCompositorType
    </wsag:Terms>
</wsag:Agreement>
```

Figure 19 WS-Agreement structure

WS-Agreement depends on some other WS-* specifications: WS-Addressing, WS-ResourceProperties, WS-ResourceLifetime and WS-BaseFaults. The last three of these specifications are components of WSRF.

WS-Policy

WS-Policy is now a W3C recommendation[30] (since September 2007). WS-Policy is a standard to describe the properties that characterize a Web service. By means of this specification, the functional description of a service can be tied to a set of assertions that describe how the Web service should work in terms of aspects like security, transactionality, and reliable messaging. According to WS-Policy, the assertion is defined as "an individual preference, requirement, capability or other property", and the WS-Policy document is in charge of composing such assertions to identify how a Web service should work. These assertions can be used to express both functional aspects (e.g., constraints on exchanged data), and non-functional aspects (e.g., security, transactionality, and message reliability). An example of WS-Policy is shown in Figure 20.

```
<wsp:Policy ..>
   <wsp:ExactlyOne>
     ( <wsp:All> ( <Assertion ..> ...</Assertion> )* </wsp:All> )*
   </wsp:ExactlyOne>
</wsp:Policy>
```

Figure 20 WS-Policy example

Apache Neethi[31] provides general framework for programmers to use WS-Policy. It is compliant with latest WS-Policy specification. This framework is specifically written to enable the Apache Web services stack to use WS-Policy as a way of expressing the requirements and capabilities.

Although WS-policy is recommended by W3C, in comparison with WS-Agreement, it actually provides no advantage for QoS specification, other than it is a standard way of associating QoS-like descriptions with service[32]. Also WS-Policy does not support negotiation or monitoring of compliance at runtime which SLA management systems need.

Web Service Distributed Management (WSDM)

The Web Services Distributed Management (WSDM)[33] standard published by OASIS contains two parts, Management Using Web Services (MUWS) and Management of Web Services (MOWS), which defines the methods, structure, and specification of a system for

managing network resources (printers, routers, servers and services, for example) and for managing Web services.

MUWS consists of two main standards: MUWS Part 1 (MUWS1) and MUWS Part 2 (MUWS2). MUWS1 defines the properties of the resource that are required to interface to the Web services. For example, MUWS1 part components in the definition might define the resource ID used to identify the system. MUWS2 defines the standard used to specify support for manageability capabilities. In MUWS2, the definition of the resource is handled through capabilities, which include functionality, properties, and other settings. The MUWS standard includes definitions for the following capabilities:

- Identity -- the identity of the resource
- Description -- defines the list of captions, descriptions, and version information used to provide a human-readable identity for the resource
- Manageability Characteristics -- describes the properties of the interface for managing this component
- Correlatable properties -- defines the properties that determine if two manageable resources with different identities are actually the same resource
- Metrics -- defines how to represent and access information about a specific property
- Configuration -- defines how to change the configuration of a resource
- State -- defines how to change the state of a resource
- Operational Status -- defines the status levels for a resource. The MUWS specification includes three basic states (available, unavailable, and unknown)
- Advertisement -- defines the event to be raised when a new manageable resource is created

The WSDM-MOWS specification is an extension of WSDM-MUWS that defines how to manage a Web service. MOWS capabilities are similar to MUWS definitions, but the standard schemas provide information about properties specific to the process of managing Web services. In particular, the MOWS standard includes definitions for the following:

- Identity -- a unique identity for the service
- Identification -- a human readable identification of the resource
- Metrics -- a number of basic metric information is defined within the standard, including NumberOfRequests, NumberOfFailedRequests, NumberOfSuccessfulRequests, ServiceTime, MaxResponseTime, and LastResponseTime.
- OperationalState -- the MOWS operational state provides two main states, Up and Down, with sub states; Busy and Idle for Up; Stopped, Crashed, and Saturated for Down.

- OperationalStatus -- a summary of the current status, based on the MUWS OPerationalStatus.

The MOWS schema also includes standard definitions for a number of metric types, including IntegerCounter (as used for the NumberOfRequests metric) and DurationMetric (as used for MaxResponseTime).

WSDM also contains an event model. WSDM events communicate information between different components in a WSDM system.

WSDM provides definitions for metrics and measurement (e.g. NumberOfRequests, operation state, operation status). However, as the aim of WSDM is about management using Web service and management of Web services, rather than focussing on defining capabilities and requirements of service providers and customers, the support for defining metrics and measurement is limited. A domain-specific model or domain ontology needs to be defined. WSDM provides an event model which could be used for SLA monitoring, but it does not provide support for SLA negotiation.

WS-Management

WS-Management[34] is a specification for managing devices, computers, web services and other applications using web services. It was proposed by the Distributed Management Task Force (DMTF) and published in 2004 with support from IT companies such as AMD, Dell, Intel, Microsoft and Sun. DMTF is a standards organisation that develops and maintains standards for systems management of IT environments in enterprises and the internet.

WS-Management has some overlapping area with the MUWS part of and there is also a mapping of the DMTF Common Information Model into WS-Management. Microsoft have implemented WS-Management as Windows Remote Management (WinRM) and use it to enable the execution of scripts on remote machines.

Similar to WSDM, WS-Management aims at management using Web service, hence has same disadvantage as WSDM does. WS-Management is not strong enough to be used for service level agreements. For example, it does not provide a negotiation model for SLA, not appropriate for defining agreement between customer and service provider.

Comparison Matrix

Specification Languages	Description	Strength	Weakness	SLA Process / Usage	Standard
HQML	The Hierarchical QoS Markup Language (HQML) developed at the University of Illinois in 2002, is an XML based language to enhance the distributed multimedia application over Web with QoS capabilities.	- Suitable for QoS representation	- Not appropriate for SLA - Not tied up to Web service - Out of date	Out of date	No
WSLA	The Web Service Level Agreement (WSLA) is a specification language for service level agreement. It was	- A framework for specifying and monitoring SLA - Fully documented and publically available. - Widely used.	- The last release version 1.0 released in 2003 - Some of content becomes part of WS-	- Support SLA negotiation, SLA deployment, measuring and	No

Specification Languages	Description	Strength	Weakness	SLA Process / Usage	Standard
	proposed by IBM and version 1.0 was released in 2003.		Agreement	reporting, corrective management actions, termination - Still being used	
SLAng	SLAng as SLA language was developed by University of College London (UCL) under the TAPAS project (2002 - 2005).	- Defines six different types of SLA, corresponding to service usages present in the model - Specify vertical SLA and horizontal SLA	- Not widely used - Further support is not guaranteed	Not widely used	No
WSML	Web Service Management Language (WSML) was developed in 2002 by HP Laboratories. It can be regarded as an extension of QoS Modeling Language (QML)	Can specify about: - when SLAs should be evaluated, - which inputs should be considered for evaluation, - where are the measurements should occur - what an how to evaluate	- Not widely used - Out of da te	Not widely used	No
WSOL	Web Service offering Language (WSOL) was developed in 2003 by Carleton University, Canada. It claims to be a language for the formal specification of various constraints, management statements, and classes of service for Web Services.	- The distinct feature of WSOL is that it has defined external ontology of QoS metrics and measurement units for the specification of QoS constraints. - Can reference WSDL file and contain information that is not supported in WSDL	- Not widely used - Out of date	Not widely used	No
WS-Policy	WS-Policy, is a W3C recommendation since September 2007. It is a standard to describe the properties that characterize a Web service. It provides a set of assertions that describe how the Web service should work in terms of aspects like security, transactionality, and reliable messaging.	- W3C recommendation - Emerging standard of SOA - Extensible specification language - Apache implantation *Neethi* - Full client-side support in AXIS2	- No advantage for SLA specification, other than a standard way to associate some descriptions with Web service. - Cannot support SLA negotiation, SLA monitoring, etc. - Need to design domain-specific schema/ontology	Used in some projects for SLA (e.g. GlueQos), usually with extension by incorporating domain ontology	Yes W3C
WS-Agreement	WS-Agreement proposed by OGF describes a protocol for establishing an agreement on the usage of services between a service provider and a consumer. It defines a language and a protocol to represent the service providers, create agreements based on offer and monitor agreement compliance at runtime.	- Proposed by OGF - More expressive power to describe service level objectives - Able to represent the service providers, create agreements based on offers and monitor agreement compliance at runtime - An extensible language, the specification of domain-specific terms is open - Widely used in Grid computing	- No semantic support - Need to provide domain-specific schema / ontology - The WSAG4J implementation seems not widely used - IBM implementation Cremona seems not public	- Support SLA negotiation, monitoring - Widely used in many projects, e.g. Assess Grid, VIOLA MSS, ASKALON, CSF, CATNETS, JSS	Yes OGF
WSDM	Web Service Distributed Management (WSDM) proposed by OASIS, is a specification about Management Using Web Service (MUWS) and Management Of Web Service (MOWS).	- A standard by OASIS - Provide definition of capabilities for resource and Web Service. - Defines a number of basic metric information - Define standard to define manageability capabilities	- Aim at system management, rather than an agreement on the usage of services between a service provider and a consumer - Need to provide domain-specific schema/ontology	- Not support SLA negotiation, etc. - Not widely in SLA	Yes OASIS

Specification Languages	Description	Strength	Weakness	SLA Process / Usage	Standard
WS-Management	Proposed by DMTF (a standard organisation) in 2004, supported by IT companies such Microsoft, DELL, Intel, SUN, etc. is a specification for managing devices, computers, Web service and other application using Web Services.	- A standard by DMTF - Compliance with CIM model proposed by DMTF - Promote interoperability between management application and managed resources	- Aim at system management, rather than an agreement on the usage of services between a service provider and a consumer - Need to provide domain-specific schema / ontology	- Not support SLA negotiation, etc. - Not widely in SLA	Yes DMTF

Table 10 Comparison of SLA specification languages

7.2. Related Specifications

WS-Agreement does not mandate a particular language for describing agreement terms such as constraints. Also, in addition to describing the SLA, management systems also need to be able to report the usage of services. Therefore we report briefly here on some related specifications for describing data.

UncertML

UncertML[35] is a conceptual model and XML encoding designed for encapsulating probabilistic uncertainties. It comes from the geographic information systems (GIS) field and is currently being evaluated by the Open Geospatial Consortium (OGC). It enables the expression in XML of uncertainties in several ways, such as probability distributions, statistics (means, standard deviations, etc) and sampled data sets. Figure 21 shows how a standard deviation is represented in UncertML.

```
<un:Statistic definition="http://dictionary.uncertml.org/statistics/standard_deviation">
   <un:value>12.08</un:value>
</un:Statistic>
```

Figure 21 An example of UncertML

UncertML may be useful in two ways: (1) when reporting usage data can be aggregated and statistical reports generated instead of reporting the raw monitoring data and (2) when negotiating an SLA it may be useful to be able to define constraints on either party in statistical terms instead of hard guarantees.

UCUM

When notating quantities in SLAs or usage reports it is important to pay attention to the units.

A useful reference for units is the Unified Code for Units of Measure[36] (UCUM). It is intended to include all units of measures being contemporarily used in international science, engineering, and business. It defines a single representation for each unit and shows how derived units are defined in terms of base units.

Ideally an SLA management system should be able to convert between units where appropriate, but that is an implementation detail. In the context of this discussion it is important to note that understanding and recording the units of the measurement is crucial.

The “unit” of a measurement does not just mean considering SI or metric units and the standard set of dimensions (i.e. length, weight, time, etc). In many cases we need to record what UCUM calls “non-units” such as “CPU” or “files”.

7.3. Conclusions

There is no clear “winner” yet in the SLA specification battle. It is also not clear that any of the contenders will ultimately be suitable for use as part of an automated SLA and management system for digital preservation. WS-Agreement stands out as having a strong following and also permitting the terms of the SLA to be defined in whatever way is necessary for the domain. This provides great flexibility but may also be a reflection on the maturity of these specifications as much as anything else: the popular specification is the one that doesn’t try to specify the most important part.

8. Example

Two of the partners in PrestoPRIME also work together commercially:

- Sound and Vision provides access to 700,000 hours of Dutch television, radio, music and film.
- Technicolor: provide the storage systems for the content.

Both Sound and Vision and Technicolor are located on the Hilversum Media Park in the Netherlands. Sound and Vision has an SLA with Technicolor for the provision of storage services. Whilst the detail of the SLA is commercially sensitive we are able to discuss some aspects of the relationship between Sound and Vision and Technicolor.

In the opinion of both Technicolor and B&G, the main tasks of an archive service provider are to:

1. Provide a responsive ingest facility.
2. Keep the material safe.
3. Deliver the material whenever requested.
4. Provide additional services like file integrity checking.
5. Provide a conservation plan or assist in generating one.
6. Provide an 'exit-strategy' to keep the material safe in case of switching service provider when necessary.

Key performance indicators (KPIs) can be created to measure these aspects.

8.1. Ingest and Delivery

For ingest, the KPI suggested by Technicolor is:

"Broadcasted material will be available in the archive and searchable within 4 hours of the broadcast."

Delivery is a little more complex. The delivery of material can be done through many different channels: a browse copy over the internet in a web browser, a copy delivered on DVD, delivery via an FTP server or physical delivery of a copy on tape.

Specifying the availability of the systems that provide download services is consistent with the approach normally offered by e.g. Amazon or ISPs but does not have any direct relation with the perception of the user. Better KPIs are:

"95% of the material requested from the archive is downloaded within a specific time (e.g. 30% of the length of the material) to the download server."

or

"Requested Digibeta tapes are delivered by 10 am the next morning to the archives desk."

Technicolor has used indicators like these in the past but has found them to be too rigid in some situations. In practice, problems arose mostly from the unpredictable behaviour of the users:

- Some users order for example 20 total programmes. Next a news reporter wants to download 1 minute from a programme and is 21st in the queue. In this case we couldn't meet the KPI for 1-minute shot because of the queue that occurred. Moreover even if we had met the KPI (and 20 programmes had been delivered on time and the 1-minute hadn't but nevertheless we would have delivered more than 95% on time, the user perception of the quality would not have been positive as the news reporter would not have received the important item.

- It was expected that user would be requesting mostly the 'newest' material however in practice 50% of the requested material turns out to be older than 1 year.

- The system is not scaled to peak time behaviour (traffic jams during rush hour). During peak times it becomes overloaded. Since the users are not aware of this overload (they are sitting behind their desks and do not notice the load on the system directly) they become agitated.

- Problems in the delivery channel. One of our services is to 'push' the content to a local server of the user instead of delivering the material on a central server where the user can download the material. More than once we faced problems with this receiving server (full storage or not up and running) so we could not deliver our files.

To counter these problems, several measures can be taken:

1. Increasing the system capacity: this improves the performance but is expensive and in isolation is limited in its effect as the load on the system often grows to fill the capacity.

2. Regulate user behaviour: a higher price can be placed on downloads during peak times and users can be asked to indicate the priority of a request.

3. Visibility of the queue: if the users can see the download queue then correct expectations are set. It is then possible for users to use an emergency procedure to bump a request up the queue.

4. Traffic management: through fast and slow lanes. Users working at a news department might automatically get access to the fast lane.

5. Intelligent prioritisation: the priority of a request can be judged somewhat according to the kind of material, the amount of material and the time of day or time of week the request is being made.

In addition, within an SLA provision can be made for different performance guarantees in a similar way to the different levels of service offered by postal services. For instance, items could be requested for delivery before 0900 or before 1200 with different prices and guarantees for each.

At this moment various 'lanes' as suggested in point four are implemented in the service from Technicolor to Sound and Vision. The news department always has the highest priority. Normal orders are handled sequentially according to the already existing functionality and policies. Batches of orders and physical DVD orders are put in a so called 'slow lane' (we are still looking for a more appealing name) that are dealt with during the night, and hence are delivered next morning. This solves two of the problems mentioned above: first the news department will always be served as soon as possible, second the large batches are postponed and cannot interfere with other (normal) orders.

These measures considerably improve the performance service. In addition whether extra hardware might also be necessary is under investigation. It makes sense to expand resources (e.g. input and output resources) whenever numbers of users expand.

Regulating the performance by varying the price is difficult in the example of the Technicolor service to Sound and Vision since groups of users are only charged a flat fee for specific content.

8.2. Keeping Material Safe

When storing valuable content such as cultural heritage audio visual material, an archive should be able to expect more from a service provider that just an availability number of a service.

An interesting KPI could be the number of programmes damaged or lost from the archive. But this figure should preferably be zero and any provider should at least say they can meet this value. If they don't the damage has already been done. So more important is that the archive can trust the service provider. This comes back to the aspects discussed here in Chapter 6 such as:

- Measures taken to prevent loss of material
- Back up and disaster recovery procedures
- Test procedures for e.g. new technology
- Quality Assurance procedures/department
- Certification

As an additional example there are also some organisational measures taken into account in the service of Technicolor for Sound & Vision. When dealing with a service provider an archive should take into account that there are some unlikely yet potential risks involved in outsourcing (part) of the archiving service. What happens for example if the service provider decides to terminate the service, or worse, the company is terminated due to internal or external causes? In this case the archive has a huge problem because a multi-petabytes archive cannot be transferred overnight to a new provider.

For these situations Technicolor and Sound & Vision have agreed on a so called "exit plan". This means that a scenario is lying on the shelf in case such a situation might occur. Arranging an exit plan is not straightforward and takes effort. Drawing one up requires legal agreements but also a detailed insight in the current services and sharing of sensitive commercial information that would not normally be disclosed. Therefore, this requires

more that just a customer-supplier relation but more a partnership. Additional effort must be expended to keep the plan up-to-date otherwise it looses its use. However, having such an agreement is important as it can guarantee the long term safekeeping of the material.

9. Conclusion

Terms in an SLA must be of relevance to the customer, so how quickly an item will be delivered is more useful than the uptime of the service. A detailed proposal for terms to be included in a preservation service provider SLA has been made here and will be refined and expanded upon during the PrestoPRIME project, including in D6.2.7. Not every aspect of a service provider's performance can be measured though, so evidence of trustworthiness for instance is important in judging whether ingested material will be kept safely. This document has reported on a survey conducted amongst the AV preservation service provider community which has found reasonable awareness of the use of audits for judging trustworthiness and a good acceptance of the principle.

We have presented how measures must be taken to balance the load on a system so that the key performance indicators are generally met. If a system is overloaded, mechanisms need to be in place to ensure that the system fails in the best way, e.g. delivers the high priority item but fails on the low priority. Data on user and system behaviour must be recorded to understand how to best manage the service. A variety of modeling techniques have been presented that can be applied to monitoring data to predict future requirements and investment.

In order to efficiently deal with the wealth of information that can be gathered from an operating service and to provide the best possible conformance to the limits defined in an SLA, the service must ideally be monitored and managed automatically by other computer systems. This leads to the requirement for machine-readable SLAs, formats for which have been investigated in this document. We recommend that the PrestoPRIME project further investigates the best way to manage SLAs and the associated services.

The example of Sound and Vision and Technicolor remind us that maintaining a successful relationship between a service provider and its customers is not just about defining and conforming to an SLA. Informing the user of the system status and developments and giving the user control where possible are both important factors in the relationship. Finally, recognising that things won't always go as planned is important, and workarounds and emergency procedures must be defined.

10. Glossary

Term	Definition
AV	Audiovisual
Competence Centre	One of the principal objectives of PrestoPRIME: Networked Audiovisual Competence Centre to gather and organise the knowledge created by the Project, as well as other previous projects, and use it to advance digital preservation activity and services.
CPU	Central processing unit: the main microchip in a computer
Demux	Demultiplexer: a device (hardware or software) that takes a single input and can output one of the many possible signals contained therein. For instance, taking a video file and outputting the audio track.
DIP	Dissemination information package (from OAIS)
DTD	Document type definition: used to describe the elements and references that may appear in an XML document
Ingest	The process of adding data into an archive
IPR	Intellectual property rights
JPEG2000	An architecture for lossless and visually lossless image compression that supports multi-resolution imaging and scalable image quality
MXF	Material Exchange Format: a container format for professional digital video and audio media defined by a set of SMPTE standards.
OAIS	Open Archival Information System: a reference model developed by CCSDS
QoS	Quality of Service: forms part of a Service Level Agreement. Quantitative definition of the service to be delivered that can be measured using a set of metrics. For example, QoS of a media streaming service might be defined in terms of acceptable bandwidth, jitter, data loss etc.
SIP	Submission information package (from OAIS)
SLA	Service Level Agreement. A formal and negotiated agreement between a service provider and service consumer. In the context of software services, SLAs are part of policy-based service governance, i.e. all terms of the service are described in the SLA and the service provider manages the service so it conforms to the SLA.
XML	Extensible mark-up language: a set of rules for encoding documents electronically

11. Annexes

11.1. Digital Preservation Services Questionnaire

The questionnaire reported upon in Chapter 2 was conducted online. For reference, screenshots of the survey are included below.

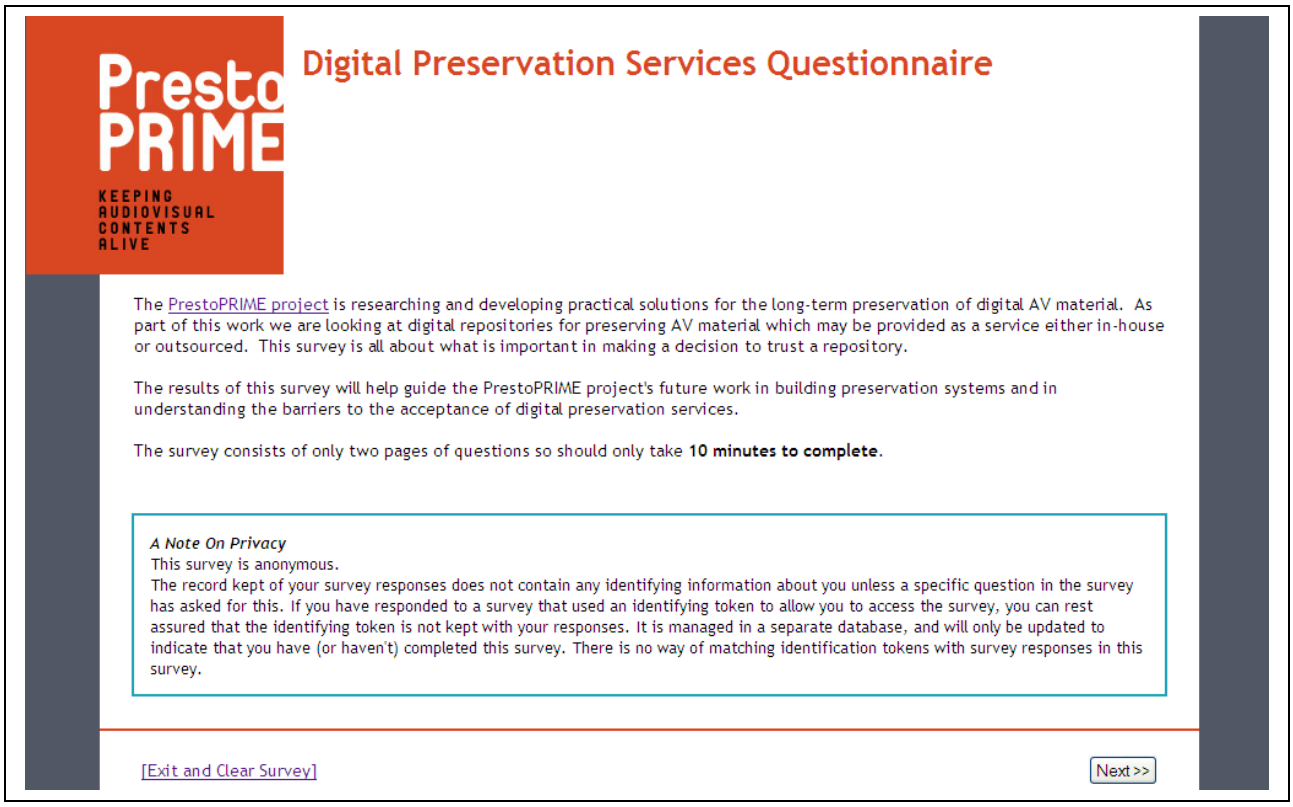


Figure 22 Introductory screen

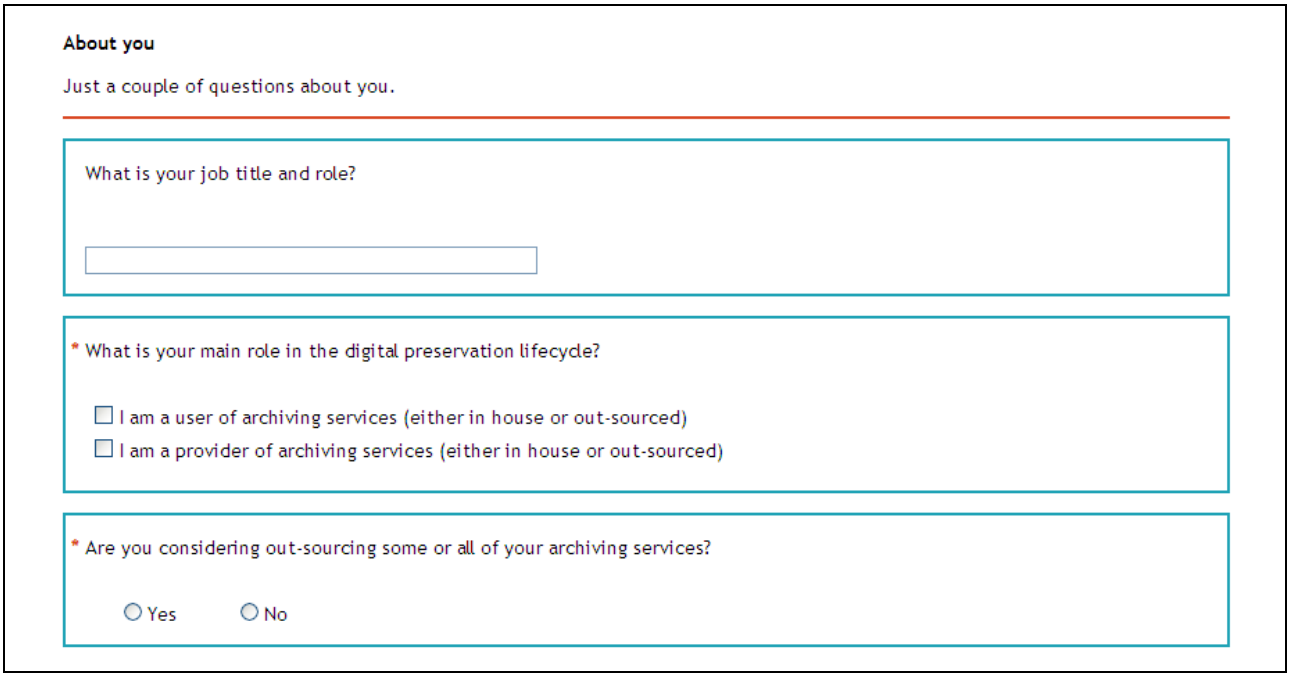


Figure 23 First page of questions.

Trusting a Service Provider

This is the final page of questions.

Placing AV material into a digital preservation service requires trust. We are interested in which aspects are most important when determining this trust. Please rate how important the following factors are to you.

* **Governance**

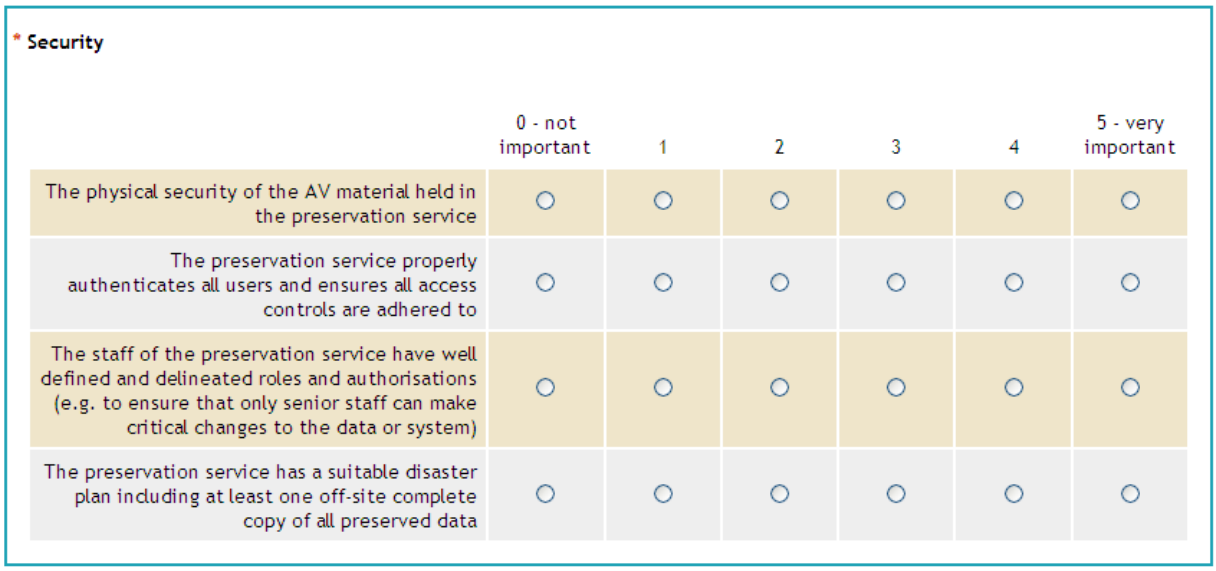
	0 - not important	1	2	3	4	5 - very important
The preservation service has long and short term business plans demonstrating financial sustainability	○	○	○	○	○	○
The preservation service has the appropriate number of staff and a professional development plan	○	○	○	○	○	○
The preservation service monitors technological developments and properly plans technical changes	○	○	○	○	○	○
The preservation service has a succession plan (what should happen if the service ceases to exist)	○	○	○	○	○	○
The preservation service tracks and manages intellectual property rights	○	○	○	○	○	○

Figure 24 Top of the second page of the questionnaire.

* **AV material management**

	0 - not important	1	2	3	4	5 - very important
The preservation service can authenticate the source of all AV material	○	○	○	○	○	○
The preservation service logs all preservation actions	○	○	○	○	○	○
The preservation service has a clear preservation plan (e.g. when to migrate)	○	○	○	○	○	○
The preservation service actively monitors the integrity of preserved AV material (e.g. using checksums)	○	○	○	○	○	○

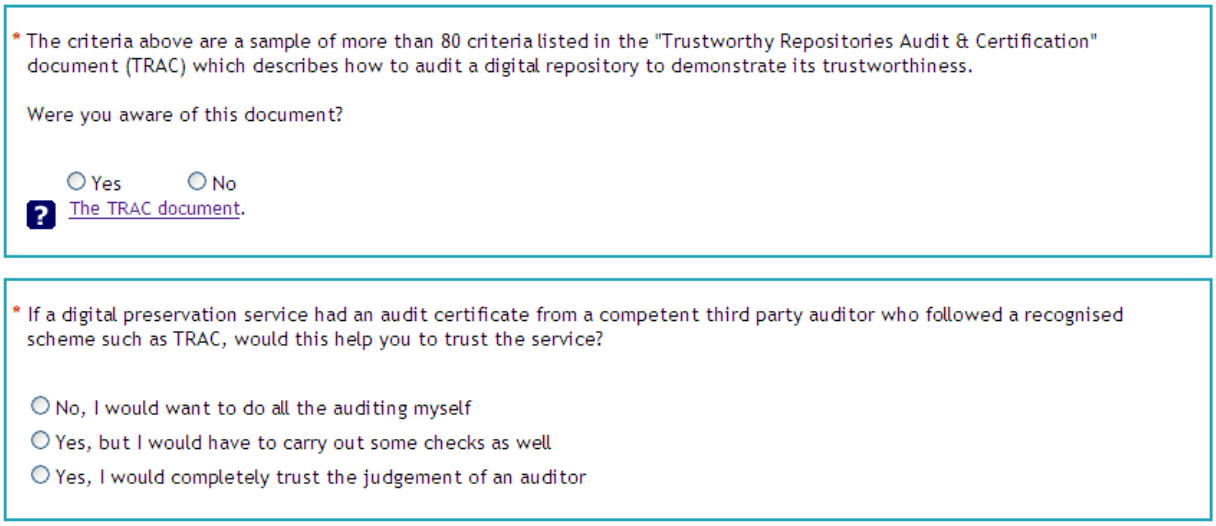
Figure 25 Questions on AV material management.



* **Security**

	0 - not important	1	2	3	4	5 - very important
The physical security of the AV material held in the preservation service	○	○	○	○	○	○
The preservation service properly authenticates all users and ensures all access controls are adhered to	○	○	○	○	○	○
The staff of the preservation service have well defined and delineated roles and authorisations (e.g. to ensure that only senior staff can make critical changes to the data or system)	○	○	○	○	○	○
The preservation service has a suitable disaster plan including at least one off-site complete copy of all preserved data	○	○	○	○	○	○

Figure 26 Questions on security.



* The criteria above are a sample of more than 80 criteria listed in the "Trustworthy Repositories Audit & Certification" document (TRAC) which describes how to audit a digital repository to demonstrate its trustworthiness.

Were you aware of this document?

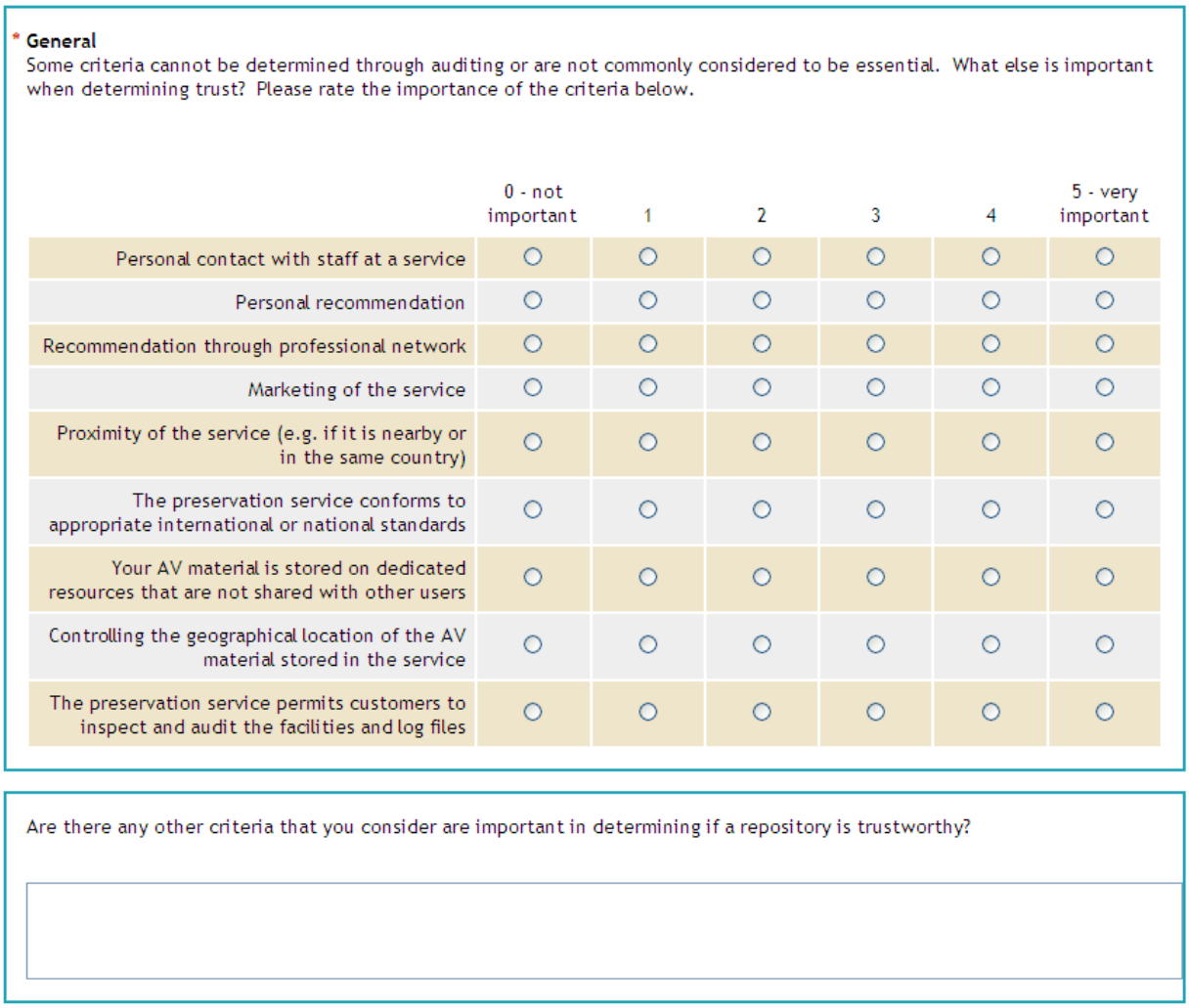
○ Yes ○ No

? The TRAC document.

* If a digital preservation service had an audit certificate from a competent third party auditor who followed a recognised scheme such as TRAC, would this help you to trust the service?

○ No, I would want to do all the auditing myself
○ Yes, but I would have to carry out some checks as well
○ Yes, I would completely trust the judgement of an auditor

Figure 27 Questions about TRAC.



* **General**
Some criteria cannot be determined through auditing or are not commonly considered to be essential. What else is important when determining trust? Please rate the importance of the criteria below.

	0 - not important	1	2	3	4	5 - very important
Personal contact with staff at a service	○	○	○	○	○	○
Personal recommendation	○	○	○	○	○	○
Recommendation through professional network	○	○	○	○	○	○
Marketing of the service	○	○	○	○	○	○
Proximity of the service (e.g. if it is nearby or in the same country)	○	○	○	○	○	○
The preservation service conforms to appropriate international or national standards	○	○	○	○	○	○
Your AV material is stored on dedicated resources that are not shared with other users	○	○	○	○	○	○
Controlling the geographical location of the AV material stored in the service	○	○	○	○	○	○
The preservation service permits customers to inspect and audit the facilities and log files	○	○	○	○	○	○

Are there any other criteria that you consider are important in determining if a repository is trustworthy?

Figure 28 General questions at the end of the questionnaire.

11.2. SLA Modeling

For convenience we report here the complete tables regarding the modelling of SLAs.

Capabilities

ID	Name	Requisite Level	Description	Remarks	Referenced metrics
CAP-01	Ingestion/Delivery	Mandatory	Capability to upload/download material and related metadata in the form of an agreed SIP, more protocols (e.g. ftp, http) and modalities (e.g. push, pull) should be provided		ME-05 ME-07
CAP-02	Validation	Mandatory	Identification and formal check of formats including wrappers, AV essence and metadata formats.	Validation can be at different level, from simple identification of wrappers to formal and structural validation of media	ME-05
CAP-03	Partial extraction	Mandatory	It is essential to have the possibility to ask for a specific portion of a stored object e.g. from frame 5000 to frame 51500 of a certain programme instance		ME-07
CAP-04	Search/Retrieve by ID	Mandatory	Possibility to ask for a specific stored object by its unique ID		ME-08 ME-09

ID	Name	Requisite Level	Description	Remarks	Referenced metrics
CAP-05	Backup/Restore	Mandatory	The provider manages periodic backups on different systems (e.g. offline supports like data tapes), the restore capability should be periodically tested independently of effective necessity.		
CAP-06	Access control	Mandatory	The ability of the owner of the content to set rules on who can access the content, when and in what form	The service provider should agree with the customer who can do which operations. What the customer will do with extracted material is a matter of rights and is not a concern of the provider.	
CAP-07	Rights management	Mandatory	This is a complex aspect to be better investigated; it is only partly related with Access control. As a minimum feature the system has to be able to understand (validate) rights management metadata and forward them as part of the delivery package to the customer.	Rights is mostly a matter of usage for publication, the responsibility thus should be of the publisher (e.g. the broadcaster) that definitively is not the preservation provider	
CAP-08	Audit/report	Mandatory	The ability to request an audit (technical, financial, process) of the contents held by the service provider on behalf of the content owner. Technical = what is there, what format is it in etc. Financial = how much is the service costing, what are the charges. Process = what actions have been performed on the content whilst at the service provider.		
CAP-09	Package update	Highly Recommended	Capability to accept revisions of metadata, but eventually also some change regarding part of AV essence		ME-05
CAP-10	Transcode /Transwrap	Highly Recommended	The capability to transcode between different audio-video formats and different containers (the same could be for metadata formats)		ME-07
CAP-11	Migration	Highly Recommended	The capability to manage massive batch processes of migration from some audio-video (and eventually metadata) formats and wrappers to others in order to preserve the usability of the content. Migration should be mandatory if does not exist an alternative mechanism like multivalent.		
CAP-12	Disaster recovery	Highly Recommended	The ability to restore the material even in case of natural disasters like fire, flooding, earthquake and partial or complete physical destruction of media and hardware systems. The typical way to face the problem is the replica of the information in a geographically distinct venue.	A customer could decide to have this protection using a different (more specialised) provider or setting up an internal system to this end	
CAP-13	Demux	Recommended	The possibility to ask and obtain only a specific track of audio / video to be extracted from the multiplex		ME-07
CAP-14	Upload/Download Resume	Recommended	The possibility to resume the download/upload operations when the process has been interrupted because of network or application troubles, the process could also have been paused by the customer		ME-10

ID	Name	Requisite Level	Description	Remarks	Referenced metrics
CAP-15	Search/Retrieve by meta	Recommended	Possibility to ask for a specific stored objects by specifying several kind of metadata		ME-08 ME-09
CAP-16	Fast preview	Recommended	Ability to show a browsing quality version of multimedia content e.g. movie at internet quality, shots with key frames		ME-11
CAP-17	Redundancy	Optional	It could be thought as an internal feature of the archive system (it directly conditions the Integrity guarantee), the customer however could be interested in knowing the level of redundancy feeling safer for example if he knows that there are 2 available copies of the same objects.	Redundancy is also a way to augment the availability	
CAP-18	Collections management	Optional	Ability to manage relations between objects e.g. a set of objects belonging to a single collection		
CAP-19	Advanced Search	Optional	The ability to perform complex queries like similarity search, search by video/audio content, semantic search		ME-08 ME-09
CAP-20	Retention of original files	Optional	The ability to require the system to retain (in the storage) the original files exactly as they were submitted. Note that by default the system gives guarantees on A/V and metadata quality while the format could change overtime according to preservation strategies.	This capability could be very important for some customers who want to reserve the possibility to access the original versions.	
CAP-21	Delivery of partially corrupted files	Recommended	The ability to delivery a file even if it is partially corrupted (think about a bad sector of a disk that affects the file). Before delivery there could be a restoration attempt with a notification of that to the customer.	In case of corruption the delivered package should be documented with the kind and entity of the damage (e.g. which frames are affected).	ME-02, ME-03

Features of Interest

ID	Name	Description	Behaviour	Referenced metric
F-01	Availability of services	Is the Boolean function of time (works or does not work) representing the fact that the service is up and effectively usable. This feature should be detailed for each atomic service making up the entire preservation service, should be considered at least: - SIP ingestion (and update) service - Search service - Browse service (if present) - DIP delivery service	For a big and differentiate amount of reasons the system could not be usable by the customer, the course of the availability is not predictable but the customer can	ME-01

ID	Name	Description	Behaviour	Referenced metric
F-02	Content Information corruption	Is a function of time that represents the corruption level of the information package (ingested A/V and metadata contents). In the case that the system has been asked to keep original files the corruption check is made directly on bits and not on A/V quality.	Typically It can only increase, because of bit rot or bit lost, affecting both essence data and metadata. Not all the bytes have the same importance e.g. the representation information is particularly important because in case of lost it could be impossible to exploit the content. In case of corruption it is actually possible in some cases to restore at least partially the audiovisual content using complex algorithms	ME-02 ME-03
F-03	Storage occupation	The actual storage space occupied by the customer (storage allocation - actual remaining space)	It depends on the number and weight of ingested and eventually deleted SIPs	ME-04
F-04	SIP Ingestion time	Is the total elapsed time from the SIP submission to the confirmation from the system that everything has been correctly acquired. This includes: - the time necessary for the upload transfer of the package - the time necessary validate the SIP The ingestion can also be an update of a pre-existing object.	This time could be affected by the total system workload (e.g. the network could be congested), it is reasonable that the ingestion time is proportional to the dimension of the package.	ME-05
F-05	Availability after ingestion (AIP preparation)	Is the elapsed time from the successful validation of ingested SIP to the full availability of the content .This includes the time for transforming the SIP into an internal representation (AIP) and to index the content in order to offer a full access.	This time could be affected by the total system workload (e.g. multiple concurrent operations), it is reasonable that this time is proportional to the dimension of the package.	ME-06
F-06	DIP Delivery time	Is the total elapsed time from the request of a specific SIP (e.g. discovered by a previous search) to the complete and correct reception of the package. This time includes: - the time necessary to extract and prepare materials with a coherent DIP wrapper - if necessary, the time for recovering a corrupted file - the time necessary for the download transfer	The time necessary for preparing the material is highly variable depending on required operations like transcode/transwrap, aggregation of separate content etc. It could be affected by the total system and network workload.	ME-07
F-07	Search time	Is the elapsed time from the query submission to the production of the result list	This time could be affected by the total system workload (e.g. multiple concurrent queries), also depends on the complexity of the query, the total amount of indexed material as well as on the actual implementation of the indexing engine.	ME-08
F-08	Search results quality	Is the goodness of results, usually expressed as recall and precision with respect of the query and the ingested material.	It depends on the type of the query (e.g. exact match should be 100% recall/precision) and the performance of the search algorithms.	ME-09

ID	Name	Description	Behaviour	Referenced metric
F-09	DIP conformance	Is a Boolean value associated to a particular DIP retrieved from the provider, it is false in the case that even just a component (wrapper, essence, metadata) is not valid according to the agreed validation procedure.	DIP in the most general case, is the product of an elaboration performed within the preservation system (e.g. transcoding, trans-wrapping, partial extraction etc.) that could introduce errors. The validation procedure (agreed between parties) can be defined with different level of deepness (e.g. a light validation on wrappers or a deeper one that also checks essence)	ME-10
F-10	Content browsing performance	Is the readiness for the fruition of the browsing quality	This performance could be affected by the total system and network workload	ME-11
F-11	Upload network performance	Is the carrying capacity of data between the customer and the provider averaged on a certain time interval usually measured in Mbits/sec	It depends on the bitrate injected by the customer and the actual availability of the network that usually is shared among several processes and maybe customers. Uploads could be served considering a priority determined automatically (e.g. shortest first) or by explicit request by privileged users.	ME-12
F-12	Download network performance	Is the carrying capacity of data between the provider and the customer averaged on a certain time interval usually measured in Mbits/sec	It depends on the bitrate injected by the provider and the actual availability of the network that usually is shared among several processes and maybe customers Downloads could be served considering a priority determined automatically (e.g. shortest first) or by explicit request by privileged users.	ME-13

Metrics

ID	Name	Description	Unit of measure	How to calculate monitoring data value from measures
ME-01	Availability over time	Given a certain time slot that can be fixed (e.g. a specific month or year) or sliding windows (e.g. the last hour) is the percentage of time where the service was available (up and working correctly) over the total time. The assumed time slot and modalities are explicitly agreed between the parties.	Percentage (calculated over a certain time slot)	On a practical point of view measures are made with a sampling approach, e.g. by testing the availability every minute. The monitoring data values are thus simply calculated for each time slot making the ratio between successful and the total number of calls. Some basic measurements can be done without ad hoc tests but only considering customer calls
ME-02	Bit Integrity	It expresses the fact that the information originally submitted by the customer has been preserved from corruption. One way to state this is to count the number of corrupted bytes over the total amount of storage occupied and the total time of retention.	Probability of byte corruption / GB*year (e.g. 10^ -3 byte/GB*year)	Calculated on fixed periods (e.g. each month or year), on this period sum the number of mistaken and corrupted bytes and divide this number by the mean storage occupation during the period. Could also be decided to calculate over a sliding window e.g. a windows of one month wide calculated each day. E.g. Analysis on January, found 2 corrupted bytes on day 2 and 1 mistaken byte on day 10. Storage follows this progression: 1000 GB from day 1 to 10, 1200 from day 11 to 15, 2000 from 16 to 25 and 1500 from 26 to 31. Mean storage = (1000 * 10 + 1200 * 5 + 2000 * 10 + 1500 * 5) / 31 = 1403 GB Integrity = (2 + 1) / 1403 = 2,14 * 10^-3 bytes/GB*month
ME-03	File Integrity	Another way to state the integrity is to express a percentage of damaged files over the total amount of retained files, in a specific lapse of time.	Probability of file corruption / numfiles*year (e.g. 10^-4 per year that is one file corrupted over 10000 in a year of retention)	Calculate on fixed periods (e.g. each month or year), on this period sum the number corrupted files and divide this number by the mean number of files during the period. E.g. Analysis on 2008, found 2 corrupted files on day 122 and 1 on day 206. Number of stored files follows this progression: 100 files from day 1 to 100, 250 from day 101 to 215, 500 from 216 to 365. Mean number of files= (100 * 100 + 250 * 115 + 500 * 149) / 365 = 310,27 File Integrity = 2 / 310,27 = 6,4 * 10^-3 files per year
ME-04	Storage occupation	Is the ratio between the actually occupied storage and the total amount reserved for that customer	percentage	Absolute storage occupation can be queried directly on the system or a running total could be kept based on ingestions and deletions. A storage occupation history could be saved in order to calculate interesting QoS parameters like the incidence of corrupted bytes along the timeline
ME-05	SIP ingestion time	It is the time necessary for ingesting the submission package, usually the time is normalised against the size of the package because most of the time factors go linearly with the size (copy, transcode, transfer)	Hours or Hours / GB	It has to be measured directly, it can be probably obtained inspecting activity logs
ME-06	Availability after ingestion	Is the elapsed time from the successful validation of ingested SIP to the full availability of the content.	Hours or Hours / GB	It has to be measured directly, it can be probably obtained inspecting activity logs

ID	Name	Description	Unit of measure	How to calculate monitoring data value from measures
ME-07	DIP delivery time	It is the time necessary for delivering the dissemination package, usually the time is normalised against the size of the package because most of the time factors go linearly with the size (copy, transcode, transfer)	Hours or Hours / GB	It has to be measured directly, it can be probably obtained inspecting activity logs
ME-08	Search time	The elapsed time from the submission of the query to the complete answer of the system search engine.	seconds	It has to be measured directly; it can be probably obtained inspecting activity logs. If promises are given as statistical percentages, statistics has to be calculated.
ME-09	Search recall/precision	Recall and precision of the result sets of the search, calculated against all the material stored in the system at the moment of the query.	percentage	It is hard to verify because it needs human intervention with specific checks, it is reasonable to open a feedback channel towards users in order to keep trace of performances on the base of real cases
ME-10	DIP conformance	Given a certain amount of delivered DIPs, is the percentage of conformant packages over the total.	percentage	The native measure directly comes from validation procedures that for each delivered DIP tells if it is good or not. Percentages could be calculated on agreed time slots (e.g. each month) on a certain number of deliveries.
ME-11	Content browsing performance	Is the readiness for the fruition of the browsing quality		A smart method for evaluating should be detected, taking into account delays from request to start of play and periodic interruptions (e.g. by network latency and buffering)
ME-12	Upload Band	Given a certain time slot that can be fixed (e.g. a specific hour) or sliding windows (e.g. the last 5 minutes) is the mean value of data carrying capacity originating from the customer to the provider. The assumed time slot and modalities are explicitly agreed between the parties.	Mbit/s averaged on an agreed interval (e.g. an hour)	Usually operative systems give easy access to band measurement over specific network ports, if necessary they have to be treated to obtain graphs and connections with specific user operations
ME-13	Download Band	Given a certain time slot that can be fixed (e.g. a specific hour) or sliding windows (e.g. the last 5 minutes) is the mean value of data carrying capacity originating from the provider to the customer. The assumed time slot and modalities are explicitly agreed between the parties.	Mbit/s averaged on an agreed interval (e.g. an hour)	Usually operative systems give easy access to band measurement over specific network ports, if necessary they have to be treated to obtain graphs and connections with specific user operations
ME-14	Number of simultaneous users	Is the number of users that simultaneously are logged into the system	positive number	Directly measured
ME-15	Number of simultaneous operations	Is the number of operations (e.g. search, browse, ingest, delivery) that simultaneously are performed by the system	positive number	Directly measured

Quality of Service

ID	Name	Description	Ref metrics	Monitoring criterion (bounds)	Monitoring frequency
QS-01	Availability	The guarantee that the service will be available (up and exploitable) at least as much as agreed. The measure of this quantity can be done as percentage of time (e.g. 99% of the time) or percentage on the number of usage (e.g. 98% of the times one tries, the service is usable)	ME-01	The availability should never go below a specific threshold, there could be more than a threshold (e.g. for business hours and night)	Once per fixed period like month or year OR on a sliding window with an appropriate width and moving with a convenient step
QS-02	Integrity	The guarantee that the ingested A/V and metadata contents has been preserved keeping an agreed quality level (assessed for example with PSNR). These probabilities have to be normalised over the amount of data and the retention time. Additionally the system could have the possibility to keep original formats if required (CAP-20)	ME-02 ME-03	The integrity should never go below a specific threshold	Once per fixed period like month or year OR on a sliding window with an appropriate width and moving with a convenient step
QS-03	SIP Ingestion time	One of the most important parameter perceived by a user when submitting a new SIP (or even for updating) is the total elapsed time from the SIP submission to the confirmation from the system that everything has been correctly acquired. This includes: - the time necessary for the upload transfer of the package (refers to QS-03) - the time necessary to extract, validate, index and transform the SIP into an internal representation (AIP) It is reasonable that the ingestion time will be proportional to the dimension of the package.	ME-05	The SIP ingestion time should never go above a specific threshold, there could be more than a threshold (e.g. for business hours and night).It can be given as percentage, e.g. 90% of deliveries are done under a threshold 1 and the rest under threshold 2.	Every time there is a DIP delivery or periodically if percentage check is assumed on the base of pre-calculated statistics
QS-04	Availability after ingestion	This is the guarantee that ingested packages (SIP) will be searchable, browsable and downloadable (as DIP) after a maximum period of time starting from successful validation.	ME-06	The SIP adaptation to make it available should never go above a specific threshold, there could be more than a threshold. It can be given as percentage, e.g. 90% of adaptations are done under a threshold 1 and the rest under threshold 2.	Every time there is an ingestion or periodically if percentage check is assumed on the base of pre-calculated statistics.

ID	Name	Description	Ref metrics	Monitoring criterion (bounds)	Monitoring frequency
QS-05	DIP Delivery time	One of the most important parameter perceived by a user when asking for some material (media + metadata packaged in a DIP) is the total elapsed time from the request to the complete and correct reception of the package. This time includes: - the time necessary to extract and prepare materials with a coherent DIP wrapper - if necessary, the time for recovering a corrupted file - the time necessary for the download transfer (refer to QS-04) The time necessary for preparing the material is highly variable depending on required operations like transcode/transwrap, aggregation of separate content etc.	ME-07	The delivery time should never go above a specific threshold, there could be more than a threshold (e.g. for business hours and night). It can be given as percentage, e.g. 90% of deliveries are done under a threshold 1 and the rest under threshold 2. Because the delivery time depends not only on dimensions but also on the kind of operations required, threshold should be given taking this into account (e.g. if transcode necessary threshold1, if not threshold2 etc.)	Every time there is a DIP delivery or periodically if percentage check is assumed on the base of pre-calculated statistics
QS-06	Search time	Search results should be obtained in a reasonable amount of time depending on the complexity of the query and on the amount of the indexed material	ME-08	One of these ways: *The search time should never go below a specific threshold *In terms of percentage of invocations and different thresholds (e.g. 90% of times time is under threshold1, the remaining under threshold2)	Every time there is a search activity or periodically if percentage check is assumed on the base of pre-calculated statistics
QS-07	Search results performance	The system has to guarantee a high level of recall and precision depending on query type (e.g. query by id or exact match should be 100% while for similarity search - if available - above a certain threshold)	ME-09	Precision and recall should be above a certain threshold that depends on the kind of search	Every time there is a search activity or periodically if percentage check is assumed on the base of pre-calculated statistics
QS-08	DIP conformance	The requested and delivered material has to be valid for all its components: wrapper, essence and metadata.	ME-10	No corrupted packages	Whenever a package is delivered
QS-09	Content browsing performance	This is the guarantee that the service of content browsing (preview) will be working well in terms of readiness as agreed	ME-11	The "smart" metrics used to objectively evaluate the performance should be above a certain threshold	Whenever the browsing service is used (typically after queries)
QS-10	Upload network performance	During ingestion operations like entire new SIPS upload or just SIPS updates where additions or modifications of some components (whether multimedia or textual) are performed, the network is used and it is essential that transfer times would be reasonable. The urgency of these operations depends on several aspects so that is advisable to have a priority for uploads, determined automatically (e.g. shortest first) or by explicit request by privileged users.	ME-12	Two possible modality of agreement: * The performance should never go below a specific threshold, there could be more than a threshold (e.g. for business hours and night). * More likely it could established in term of percentage over time e.g. for 90% of the time, performance will be no less than threshold1 and for the rest not less than threshold2	Once per fixed period like each minute OR on a sliding window moving with a convenient step

ID	Name	Description	Ref metrics	Monitoring criterion (bounds)	Monitoring frequency
QS-11	Download network performance	During access operations like search and DIPS download , the network is heavily used and it is essential that transfer times would be reasonable. The urgency of these operations depends on several aspects so that is advisable to have a priority for downloads, determined automatically (e.g. shortest first) or by explicit request by privileged users.	ME-13	Two possible modality of agreement: * The performance should never go below a specific threshold, there could be more than a threshold (e.g. for business hours and night). * More likely it could established in term of percentage over time e.g. for 90% of the time, performance will be no less than threshold1 and for the rest not less than threshold2	Once per fixed period like each minute OR on a sliding window moving with a convenient step
QS-12	Restore time in case of disasters	In case of natural disasters like fire, flood the provider could have the ability to restore information from a remote backup copy or system	NA	Restore time has to be under a certain threshold that could depend on the total amount of stored material/metadata.	In case of disaster only.

Constraints

ID	Name	Description	Measure unit	Ref metrics	Monitoring criterion (bounds)	Monitoring frequency
C-01	Authorised formats (wrapper, essence and metadata)	Provider and customer agree on a set of predefined formats accepted for essence and metadata e.g. MXFD10 and DV50 for essence and Mets as xml for metadata. Adhering to OAIS model also a proper definition of SIP and its validation has to be established.	NA	NA	A package that is not in the form of an authorised format should be discovered by the validation phase that has to be applied for each complete or partial ingestion.	At every package ingestion or update, otherwise adopt a sample criteria
C-02	Maximum amount of storage	The maximum amount allowed to a specific customer by contract	GBytes	ME-04	The occupied storage should never cross a specific threshold, partial exceeding for a limited period of time could be tolerated	Periodically check (e.g. once per hour) that the storage occupation of the customer is below the agreed maximum. If a mechanism like quotas is used the check is automatic but should be desirable to advise the customer when he is reaching the limit (e.g. by e-mail)
C-03	Maximum number of simultaneous users	Maximum number of users logged in at the same time	positive integer	ME-14	The actual number of logged users should never cross a specific threshold, partial exceeding for a limited period of time could be tolerated	Check when users ask for login, in case deny if the maximum has been reached
C-04	Maximum number of simultaneous operations	Maximum number of simultaneous operations, e.g. No more than 3 simultaneous ingestions or 2 simultaneous transcoding	positive integer	ME-15	The actual number of simultaneous operations should never cross a specific threshold, partial exceeding for a limited period of time could be tolerated	Check when users ask for operations, in case deny if the maximum has been reached

Pricing Terms

ID	Name	Description	Units	References	Possible ways to calculate
PRC-01	Fixed price	This is the fixed component of price, it is determined by service characteristics, promised quality, constraints, variable cost model and market competition. Usually is in form of annual fee. Example: The service provides a storage capacity of 500 TB for a period of 5 years with the possibility to vary this amount ongoing. The maximum number of users is 50 (30 simultaneous). The service is 24h and Availability of each service (ingest, search etc.) is guaranteed for 99,9 % of time, integrity is assured with a byte corruption probability of 5*10-3 bytes/GB*year. Additional capabilities include upload-download resume, transcoding, demux.	Euros / year	All constraints All QoS guarantees	Agreed at contract signature, eventually revised periodically (e.g. yearly or when some conditions change)
PRC-02	User charge	Usually included in fixed price with a maximum number of licensed and contemporary users. Some cost model can foresee a variable charge depending on number of licensed users or their connection time (login logout)	Euros / user Euros / hour	ME-14	A fixed cost for hour or a nonlinear cost with discount over a certain amount
PRC-03	Hit charge	Assign a charge for each specific service invocation e.g. 0.2 euro each ingestion, 0.5 euro for each export	Euros / service usage		A fixed cost for single invocation or a nonlinear cost with discount over a certain amount. Example from Amazon/S3: $0.01 per 1,000 PUT, COPY, POST, or LIST requests $0.01 per 10,000 GET and all other requests
PRC-04	Data movement charge	A charge for uploaded and dowloaded quantities (e.g. 0.1 euro for each uploaded GB)	Euros / GB	ME-04	The cost for ingest could be less than that for upload, because uploads foster the usage of the system with the paradigm of one upload multiple search/download. Example from Amazon/S3: $0.170 per GB – first 10 TB / month data transfer out $0.130 per GB – next 40 TB / month data transfer out $0.110 per GB – next 100 TB / month data transfer out
PRC-05	Storage usage	Even if in the fixed cost takes into account the maximum storage made available, the cost model could also consider the effective storage usage	Euros / GB for month of storage usage	ME-04	A non-linear approach can be used like Amzon/S3: $0.100 per GB – data transfer out / month over 150 TB $0.150 per GB – first 50 TB / month of storage used $0.140 per GB – next 50 TB / month of storage used $0.130 per GB – next 400 TB /month of storage used $0.120 per GB – storage used / month over 500 TB
PRC-06	CPU usage	Some kind of operations like transcoding (e.g. for migration) are CPU intensive and it could make sense to have a variable price component based on it	Euros / (CPU * month)	ME-15	E.g. in a month has been fully used on average 2.3 CPUs

Penalties

ID	Name	Description	References	Possible ways to calculate
PTY-01	Broken contract	The customer decide to stop the relationship with the given provider prior of the natural expire date of the contract		Penalty amount could depend on the date of the breaking with relation of the contract duration.
PTY-02	Lack of availability	The provider does not respect the promise with respect of availability and gives a refund for this reason	ME-01	Whenever availability guarantee is not respected pay a fixed amount of Euros
PTY-03	Data corruption	The provider does not respect the promise with respect of integrity guarantee and gives a refund for this reason	ME-02	Whenever integrity guarantee is not respected pay a fixed amount of Euros
PTY-04	Lack of band	The provider does not respect the promise with respect of upload and/or download performances	ME-05 ME-07 ME-11	Whenever band guarantees are not respected pay a fixed amount of Euros
PTY-05	Poor Search	The provider does not respect the promise with respect of search performances or recall/precision	ME-08 ME-09	Whenever search guarantees are not respected pay a fixed amount of Euros
PTY-06	Maximum storage exceeded	The customer continues to upload material even if the maximum agreed space has already been used	ME-04	If this happens the provider could decide to allow anyway further upload but asking for a certain amount of money with the possibility to contract for an enlargement of the available space
PTY-07	Invalid SIP submitted	The submission package has some problems, e.g. some xml not well formed or invalid or unsupported or corrupted formats for essence	ME-10	If the number of submitted and corrupted packages goes beyond a certain number pay a fixed charge or pay a certain amount for every invalid SIP submitted

12. References

[1] PrestoPRIME D2.3.1, Service-Oriented Models for Audiovisual Content Storage, Feb 2010, http://www.prestoprime.eu

[2] PrestoPRIME D2.2.1, Preservation Process Modelling, Feb 2010, http://www.prestoprime.eu

[3] Recommendation for a Producer-Archive Interface Methodology Abstract Standard, CCSDS 651.0-B-1, Blue Book, May 2004, http://public.ccsds.org/publications/archive/651x0b1.pdf

[4] PrestoPRIME D2.2.1: Preservation Process Modelling, Feb 2010, http://www.prestoprime.eu

[5] http://www.dcc.ac.uk/tools/trustworthy-repositories/

[6] Report on Trusted Digital Repositories, RLG, http://www.rlg.org/en/page.php?Page_ID=20769

[7] How does one know which repository is worth its salt? David Giaretta, http://www.ais.up.ac.za/digi/docs/giaretta2_paper.pdf

[8] TRAC, http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

[9] DRAMBORA, http://www.repositoryaudit.eu/

[10] Nestor, http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf

[11] International Standards Organisation, http://www.iso.org

[12] OAIS Blue Book, CCSDS 650.0-B-1, http://public.ccsds.org/publications/archive/650x0b1.pdf

[13] http://wiki.digitalrepositoryauditandcertification.org/bin/view and http://www.ais.up.ac.za/digi/docs/giaretta2_paper.pdf

[14] Information Technology Infrastructure Library (ITIL), http://www.itil-officialsite.com

[15] Control Objectives for Information and Related Technology (COBIT), http://www.isaca.org

[16] Statement on Auditing Standards number 70 (SAS 70), http://www.sas70.com

[17] Nirvanix, http://www.nirvanix.com/

[18] SERSCIS, EU ICT FP7 project reference 225336, http://www.serscis.eu

[19] Edutain@Grid, EU IST FP6 project reference 034601, http://www.edutaingrid.eu

[20] GRIA, http://www.gria.org

[21] PrestoPRIME D2.1.1, Audiovisual preservation strategies, data models and value chains, Feb 2010, http://www.prestoprime.eu

[22] SIMDAT, EU IST FP6 project reference 511438, http://www.simdat.org

[23] IRMOS, EU IST FP7 project reference 214777, http://www.irmosproject.eu

[24] HQML: Xiaohui Gu et al. “An XML-based Quality of Service Enabling language for the Web”, Journal of Visual Language and Computing, Special Issue on Multimedia Language for the Web, 2002.

[25] WSLA, http://www.research.ibm.com/wsla/

[26] SNAng: D.D.Lamanna, J. Skene and W., Emmerich, “SLAng: A Language for Defining Service Level Agreements”, Proceedings of the The Ninth IEEE Workshop on Future Trends of Distributed Computing Systems (FTDCS’03).

[27] QML, M. A. de Miguel, "QoS Modeling Language for High Quality Systems," Eighth IEEE International Workshop on Object-Oriented Real-Time Dependable Systems (WORDS'03), 2003

[28] WSOL, V. Tosic, B. Pagurek, K. Patel, “WSOL – A Language for the Formal Specification of Various Constraints and Classes of Service for Web Services ”, The International Conference On Web Services, ICWS’03

[29] WS-Agreement, D.Garcia, M. Toledo, “Semantic-enriched QoS Policies for Web Service Interactions”, Web Media ’06, Natal, Brazil.

[30] WS-Policy, http://www.w3.org/TR/ws-policy/

[31] Apache Neethi, http://ws.apache.org/commons/neethi/

[32] G. Dobson. “Quality of Service in Service-Oriented Architecture”, http://digs.sourceforge.net/papers/qos.pdf

[33] OASIS WSDM, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsdm

[34] WS-Management, http://www.dmtf.org/standards/wsman

[35] UncertML, http://www.uncertml.org/documents/UncertML.pdf

[36] UCUM, http://unitsofmeasure.org/