Analysis of Functional User Requirements and System Requirements in the context of Long Term Data Preservation (LTDP)

Rubén F. Pérez [(1)], Luis G. Gutiérrez [(1)], Óscar Pérez [(1)], Raffaele Guarino [(2)], Rosemarie Leone [(3)], Mirko Albani [(3)]

[(1)] GMV
Isaac Newton 11, 28760, Madrid, Spain
EMail: rfperez@gmv.com, lgutierrez@gmv.com

[(2)] CAPGEMINI
Via di Torre Spaccata 140, 00173, Roma, Italy
EMail: raffaele.guarino@capgemini.com

[(3)] ESA-ESRIN
Via Galileo Galilei 64, 00044, Frascati, Italy
EMail: Rosemarie.Leone@esa.int, Mirko.Albani@esa.int

ABSTRACT

The Long Term Data Preservation (LTDP) functional user requirements and system requirements are presented in this paper, showing the most appropriate architectures in relation to main scenarios as identified by ESA FIRST and LAST activities.

According to classical engineering process, functional user requirements concerned with long term data preservation of scientific data (Earth Science) are captured and analysed, finally deriving in the definition of the System Requirements and consequently the most appropriate system designs and architectures. In this context, the users of the system have different roles, mainly classified in consumers and data holders of EO archives.

Initially, the functional user requirements involving consumers were identified in the context of FIRST activity, including an analysis of business needs requiring and justifying long term preservation and performing an analysis of typical scientific mission phases, along with a preliminary contribution to the identification of the possible content to be preserved.

During the LAST activity, a further analysis of user requirements involving data holders was performed by means of a survey of current LTA Systems of reference worldwide, the analysis of LTDP common guidelines and other standards (i.e. OAIS, ISO) deriving finally in a set of fundamental use cases which gave rise to the elucidation of the LTDP common system requirements. This set of requirements can be considered as a system of reference related to the state of the art in 2010-2011, allowing data holders to evaluate the most appropriate technologies and architectures in relation to LTDP guidelines and their specific constraints and preferences.

Keywords: LTDP, User Requirements, System Requirements, Architectures, Long Term Archiving, Earth Observation, Evaluation, Assessment

INTRODUCTION

The main objective of the ESA's proposed Long Term Data Preservation (LTDP) initiative is to guarantee the preservation of the data from all Earth Observation (EO) ESA and Third Parties ESA managed missions on the long term, also ensuring their accessibility and usability, as part of a joint and cooperative approach in Europe aimed at preserving the EO European data from member states' missions. The need to ensure the preservation of the Earth Observation data has been expressed by practically all environmental monitoring programmes and recently again through the Climate Change Initiative. Following

consultations with space Agencies and workshops with the owners and holders of other Earth Observation data archives, ESA member states, as part of ESA's mandatory activities, approved a three year initial programme with the aim to establish a full long term data preservation concept, and a later programme beyond 2011 [1-3]. ESA started the set-up of a cooperation framework with other European space agencies and EO satellite operators to address LTDP issues from a technical point of view and to pursue a stronger coordination at European level. Over the last years, a set of European LTDP common guidelines have been defined. These initial guidelines are being consolidated and promoted within the Committee on Earth Observation Satellites (CEOS) [4] and Group on Earth Observation (GEO) [5], and constitute the basis for the ESA's EO data preservation approach and for the further cooperation with other European EO data archive holders.

LTDP/FIRST project was started in June 2010. The project was aimed at understand user requirements, possible evolution of former preserved data-set composition and impacts on guidelines was completed in July 2011. During the project cycle a significant workshop was held on October 2010 at ESRIN. All the material produced during the project is available on ESA LTDP website.

LTDP/LAST project (Long term data Archive Study on new Technologies) was started by ESA in order to perform an independent assessment on the best practices and the many different archiving technologies for archive management and operation in the short and mid-term time frame, or available in the long-term, suited to satisfy the requirements of ESA Earth Observation Space data digital information preservation. This activity has performed a wide analysis of the technological areas involved in archiving, identifying the different technological areas and the system requirements related to each one. Additionally, an evaluation model has been defined in each technological area in order to evaluate the last technological products in the most relevant ones. Finally, after identifying the most important elements of archiving systems, an analysis of the different architectures that suit with the LTDP guidelines was also provided.

USER REQUIREMENTS

The user requirements have been classified into data consumers, who represent the users making access to the Long Term Archive (LTA) system, and data holders, who represent the users that produce and administer the system providing data to the consumers as demanded. The objective of specifying the user requirements of both types of users involved in archiving was to define a set of common system requirements to identify and evaluate the suitable architectures and technologies available in the market.

Data Consumers

Capture and analysis of user requirements concerned with scientific data series was aimed at understand what it is necessary in terms of data and information to the scientific communities in order to achieve their business objectives. This capture and understanding was focused on following aspects: what, why, together with what else, for how long time.

The first and most important issue was the identification of user communities having some kind of interest in long term data series as part of their scientific analysis of phenomena.

Communities of Interests (CoI), i.e. scientific users, for such an analysis were all scientists and investigators having some kind of direct or indirect interest in Earth Science (ES) domains and using series of data of any kind (e.g. satellite based remote sensed data, ground based sensors, water related data, etc.).

User requirements concerned have been captured analyzing needs of scientific domains through publications , reviewing similar studies done in the recent past , through the analysis the of activities performed by users, through questionnaires, and having interviews with scientists and investigators.

Participation to international scientific events and contacts was used to capture and understand even more.

Almost all Earth Science domains have been contacted and a great amount of sent questionnaires have been returned with all questions answered (Figure 1).

Captured user requirements have been refined through cross-verifications and cross-questions eliciting duplications and overlaps, doing a reverse simulation, and by interviews using a complementary approach (e.g. possible alternatives, complementary/opposite, if-not method).

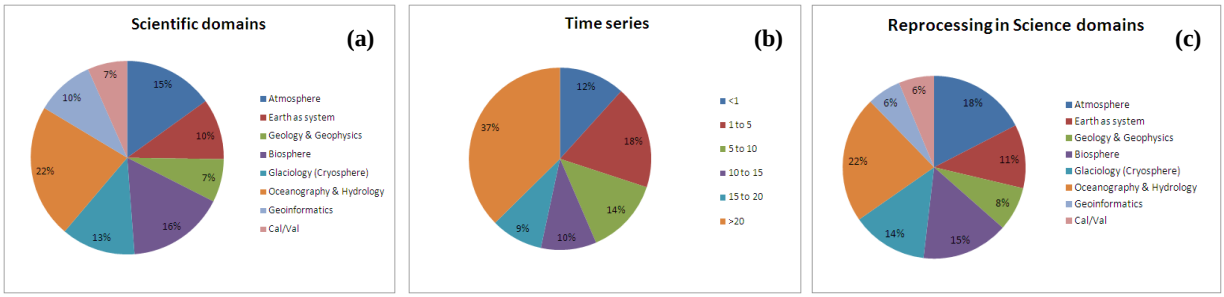

Figure 1: Distribution of information and answers from different scientific domains (a), Expected/Used times series (b), Relative amount of reprocessing (c)

The user requirement understanding was carried out for the necessity to extend data categories and classifications used in the first version of the European Common Guidelines Long Term Preservation of Earth Observation Space data [1].

Data set have been extended to nine categories embracing from SAR data (C1) to water related sensors/measures (C9). These categories cover from satellite based and remote sensed data to airborne based sensors, and in-situ sensors/measures both on ground and water.

Another result of this first step is an effective list of what is necessary to preserve, why, for how long time and what's the additional information to be taken into account in any preservation process.

One of the outcomes is the updating and extension of "preserved data set composition" with specific information and attributes. The "preserved data set" composition defines a consistent and complete set of data enabling current and possible future utilization.

It has been strengthened the necessity to add some more attributes to the preserved data set composition adding information concerned with the context of capture or transformation, provenance of the dataset, stability of the information contained and in general with attributes supporting relevant trustability and reliability.

Special importance and significance of user requirement is concerned with maintenance of knowledge and experience acquired along the development of each mission or campaign aimed at capture scientific data. Pertained point-of-view is focusing to allow future processing and/or reprocessing of today scientific missions and of the past ones where feasible.

The general expectation of the scientific community remains to preserve everything is today captured or generated and to uphold forever being impossible deciding today what will be of interest in any future research.

A synthesis of user requirements can be summarized as follows:

a) Access to data and information including documentation for scientific purposes should be free and open.

b) The owners or providers of Earth Science data and information including documentation should guarantee their preservation without limits (all forever).

c) Access mechanism to data should be simple, easily available, easily deployable, and economical for the user.

d) Data, products and information should be made available on request at any time.

Data Holders

During the first stage of LAST, a survey was performed by distributing a questionnaire to ESA and a number of EO partners, plus selected parties from outside of the EO domain.

The total number of organizations contacted all around the world (i.e. about 54) made any individualized scheme impractical from the point of view of the costs, resources, and of the amount of time required to complete the task. Therefore, a questionnaire was circulated among all parties to gain a broad perspective over the different issues of interest, their approaches, and the overall situation within the field.

Since the archival and management of vast amounts of data for (very) long periods of time is not a challenge exclusive of the Earth Observation domain, a number of organizations from other different fields, where similar Long Term Data Preservation issues are being addressed, were also contacted (to a lesser extent). These fields included, among others:

- Big science (e.g. astronomy, high energy physics)
- Supercomputing centres
- Digital libraries and repositories
- Online storage and services

In the end, from the approximately fifty-four (54) organizations contacted, twenty (20) replied the questionnaire and/or were interviewed.

Among the results of the survey, it was commonly highlighted the need for managing increasing volume and access, with the concern for flexible, robust, and less expensive solutions (Figure 2). Note that the volume Y in axis (Figure 2c) is in logarithm scale, which means that the growth is exponential. The reprocessing campaigns are also a key issue for the holders. The holders transmitted their concerns about the feasibility of future campaigns of reprocessing, in case the data is not accessible at campaign time. Past lessons show that data in obsolete media might not be accessed appropriately, making difficult the automatic reprocessing campaigns.

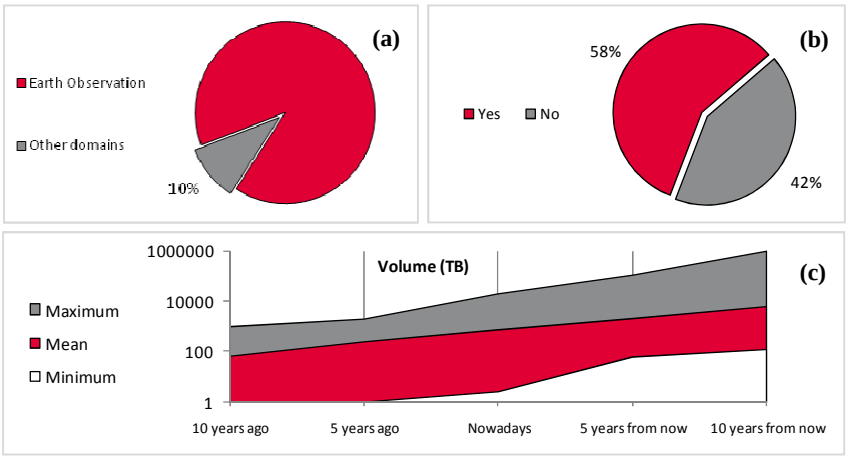

Figure 2: Data holders survey results for domains interviewed (a), data holder performs reprocessing campaigns (b), and growth trends (c)

A set of use cases were extracted initially from this survey, compiling the current status of the archiving along with the needs and wish-list provided by the data holders.

SUMMARY OF USE CASES INVOLVED

In addition to the use cases extracted from the technological survey with data holders, other use cases were extracted and harmonized according to the current standards of relevance in Earth Observation (Table 1).

Identifier	Description
ESA LTDP	Long Term Preservation of Earth Observation Space Data: European LTDP Common Guidelines
ISO 14721	Open archival information system (OAIS) - Reference model
ISO 15489	Information and documentation - Records management
ISO 19115	Geographic information - Metadata
ISO 19119	Geographic information - Services

Table 1: Summary of main standards analysed

The context and overall structure of a long term archive are well represented by the OAIS reference model [6], whose high-level functional decomposition is depicted in Figure 3. The following actors interact with the archive:

- Management: the role played by those who set overall archive policy as one component in a broader policy domain. Management is not involved in day-to-day archive operations, since this responsibility is included by OAIS in an administrative functional entity.
- Producer: the role played by those persons or client systems, which provide the information to be preserved. The producer and manager are usually addressed by the same EO data holder, although an external data holder can also make this.
- Consumer: the role played by those persons or client systems, which interact with archive services to find and acquire preserved information of interest. A special class of consumers is the designated community. The designated community is the set of consumers who should be able to understand the preserved information.

The reception of the data provided by a producer is represented by Submission Information Packages (SIP); its storage and preservation as Archival Information Packages (AIP), which include the Preservation Description Information (PDI) necessary for their adequate conservation and maintenance, and the Representation Information needed to translate the Data Objects into understandable forms; and the access to this data by the consumers and members of the designated community, to whom the applicable Data Objects will be provided as Dissemination Information Packages (DIP), delivered in a set of media or through telecommunications means. Furthermore, the access and ordering aids used by the consumers to find, order and retrieve data is supported by Descriptive Information, managed as well within the archive in e.g. a database.

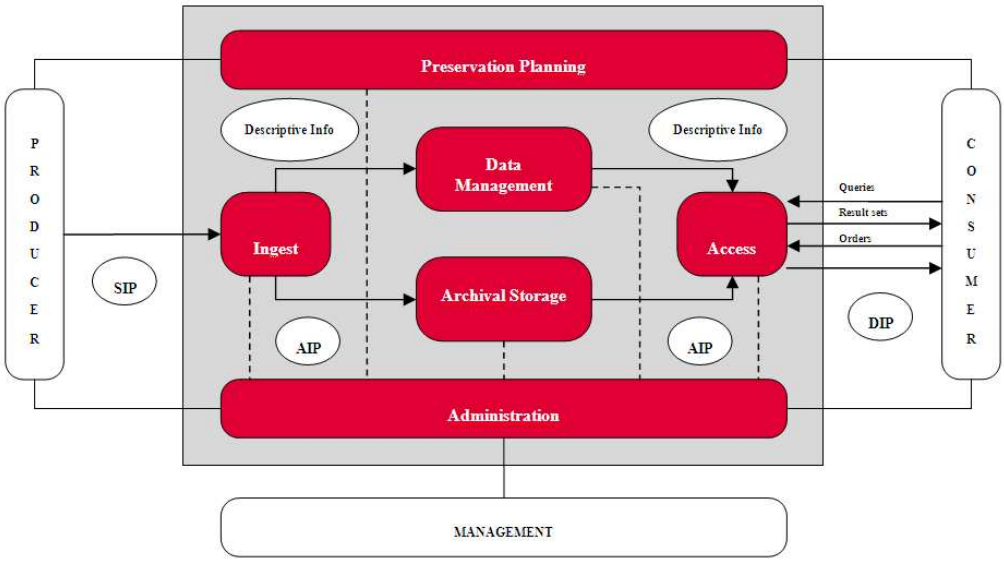

Figure 3: OAIS Functional Model.

The Use Case Model depicted in Figure 4 was obtained reflecting the main interactions behind the OAIS functional and establishing the system boundaries around the aforementioned entities which together constitute the archive as a whole.

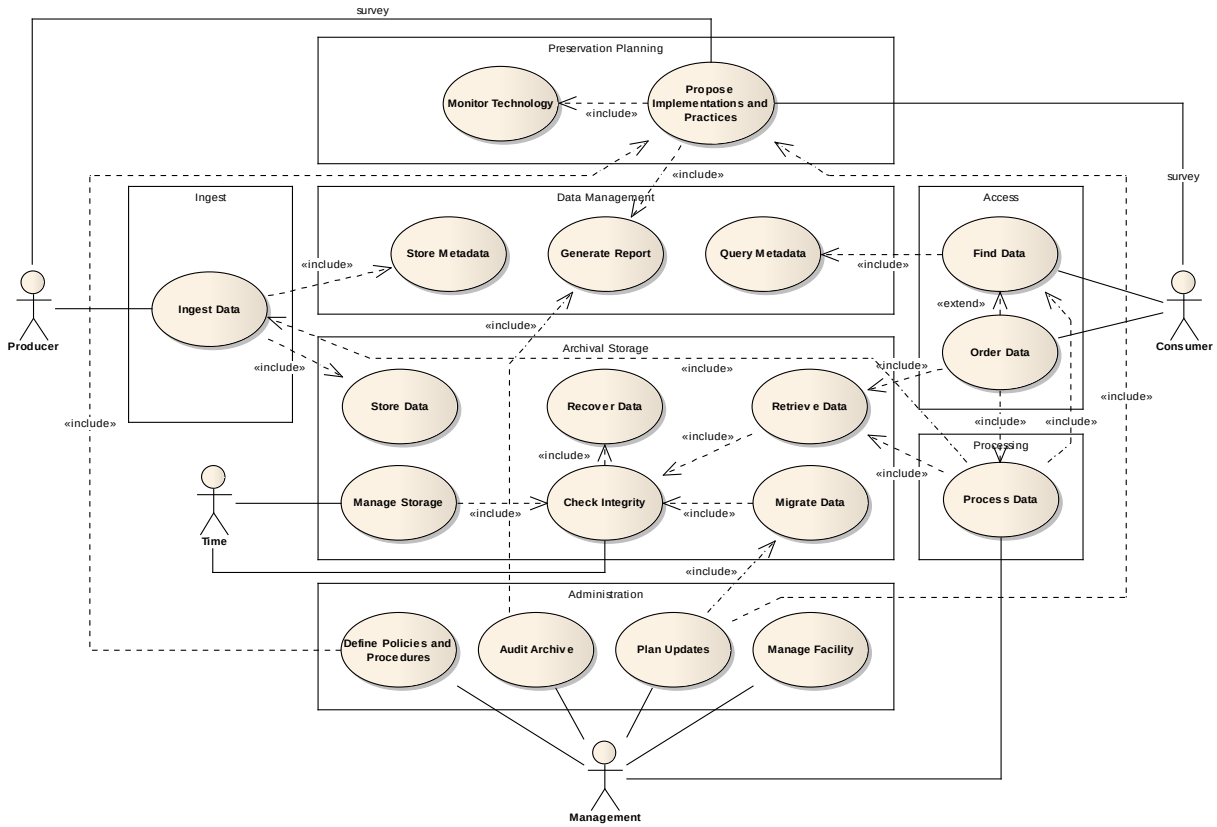

Figure 4: General Use Case Model

A further analysis revealed a variety of use cases that can be summarized according to the following categories:

- **Ingest Data**: One or more Submission Information Packages (SIP) from a producer, or a processing system internal to the archive are received. The integrity of these SIP is verified and, once enough of them are available (i.e. for the composition of the subsequent Content Data Objects to be archived), one or more Archival Information Packages (AIP) are created and stored, registering the associated metadata as appropriate.
- **Store Data**: One or more Archival Information Packages (AIP) are received. These are copied into permanent storage, along with other items (e.g. any accompanying ancillary and auxiliary data). In addition, AIP are replicated in different media according to the archive implementation and preservation policies.
- **Store Metadata**: One or more Descriptive Information packages associated to a similar amount of Archival Information Packages (AIP) are received. These are registered into persistent storage, along with any accompanying Preservation Description Information (PDI).
- **Define policies and procedures**: A request for the establishment and/or update of the standards, policies and procedures to be followed by the archive systems and operations is received from the management of the archive. A set of proposals on the best applicable practices is elaborated and validated.
- **Order Data**: A query about the data available in the archive is received from a consumer. The appropriate Descriptive Information is accessed and a result set (including any access and delivery

costs associated) is presented, allowing the corresponding Archival Information Packages (AIP) to be ordered. Those AIP selected are retrieved and, after generating the respective Dissemination Information Packages (DIP), the latter are sent to the consumer. Note that, depending on the actual archive implementation, some or all of the AIP may not be stored as such, but generated on demand from the stored information. If so, the order is satisfied by triggering the appropriate processing.

- **Proposed implementations and practices**: A request for proposals on archive practices, designs and implementations is received. A survey on the applicable media and archiving technologies is carried out to determine the available options and the state of the art in the appropriate subjects, plus any standards, conventions and guidelines of interest. The archive producers, consumers and designated community are polled, identifying their requirements, needs, approaches and any related trends. The best practices for the archive implementation and operations are then elucidated, taking into account the aforementioned factors and details, and a number of statistics from the archive itself. All this information is formalized in one or multiple reports, and returned to the requester.

SYSTEM REQUIREMENTS

From the standards analysis and use cases previously described, a set of requirements were extracted from the survey and the resulting Use Case Model. The set of requirements were classified according to the following main topics [3]:

- **Standardization:** Relation of the archive system with any applicable standards, e.g. European LTDP Common Guidelines [1], the OAIS reference model [6], International Standards for Information and documentation [7], and Geographic Information for Metadata and Services [8].
- **Reliability**: Oriented to the appropriate backup and redundancy systems, type of access to the data (i.e. on-line, near-line, off-line) and other factors that contribute to assure the quality of system services.
- **Maintenance**: Oriented to define the most appropriate maintenance practices and conditions (e.g. building safety, protection against electrical disruptions, hardware maintenance etc). Since some systems may need to support legacy technologies for extended periods of time, it may prevent them to adopt new systems.
- **Migration**: Oriented to the periodic migration of data to new media and the mechanisms that shall be involved (e.g. integrity, process automation etc).
- **Interface**: Oriented mainly to the use of standard interfaces in all the services of the archive, way of handling of nominal user requests, and types of access mechanisms.
- **Performance**: Aimed at meeting the required level of service, taking into account the maximum number of simultaneous requests and related parameters.
- **Security**: Oriented to monitor, control, and restrict access to archive data, which should only be granted to authorized personnel and users for the different operations, arranging firewalls appropriately to improve network security.
- **Operations**: Providing a number of guidelines for the management of the archive and, in particular, its operations, policies and procedures.
- **Procurement**: Addressing the selection of new technologies and media, and the associated vendors, to guarantee the long-term continuity of the archive, including tests of new systems and technologies, type of software used for the archive operations, and some preferred hardware implementations.

The resulting set of requirements was used for the elaboration of a trade-off of the last technologies available in the archiving domain by means of the definition of evaluation models (Figure 5). A further description of the technique applied can be found in [3].

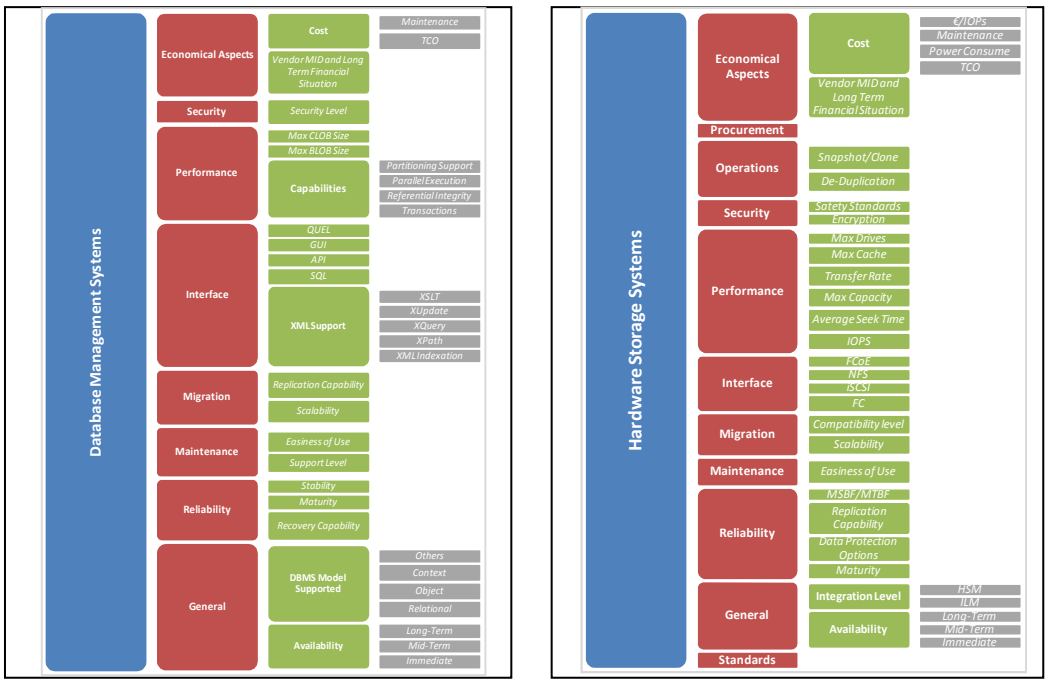

Figure 5: Evaluation models for Database Management and Hardware Storage Systems [3]

ARCHITECTURES ASSESSMENT

Since LTDP guidelines are described in terms of policies, allow to make use of several types of architectures. However, the cost (not only economical, but in terms of resources and expertise) can differ from one to other approach. From the point of view of system requirements obtained in this study, the specific use of the different architectures in the implementation of LTA systems may also involve a different cost in terms of security, operations, etc. Therefore, the architectures broadly used in relation to archive have been also analysed in relation to the LTDP guidelines and System Requirements. This assessment has been driven by two inputs:

- LTDP common guidelines: This allowed the identification of the difficulties that may appear in each type of architecture in relation to the LTDP guidelines.
- System Requirements extracted within the study: This allowed the identification of the architectures that may better suit the system requirements and have helped to evaluate the last technologies in relation to archiving systems.

It shall be noticed that all the potential risks found in this analysis have proved to be solved with the appropriate resources and expertise. However it involves a cost that may be difficult to assume with an important growth of the archive. In relation to the performance, it may be clearly limited if the architecture is not the most appropriate. This assessment identifies potential risks, limitations, disadvantages and advantages derived from the potential application of the themes classified by LTDP guidelines and technological aspects driven by the system requirements.

The following families of architectures were analysed:

- **Storage Architectures**: Server-Centric and Storage-Centric architectures.
- **Basic Distributed Architectures**: Client-Server (C/S) and Peer-to-Peer (P2P) architectures.
- **Integration Architectures**, comprising the architectures that allow the integration of the access to data archives with the business logic and the delivery to the consumers: Multi-Tier, Service Oriented Architectures (SOA) and Cloud architectures.

Storage Architectures

In the case of Storage Architectures, after comparing the results of Server-Centric and Storage-Centric architectures (Figure 6), the Storage-Centric is clearly more appropriate for the implementation of LTDP

guidelines. The main reason is the flexibility and scalability of this architecture, allowing delegate some capabilities (i.e. duplication, replication etc) directly on the technologies. On the other hand, the cost of such architecture is higher in some aspects, since it shall be ready to interact with more servers simultaneously involving an additional effort in security and infrastructure safety and security. In spite of this result, it does not mean that Server-Centric architectures cannot be used for long term archiving. The meaning of the results obtained is that the implementation of such architecture would involve more costs in the main aspects of the archive (e.g. Effort of development, Maintenance Costs, and Capability of growth) with a lack of scalability and flexibility that could be ignored in case the system has not a considerable growth or change of the requirements [9].

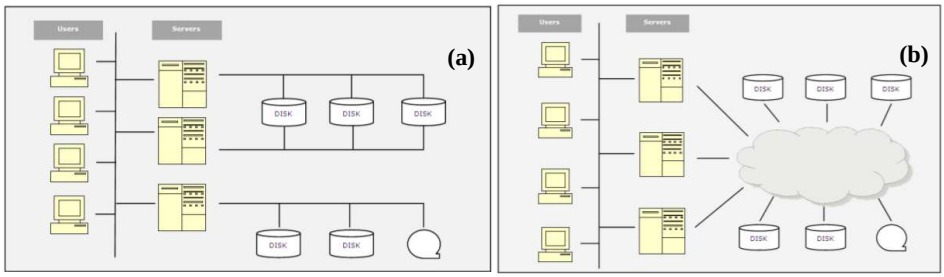

Figure 6: Server-Centric architecture (a) versus Storage-Centric (b) architecture

Basic Distributed Architectures

The results of the evaluation for basic distributed architectures revealed that Security and Reliability are the main weaknesses of P2P architecture. On the other hand, in the case of C/S architecture it is assumed that a control of the dimension of the system and appropriated failover mechanisms has to be performed involving an increase of the cost for the implementation. In this case, the main weakness is the maintenance and the cost of the implementation. It shall be noticed that the costs of maintenance and monitoring of the system are higher in the case of P2P, as all the nodes of the network must be secured and controlled. In the case of C/S only the server node must be secured and the servers correctly dimensioned. Although C/S is more extended at this moment with more tools available in the market, P2P is gaining an important number of users mainly in relation to streaming services. If the P2P is preferred (e.g. given the advantages in the scalability of the system), it shall be considered a hybrid implementation in order to avoid the main disadvantages of security and reliability in all the nodes of the system.

Integration Architectures

The Integration Architectures have revealed a good compatibility of SOA with the rest of architectures (i.e. Cloud, N-Tier, Basic Distribution Architectures, and Storage Architectures). It can be combined with other integration architectures either by means of a C/S (e.g. middleware solution) or P2P. However, since P2P is less extended, there are fewer tools available in the market for such implementation. Note that the potential risks are always minor than the potential advantages (Figure 7).

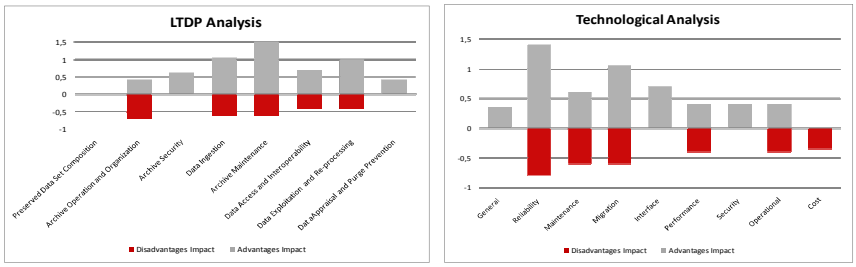

Figure 7: Summary of potential risks, limitations and advantages of SOA architecture

The expertise and the resources to manage the system normally allow taking the most from this architecture with the corresponding cost and skipping the highlighted risks. In relation to the N-Tier architectures the main weaknesses are related to the performance, ingestion capability, and interoperability when the archive tends to grow (Figure 8). In the case of Cloud, the public Cloud has been discarded, given the low security and confidentiality of the solution. However, it shall be noticed

that this could be perfectly used when the information is public and the replication in a different public Cloud is guaranteed to avoid data loss or possible trouble derived from the Cloud vendor business.

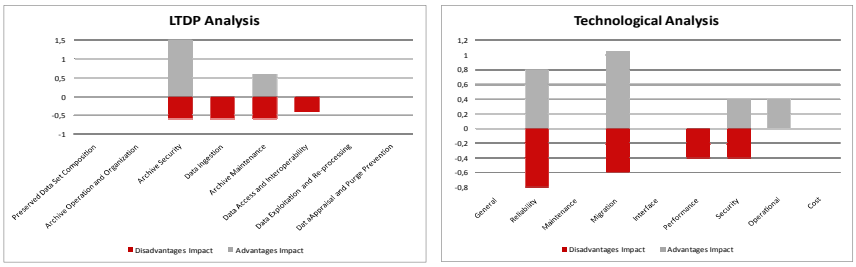

Figure 8: Summary of potential risks, limitations and advantages of N-Tier architecture

The private Cloud avoids the lack of confidentiality and security, but with a necessary increase of the cost (Figure 9). As in the case of SOA, the different risks can be mitigated with the resources and expertise which implies an impact in the cost.

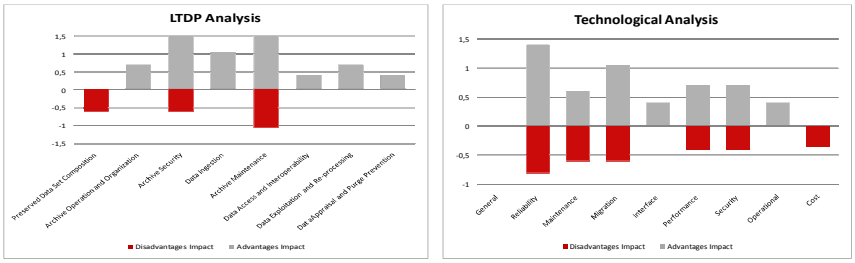

Figure 9: Summary of potential risks, limitations and advantages of Cloud architecture

CONCLUSION

LTDP activities coordinated by ESA (FIRST and LAST) have contributed to the consolidation of a rigorous basis of user requirements and use cases that have led to the definition of a set of system requirements for LTA systems [10]-[11]. This set of system requirements are related to the management of technologies and can be considered as a bridge between LTDP guidelines and the technologies available in the market.

FIRST activity has contributed to identify user communities interested in long term data series as part of their scientific analysis of phenomena, and retrieving their user requirements as data consumers. Another result of relevance of this activity consists in an effective list of what is necessary to preserve, why, for how long time and what is the additional information to be taken into account in any preservation process.

On the other hand, LAST activity has contacted data holder users in order to elaborate a set of system requirements of reference to cover present and future technological needs for massive storage. Since technology advances very quickly (i.e. and as soon as a new device is ready, the next generation is about to see light), a number of evaluation models has been defined to compare the different technologies, using metrics that help to make appropriate decisions on the technologies and architectures to be used. Finally, an assessment of architectures that met LTDP guidelines and the system requirements was done, analysing the risks in each case. The following conclusions about the architectures assessment can be summarized:

- The storage-centric architecture is recommended for high scalable and reliable systems.
- Although P2P and C/S architectures are compatible with LTDP guidelines and LAST system requirements, the main differences consist in the security, cost, and scalability of the systems. A control of the dimension and growth rate of the archive is recommended for C/S, and a control of the security is recommended for P2P.

- The N-Tier architecture is recommended to be implemented in the context of small/medium systems, oriented to fast accessibility, specific use and low scalability. These systems can be perfectly integrated in the context of more scalable systems like Cloud or SOA.
- Public Cloud is recommended when the information can be publicly available and shall be replicated at least in 2 different Cloud providers. In case the information is confidential, the recommendation is private cloud, which involves a major cost, expertise and resources.
- SOA is highly recommended for high scalable systems. However, the related cost to mitigate risks and the necessary expertise is relevant. The main advantage of SOA is the interoperability, which allows combining heterogeneous systems that may follow other types of architectures, like Cloud (Private or Public, depending on the specific requirements) or N-Tier (e.g. legacy systems with low growth and an appropriate interface for connecting with SOA system).

REFERENCES

[1] - Long Term Preservation of Earth Observation Space Data, European LTDP Common Guidelines Issue 1.1., 10 September 2010

[2] - European Strategy for Long term EO data preservation and access, ESA/PB-EO/DOSTAG(2007)2, 8 October 2007

[3] - Pérez R.F., Pérez O., Portela O., Saenz A., Nieto A., Leone R., Albani M., Beruti V., "Technological evaluation model for long-term data archive systems in the context of Earth Observation", MSST 2011 Conference, Denver (2011)

[4] - Committee on Earth Observation Satellites: http://www.ceos.org/

[5] - Group on Earth Observations: http://www.geoportal.org/web/guest/geo_home

[6] - Reference Model for an Open Archival Information System (OAIS), Blue book, Issue 1, 2002

[7] - International Standard ISO 15489-1 "Information and documentation - Records management -", September 2001

[8] - International Standard ISO 19115:2003 "Geographical Information and services - Metadata -", May 2003

[9] - U. Troppens, R.Erkens, W.Müller-Friedt, R.Wolafka, N. Haustein, Storage Networks Explained, John Wiley and Sons, 2009, pp.2-8

[10] - LAST Requirements Document, vol 1, Common Set of Requirements, September 2010

[11] - LAST Requirements Document, vol 2, Due Diligence Feedback, August 2010