Evaluation model for long term data archiving systems in the context of Earth Observation

Ruben F. Perez, Oscar Perez, Oscar Portela, Antonio Saenz, Amalio Nieto
GMV Innovating Solutions
Madrid, Spain
rfperez@gmv.com, operez@gmv.com, oportela@gmv.com, ansaenz@gmv.com, anieto@gmv.com

Rosemarie Leone, Mirko Albani, Vincenzo Beruti
ESRIN
European Space Agency
Rome, Italy
Rosemarie.Leone@esa.int, Mirko.Albani@esa.int, Vincenzo.Beruti@esa.int

***Abstract*— The Long term data Archive Study on new Technologies (LAST) project aims to assess and benchmark long term data archival technologies supporting the European Space Agency Earth Observation Long Term Data Preservation Program.**

A classification of technologies is performed in the main technological areas involved in Long Term Archiving. An evaluation method based in Analytic Hierarchy Process is used to identify the most appropriate technologies in each technological area, addressing with the specific user preferences and the identification of relevant evaluation criteria (i.e. evaluation model). As a case of study, an evaluation model is defined for Storage Hardware Systems, considered as a main Technological Area in Long Term Archiving.

LAST; Archiving (Long-Term); Essential Climate Variable; ECV; ESA; LTDP; Technologies (Archiving); Method (Evaluation); Model (Evaluation); Analytic Hierarchy Process; HPA

I. 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 [1, 2].

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.

Long term data preservation includes the continuous consolidation and technical evolution of archives, archive management systems and data access systems to guarantee the basic data preservation and proper data accessibility. Beyond and even more importantly, archived data can be used only if also the processing chains, the algorithms and the data access technology are maintained and evolve such that users can actually receive and process the data products always with up to date technology. Archive management includes as well interoperability, standardization issues, archive data security and archive certification processes.

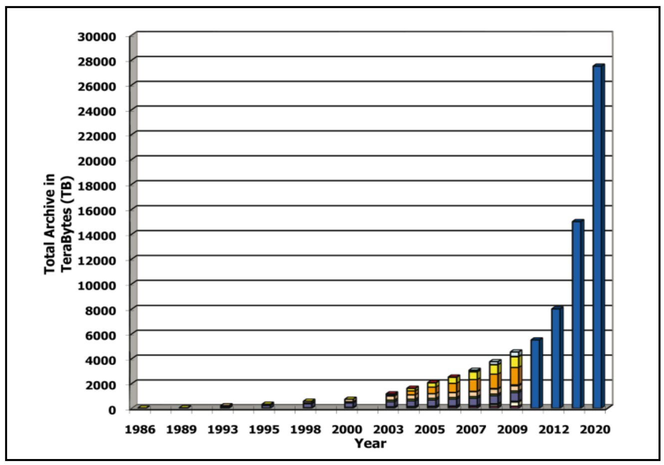

Figure 1. Evolution of ESA's EO data archives (source: ESA).

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 has been established. These initial guidelines are being consolidated and promoted within the Committee on Earth Observation Satellites (CEOS) [3] and Group on Earth Observation (GEO) [4], and constitute the basis for the ESA's EO data preservation approach and for the further cooperation with other European EO data archive holders.

In this context, the goal of the LAST project is 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.

978-1-4577-0428-4/11/$26.00 ©2011 IEEE

II. REQUIREMENTS ANALYSIS

The first stage of the LAST project scope was focused on the following activities:

- Definition of the functional and system requirements (e.g. architecture, performance, interoperability, etc.) which are mandatory or recommended for a Long Term Archive (LTA) implementation in accordance to the European Long Term Data Preservation of ESA Earth Observation Space data common guidelines (1).
- Survey of LTA archival requirements and their implementations at facilities operated on behalf of ESA.
- Verification and validation of the functional and system requirements to ensure their completeness, accuracy and applicability by means of auditing techniques, such as formal reviews of the technical documentation available and expert consultation meetings.

A case study for the LAST project is the demand for scientific information in the frame of the Climate Change programme initiatives. Particularly one of the Essential Climate Variable (ECV) exploitation projects which aims at carrying out a study on ECV bulk processing and prototype implementation. In the frame of the latter project more than 14 years of data shall be processed and archived and data shall be made available to user in near real time. The archiving system shall be deployed on a near-line basis and be stepwise upgraded on a three years basis. The case study project aims at implementing an operational environment where thousands of Terabytes of data will be incorporated along the life of the different missions. For this case study the technological evaluation model defined in the frame of the LAST project has provided valuable inputs about the LTA architectural concepts supporting the ECV exploitation project.

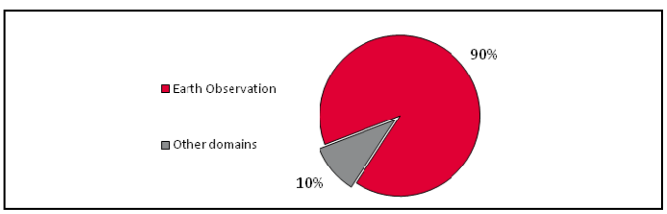

Figure 2. Distribution of answers by domains.

The archival and management of vast amounts of data are not exclusive of the EO domain. As part of the project Due Diligence activities, many organizations on different fields were contacted for a survey on their archival technologies:

- Astronomy, high energy physics scientific organization.
- Supercomputing centers.
- Digital libraries and repositories.
- Online storage and services.

A questionnaire was sent to the identified list of organizations and a significant percentage of the latter replied to the questionnaires and/or were interviewed. The percentage of those organizations outside the Earth Observation domain that replied to the questionnaire is illustrated by Fig. 2.

From the assessment of the surveys, questionnaires and interviews a common set of functional and system requirements was derived. These requirements were classified according to the following main topics:

- Standardization: relation of the archive system with applicable standards, e.g.: European LTDP Common Guidelines [1], the OAIS reference model [5], International Standards for Information and documentation [6], and Geographic Information-related Metadata and Services [7, 8].
- Reliability: oriented to the appropriate backup and redundancy mechanisms, the 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.). It's worth noting that some archives may need to support legacy technologies for extended periods of time, preventing the adoption of new systems.
- Migration: addressing 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, the way of handling of nominal user requests, and the particulars of access and retrieval aids.
- 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 physical and logical access to archive data, which should only be granted to authorized personnel and users for the different operations.
- Operations: providing a number of guidelines for the management of the archive and, more precisely, 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, the type of software used for the archive operations, and some preferred hardware implementations.

These requirements were validated and traced against the most relevant applicable standards recommended by the LTDP common guidelines and identified during the Due Diligence (Tab. I) survey, and discussed with a group of experts from various institutions -and the industry- during a technical workshop. Among these, it shall be noted that the OAIS reference model represents the foremost standard with respect to the long term preservation of data and information, and is highly regarded -and widely adopted- across the Earth Observation community. Moreover, it provides a common framework of terms and concepts which establishes a shared baseline among LTDP-aware parties, enabling the subsequent discussion and exchange of information.

Additional standards not considered for the validation and traceability of the requirements, but followed by some of the EO partners, often represent subsets, supersets or refinements of some of those mentioned here (Tab. I). This is the case, for example, of the INSPIRE profile of ISO 19115 and ISO 19119, and of the WMO Information System (WIS) specifications. Furthermore, the Open Geospatial Consortium (OGC) has a close relationship with the ISO working group behind the ISO 19100 series (i.e. TC211), and some of the standards in these series have in fact been jointly developed by both groups, even superseding the OGC abstract specifications in some cases.

TABLE I. APPLICABLE STANDARDS

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

The resulting set of requirements represented the initial baseline intended to support the remaining tasks in the LAST study, serving as guiding principles for the evaluation of the surveyed technologies and the resolution of different trade-offs to be made. Finally, the definition of the proposals and recommendations to be issued at the end of the project will be also reviewed according to them, as a means of validation regarding their suitability.

III. EVALUATION METHOD

The complexity of LTA systems requires evaluation methods that deal with several areas of archiving systems technologies and a number of parameters of interest that may be more relevant from one system to other, depending on the specific requirements and preferences of the final users and teams involved in the management and maintenance of the archival system. Therefore, it is not possible to define an absolute mark for each technology in relation to LTA archive systems in general, but a level of suitability in relation to each particular LTA context. As a consequence, the technologies to be evaluated in each Technological Area (TA) of an archiving system are determined according to the relevance of the aspects and parameters in each specific implementation.

Taking into account that the evaluation of technologies depends on several evaluation criteria (i.e. aspects and parameters to be taken into account) and given that each of those criteria may count with more or less relevance depending on the specific system to be implemented, a method based in the Analytic Hierarchy Process (AHP), which may be used to evaluate different types of criteria -even if they are structured or nested- was applied to all the TAs involved. AHP has been broadly used in literature for the mathematical adjust of structured information aimed to support general decision-making [9]. The main benefits of this method are its easiness of use, descriptive and comprehensive approach, orientation to hierarchical models, and mathematical evaluation. The application of this method allows obtaining an evaluation mark for each product depending on the independent evaluations provided by the experts for each of the evaluation attributes. The weights, on the other hand, are defined in cooperation with the LTA stakeholders, according to the relevance or preference assigned to each attribute.

The structured parameterization of the evaluation criteria defines an evaluation tree, in which a weighted evaluation by levels is adopted in order to allow a maximum flexibility and applicability to all the TAs involved. Thus, according to the evaluation method proposed, the following elements shall be defined in order to evaluate the candidate technologies:

- List of technologies to be compared, filtered according to the scope of the evaluation.
- Evaluation criteria, i.e.: comparison parameters and aspects of evaluation.
- Weights of the comparison parameters, assigned according to the preferences and requirements of final system to be implemented.
- Marks of the technology in each of the evaluation criteria.

The evaluation at any level of the tree is obtained by means of a function (i.e. Value of Benefit) which adjusts the different marks obtained for the sub-nodes according to the weights assigned in each case. This method allows defining a technological gap between LTA systems that are identical except for the technologies used. In the LAST project, this definition has allowed to measure the gap between the current ESA LTA implementations and the best solutions identified.

A. *Evaluation Criteria*

The evaluation criteria (aspects and parameters to be evaluated) that may apply to each different TA have to be identified and properly defined, which may follow a structure on different levels as in example shown in Fig. 3, where a particular TA (i.e. Hardware Storage Systems) includes several aspects (e.g. Performance, Cost etc) at the first level, and similarly for the rest of levels.

B. *Weights of Evaluation Criteria*

The definition of the weights for the criteria involved at each level are defined by means of percentages, providing with a level of relevance for each of the criteria. Those weights depend on the LTA system to be implemented and are set in agreement with the archiving implementers to fit with their specific requirements and preferences. The weights are normalized at each node of the evaluation tree so the different criteria of immediate lower level sum 1.0.

C. Technology Evaluation Marks

All the measurements that are obtained to assign a mark to an evaluation parameter (e.g. Economical Aspects in Fig. 3) are defined as pre-normalized values and are classified according to numeric and nominal types. The normalized marking scale is set by definition to the range 0-10, being 0 associated to the worst mark and 10 to the best one.

D. Value of Benefit

Based on the marks obtained in the different parameters of evaluation and weights of criteria, an adjusted Value of Benefit (VB) is calculated according to (1), where M represents the mark of each parameter of evaluation in the node and W the weight assigned, the sum applied to all the parameters nested to the tree node. Note that the Value of Benefit provides with an adjusted mark with a value that ranges from 0 to 10. This mark may contribute at the same time to the calculation of a value of benefit at higher levels in the evaluation tree by means of the same formula and the corresponding upper weights assigned until a global value of benefit is obtained, providing with the adjusted mark for the technology in relation to the aspects and evaluation criteria taken into account.

$$VB = \sum M_k W_k \quad (1)$$

E. Technological Gap Analysis

Given a structured evaluation criteria (i.e. an evaluation tree and associated weights defined), and the corresponding evaluation marks obtained for the criteria of each technology evaluated, a list of technologies is obtained, where the best technologies correspond to the highest VB obtained. The relative technological gap of a technology in relation to the best technologies is defined according to (2), where VB_{Tech} corresponds to the global mark of this technology and $VB_{TopTech}$ the global mark of the best evaluated technology.

$$T_{GAP} = 1 - VB_{Tech} / VB_{TopTech} \quad (2)$$

Therefore, the measurement of the technological gap is obtained from the comparison between the best evaluated technologies and the technologies currently in use for a given set of criteria and weights. This can be used as an estimator of a system's potential for improvement when one or more of its technologies are to be replaced by new ones, constituting one of the most important benefits of the method used (i.e. AHP).

IV. ANALYSIS OF TECHNOLOGICAL AREAS

During the first phase of the LAST project, an identification of the various TAs of interest was carried out in order to later provide an assessment of technologies in each one, e.g.: servers and computing platforms, operating systems, databases, information architectures, communication infrastructures and, first and foremost, hardware storage systems and technologies.

All of the evaluation criteria analyzed were structured (e.g. Fig. 3), following the classification of main aspects of LTA systems suggested by the requirements analysis (i.e. Standardization, Reliability, Maintenance, Migration, Interface, Performance, Security, Operations, Procurement) plus some common evaluation parameters which are transversal to the rest of aspects (e.g. Cost, Availability, etc).

A. Common Evaluation Parameters

After an analysis of the different TAs involved in LTA systems, the following common parameters were established, not being covered by the technical parameters identified during the analysis of the system requirements:

- Cost of the Technology: involving the licensing model of the technology under study and rest of costs involved in its usage and maintenance (e.g. Total Cost of Acquisition, Total Cost of Ownership, etc).
- Vendor Mid and Long Term Financial Situation: the financial situation and the future prospects in the mid and long term of a manufacturer shall also be taken into account when choosing a technology.
- Availability: expected market availability of those technologies mature for operation in the short and mid-term time frame, or foreseen to be ready in the long-term.

B. Evaluation Model for Hardware Storage Systems

Hardware Storage Systems are used to perform read/write operations over the media that store (or hold in a logical or physical way) the data and files of the LTA. Such systems are composed of more elementary devices, either directly or as aggregations of these, forming larger storage entities which can be shared among multiple systems and which usually offer different capabilities (e.g. storage, backup, data replication, etc.) oriented to massive information storage.

TABLE II. HDD STORAGE SYSTEMS SELECTED FOR EVALUATION

Vendor	On-Line Products	
	Name	*Capacity*[a]
EMC	Symmetrix DMX-4	1920 TB
Fujitsu	Storage System ETERNUS DX8400	2008 TB
Hitachi	Universal Storage Platform® V	2269 TB
HP	StorageWorks P9000	1200 TB
Huawei	Symantec OceanspaceTM S8100	2400 TB
IBM	System Storage DS8800	634 TB
Netapp	FAS6200	2880 TB
Oracle	Z FS Storage 7740	1150 TB

a. Approximated capacity per unit (as marketed by the vendor, subject to specific implementation).

TABLE III. TAPE STORAGE SYSTEMS SELECTED FOR EVALUATION

Vendor	Near-Line Products	
	Name	*Capacity*[a]
Fujitsu	ETERNUS LT270 Tape Library	1PB
HP	StorageWorks ESL E-series Tape Library	15 PB
IBM	System Storage TS3500 Tape Library	10 PB
Oracle	StorageTek SL8500 Modular Library System	10 PB
Quantum	Scalar i6000 Tape Library	15 PB
Spectra	T-finity Tape Library	5 PB

a. Approximated capacity per unit (as marketed by the vendor, subject to specific implementation).

Given the broad and heterogeneous range of systems involved in the present TA, a selection of products -representative of current market and technology trends- was chosen for evaluation (i.e. HDD Libraries in Tab. II, and Tapes Libraries in Tab. III). Most manufacturers allow the replication of data (either synchronous or asynchronous) between storage systems geographically separated contributing the platforms reliability. A fundamental piece for integrating different storage systems is Hierarchical Storage Management (HSM), a technique broadly used in LTA which automatically moves data between high-cost and low-cost storage media. Additionally, Information Lifecycle Management (ILM) can be combined with HSM in order to have a cost-effective storage of the data over its entire life cycle [12]. This is not the case of most of EO archiving systems interviewed, as the information is preserved with the same requirements of access indefinitely.

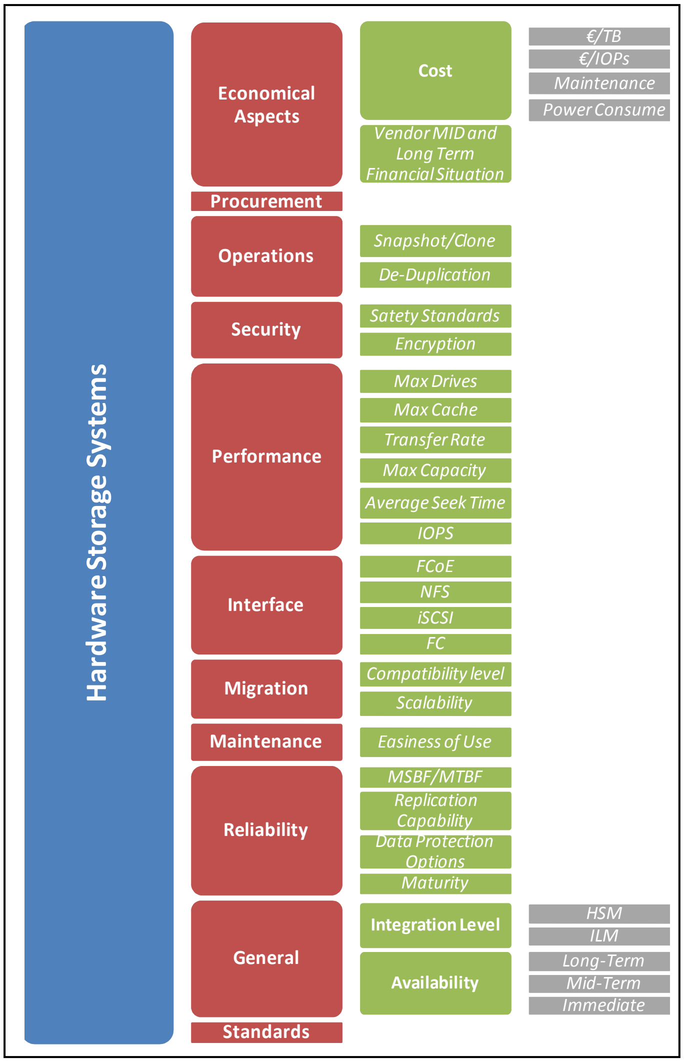

Figure 3. Evaluation Model for Hardware Storage Systems.

Last, a number of evaluation parameters of interest and their relationships (in line with literature recommendations [10-12]) were defined and structured, forming the evaluation model depicted in Fig. 3, allowing the different weights to be set and tuned by ESA according to its preferences on each of the aspects to be evaluated.

V. SUMMARY AND FUTURE WORK

After defining a reference LTA system for ESA in line with the applicable standards (Tab. I), a set of key aspects was identified for the evaluation of all TAs in a general LTA system. A method (i.e. AHP) that takes into account the preferences and specific requirements on the final LTA system to be implemented has been described, involving firstly the identification of general TAs, the possibility of using a specific set of evaluation criteria in each TA, setting of weights defining the relevance of the different evaluation criteria in each TA, and providing with a mark of evaluation by means of the *VB* (1). Additionally, a definition of a metric has been provided by means of this method, related to the technological gap between a current system and the best upgrade path concerning a TA (2).

Having described the method employed for the evaluation of LTA systems, a model has been proposed for the analysis of a particularly relevant TA (i.e. Hardware Storage Systems) along with the corresponding evaluation criteria, making use of specific aspects and attributes defined in the scope of LAST project. The main advantage of the proposed model is that all attributes are classified according to the different aspects in the LTA, represented by a weight which may be modified in the short-term as the preferences of the final users are further discussed and possibly changed.

In addition to the TA described in this article, other TAs are undergoing a detailed analysis at the moment (i.e. Operating Systems, Hardware Platforms, Middleware, Communication Protocols, Databases and Communication Networks) aiming at providing ESA with an updated and more comprehensive assessment concerning the most appropriate technologies and architectures for LTAs in relation to LTDP guidelines.

VI. REFERENCES

[1] Long Term Preservation of Earth Observation Space Data, European LTDP Common Guidelines Issue 1.1., 10 September 2010.
http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue1.1.pdf

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

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

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

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

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

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

[8] International Standard ISO 19119:2005(E) "Geographical Information and services - Services", February 2005.

[9] T.L. Saaty and L.G. Vargas, "Models, methods, concepts & applications of the analytic hierarchy process", Kluwer Academic Publishers Group, 2001, pp. 27–34.

[10] Storage Network Industry Association: http://www.snia.org/home/

[11] T.C. Jepsen, "Distributed Storage Networks", John Wiley and Sons Ltd., 2003.

[12] U. Troppens, R. Erkens, W. Müller-Friedt, R. Wolafka, N. Haustein, Storage Networks Explained, 2nd ed., John Wiley and sons Ltd., 2009.