ESA USE CASES IN LONG TERM DATA PRESERVATION

Mirko Albani Rosemarie Leone Calogera Tona

ESA-ESRIN
Via G. Galilei
CP 64,00044 Frascati
Italy
name.surname@esa.int

ABSTRACT

Long Term Data Preservation (LTDP) aims at ensuring the intelligibility of digital information at any given time in the near or distant future. LTDP has to address changes that inevitably occur in hardware or software, in the organisational or legal environment, as well as in the designated community, i.e. the people that will use the preserved information. A preservation data manages communication from the past while communicating with the future. Information generated in the past is sent into the future by the current preservation data. European Space Agency (ESA) has a crucial and unique role in this mission, because it maintains in its archives long time series of Earth Observation (EO) data. In order to ensure to future generations data use and accessibility of this cultural heritage is needed to define a systematic approach, accompanied by different use cases.

Keywords

Long Term Data Preservation (LTDP), Data Curation, ESA, EO data, Preserve Data Set Content (PDSC).

1. INTRODUCTION

The main objective of the Long Term Data Preservation (LTDP) initiative is to guarantee the preservation of the data from all 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].

The concept of LTDP can be characterized as communication with the future. In the future new technology will be used that is more cost effective and more sophisticated than current technology. Communication with the future then corresponds to moving records onto new choices of technology. The preservation environment will need to incorporate new types of storages systems, new protocols for accessing data, new data encoding formats, and new standards for characterizing provenance, authenticity and integrity. The long term preservation of Earth Observation data is a major issue today as monitoring of global change processes has led to increasing demand for long-term time series of data spanning 20 years or more also in support to international initiatives such for example the United Nations Framework Convention on Climate Change (UNFCCC), the ESA Climate Change Initiative (CCI) and the Global Monitoring for Environment and Security program (GMES). The large amount of new Earth Observation missions upcoming in the next years will lead to a major increase of EO space data volumes and this fact, together with the increased demands from the user community, marks a challenge for Earth Observation satellite operators, Space Agencies and EO space data providers regarding coherent data preservation and optimum availability and accessibility of the different data products. The preservation of EO space data can be also in the future as a responsibility of the Space Agencies or data owners as they constitute a humankind asset.

In 2006, the European Space Agency (ESA) initiated a coordination action to share among all the European (and Canadian) stakeholders a common approach to the long term preservation of Earth Observation space data. During 2007, the Agency started consultations with its Member States presenting an EO Long Term Data Preservation strategy targeting the preservation of all European (including Canada) EO space data for an unlimited time-span ensuring and facilitating their accessibility and usability through the implementation of a cooperative and harmonized collective approach (i.e. a European EO LTDP Framework) among the EO space data owners in order to coordinate and optimize European efforts in the LTDP field and to ultimately result in the preservation of the Completed European EO space data set for the benefit of all European countries and users and with a reduction of overall costs.

The Long Term Data Preservation Working Group with representatives from ASI, CNES, CSA, DLR and ESA was formed at the end of 2007 within the Ground Segment Coordination Body (GSCB) [1], with the goal to define and promote, with the involvement of all the European EO space data and archive owners, the LTDP Common Guidelines and also to increase awareness on LTDP. The LTDP guidelines were published at the end of 2009 and constitute a basic reference for the long term preservation of EO space data. Their application by European EO space data owners and archive holders is fundamental in order to preserve the European EO space data set and to create an European LTDP Framework. The application of the identified guidelines is not a requirement or a must for European EO space data owners and archive holders but is strongly recommended also following a step-wise approach starting with a partial adherence.

This paper is organized as follows: Section 2 presents state of the art, Section 3 shows LTDP architecture, Section 4 provides a use cases overview and Section 5 presents the conclusions and future developments

2. PRESERVATION OF EO SPACE DATA: STATE OF THE ART

The main milestones in LTDP development are:

- LTDP Framework
- LTDP Common Guidelines

- LTDP Preserve Data Set Content

2.1 LTDP FRAMEWORK

LTDP Framework, and others such as European LTDP Common Guidelines, was produced by the LTDP working group. The concepts contained in this document were presented to the EO data owners and archive holders community at the 1st LTDP workshop held at ESA/ESRIN in May 2008 [2].

Its main goal is to define a "European LTDP Framework" aimed at providing a practical way of carrying on LTDP activities at European level. The initial concepts and ideas contained in this document should help the establishment of a European LTDP Framework to coordinate and optimize European efforts in the LTDP field, that, in turn, would ultimately result in the preservation of the complete European data set with a coherent and homogeneous approach for the benefit of all European countries and users and with a reduction of overall costs.

Main goals of the European EO Long Term Data Preservation Framework are to:

- Preserve the European, and Canadian, EO space data set for an unlimited time-span.
- Ensure and facilitate the accessibility and usability of the preserved data sets respecting the individual entities applicable data policies.
- Adopt a cooperative and harmonised collective approach among the data holders and archive owners (European LTDP Framework), based on the application of European LTDP Common Guidelines and sustained through cooperative (multi-source) long term funding schemes.
- Ensure, to the maximum extent, the coherency with the preservation of other non-space based environmental data and international policies.

The European LTDP Framework is open to all possible members and is to be intended as a collaborative framework consisting of distributed and heterogeneous components and entities cooperating in several areas to reach a harmonized preservation of the European EO space data set. The framework is based on the contribution of European EO space data owners through their ideas and possibly their infrastructure in accordance to the commonly agreed LTDP Guidelines and should follow a progressive implementation based on a stepwise approach (short, mid, long-term activities). A common approach in the field of Long Term Data Preservation should aim at the progressive application of the European LTDP Common Guidelines but also at cooperation of the archive owners in several areas for a progressive development and implementation of technology, methodology, standardization, operational solutions and data exploitation methodologies as key aspects for the set-up of the framework.

A cooperative framework can facilitate for EO space data owners and archive holders the achievement of the common goal of preserving and guaranteeing access to the own data through benefiting from proven technologies, procedures and approaches and through the possibility to reuse and share infrastructure elements in the long term. The adoption of standards (e.g. for data access interfaces and formats, procedures, etc...) and common technical solutions can also allow to significantly reduce preservation costs.

The European LTDP Framework should be sustained through a cooperative programmatic and long term funding framework based on multilateral cooperation with multiple funding sources from at least the European EO space data owners.

The existence of a European LTDP Framework will also increase the awareness on data preservation issues favouring the start of internal processes at private or public European EO space data owners and providers. A European framework could also trigger the availability in the long term of additional permanent funding sources (e.g. European Commission) and can increase the possibility for any European (including Canada) EO space data owner to preserve missions data beyond their funding schemes into the cooperative and distributed framework.

2.2 LTDP COMMON GUIDELINES

The European LTDP Guidelines are intended to cover the planning and implementation steps of the preservation workflow and have been defined on the basis of a high-level risk assessment performed by the LTDP Working Group on the Preserved Data Set Content [3] and its composing elements.

The LTDP guidelines and the underlying data preservation approach should be applied not only to future missions, where they can be easily and systematically included in the mission operations concept starting from the early phases with consequent cost savings and better achievable results, but also to the missions currently in operation or already disposed. In those last cases their application and the recovery of the full EO PDSC content could be trickier and not completely achievable and tailoring might be necessary. For current and not operational missions in any case, an incremental application approach should be pursued; the approach should consist in auditing the archives versus the LTDP Guidelines and PDSC document to be followed by the implementation of the highest priority and by the recovery of critical missing data/information.

In the field of Earth Observation, the data landscape is complex and there will naturally be different user communities with divergent needs for the long term reuse of the data. In case a more specific designated user community has to be addressed wrt, more specific preservation objectives and PDSC content should be defined and the LTDP Guidelines might need to be refined and augmented accordingly. In those cases it is recommended to follow the steps using the PDSC and the LTDP guidelines as starting point for the definition of a more specific approach to be properly documented in the form of "preservation approach and strategy" documents.

2.2.1 Preservation analysis workflow

Preservation of Earth Observation data should rely on a set of preservation actions properly planned and documented by data holders and archive owners, and applied to the data themselves and to all the associated information necessary to make those data understandable and usable by the identified user community. Data holders and archive owners should follow the "Preservation Analysis Workflow" procedure to define the proper preservation strategy and actions for their Earth Observation data collections. The result of the procedure application should consist of a set of documents describing the preservation strategy, implementation plan and activities for each individual mission dataset. Such document(s) should refer to the LTDP guidelines and clearly

define current compliance and future plans to improve adherence. The procedure consists of the following steps:

- Definition of preservation objective and designated user communities.
- Definition of Preserved Data Set Content (PDSC) for Earth Observation missions.
- Creation of PDSC Inventory for each own EO mission/instrument dataset.
- Risk assessment, preservation planning and actions, risk monitoring.

These steps are applicable to any digital data repository and are shortly described below for a generic Earth Observation case:

The preservation objective considered here for an Earth Observation data holder and archive owner consists in maintaining the own full data holdings accessible and usable today and in future, theoretically for an unlimited time, for its designated user communities. Long-term accessibility and usability of Earth Observation data requires that not only sensed data but also the associated knowledge (e.g. technical and scientific documentation, algorithms, data handling procedures, etc.) is properly preserved and maintained accessible. This implies the availability and archiving of metadata and data products at all levels specified by each owned mission or the capability to generate them on user request through proper processing. Data products need moreover to be provided with known quality to end-users together with the information necessary to understand and use them.

Different designated user communities are addressed through the preservation objective defined above. Earth Observation data users are today, as an example and among others, Scientists and Principal Investigators, researchers, commercial entities, value adders, and general public. These communities can be further differentiated on the basis of the respective application domain and area of interest (e.g. ocean, atmosphere) and generally have different skills, resources and knowledge. The data product levels and the information associated to the data necessary for their understandability and use is different for each of the above communities and even for individuals inside each community. Earth Observation data holders and archive owners generally serve today more than one user community and therefore need to be able to address the needs of all of them in terms of data and associated information availability and access. In addition, the preservation objective includes the utilization of the data products also in the future by user communities that might have completely different skills and knowledge base wrt the ones identified today but also different objectives for the use of the data. This means that the best approach for Earth Observation data holders and archive owners today would be to consider a "designated user community" generic and large enough so that the identified content to be preserved in the long term for that community will allow also other users, not considered at the time preservation was initiated, to make use of the data in the future. The generic designated user community is assumed to be able to understand English, to work with personal computers and basic programs provided with them, and to analyse and interpret the data products when available together with the full amount of additional information necessary to understand them without additional support from the archive.

In Earth Observation, the "Preserved Data Set Content" should be comprised as a minimum, in addition to the EO data, of all information which permit the designated user community to successfully interact, understand and use the EO data as mandated by the preservation objective. The Earth Observation Preserved Data Set Content has been defined on the basis of the preservation objective and generic designated user community.

For past and current missions, the next stage to be implemented by data holders and archive owners is to tailor the PDSC for each EO mission/instrument, and to appraise each of the resulting elements comprised in the preserved data set content in terms of physical state, location and ownership. The result is the mission/instrument inventory document. For future missions, the definition of the PDSC shall be initiated during the mission definition and implementation phases and continuously maintained in the following phases.

Risk assessment in terms of capability of preservation and accessibility for each element of the inventory should be then performed and the most appropriate preservation actions identified and planned for implementation. The result of this activity should consist of one or more "preservation strategy and approach" documents. These documents could be drafted with different levels of detail and should generally contain Preservation Networks for each EO mission data collection consisting of all the PDSC Inventory elements, the elements on which they are dependent or necessary to understand and use them (e.g. the operating system underlying an EO data processor) and the associated preservation actions identified for each of them. Preservation networks should also identify the preservation state of each element of the PDSC inventory. Such document(s) should refer to the LTDP guidelines and clearly define current compliance and future plans to improve adherence. The identified preservation actions should be then implemented and the risks associated with inventory elements preservation properly and continuously monitored.

2.2.2 LTDP Guidelines Content

The guiding principles that should be applied to guarantee the preservation of EO space data in the long term ensuring also accessibility and usability are:

- Preserved data set content
- Archive operations and organization
- Archive security
- Data ingestion
- Archive maintenance
- Data access and interoperability
- Data exploitation and re-processing
- Data purge prevention

The LTDP guidelines constitute a basic reference for the long term preservation of EO data. Their application by European Earth Observation space data holders and archive owners is fundamental in order to preserve the European EO space data set and to create a European LTDP Common Framework. The application of the identified guidelines is not a requirement or a must for European EO data holders and archive owners but is strongly recommended along with following a step-wise approach starting with a partial adherence. The key guidelines should be

intended as a living practice and as such might evolve following specific research and development activities (e.g. outcome of cooperation in LTDP in Europe). Each key guideline could also have associated a set of technical procedures, methodologies or standards providing technical details on the recommended practical implementation of the guideline. Their selection has been made considering the results of cooperation activities in Europe with the goal to favour convergence in Europe on the LTDP approach and implementation.

Similarly to the key guidelines, these procedures or standards could be further evolved and improved with time or even developed or defined if missing. This can therefore also be intended as a starting point to support the establishment, and aid the implementation, of such detailed procedures or methodologies when missing, favouring active cooperation in Europe in the LTDP field. LTDP principles and key guidelines considered necessary to initiate this process and enable more detailed, specific and technical guidelines to be established by appropriate technical experts. The LTDP Common Guidelines document will be periodically updated to reflect the advances of activities carried out in the LTDP area and will be submitted, in the framework of the periodical updates, to public reviews to collect feedback and comments.

2.3 LTDP PRESERVE DATA SET CONTENT

LTDP Preserve Data Set Content (PDSC) indicates what to preserve in terms of data and associated knowledge and information during all mission phases to be able to satisfy the needs of the Earth Science Designed community today and in the future [5]. LTDP PDSC addresses the Earth Science context (i.e. Earth Science Domains) and the specific Earth Observation domain, based on the data categorization taxonomy described below.

Methods, standards and sound criteria are needed to certify whether the preserved data set content is complete and will meet the needs of future users. Long – term preservation requires solving to complementary yet coordinated problems:

- Preservation of the data records itself (the raw data bits acquired from an Earth Science instrument);
- Preservation of the context surrounding the data records (the meta-information needed to interpret the raw data bits).

The acceleration increase of the amount of digital information sensed by Earth Science instrument coupled with the aging of our existing digital heritage and well published examples of the impacts of its loss have raised the criticality and urgency of the sensed data record stream preservation.

In the frame of FIRST survey, [4] the user community has clearly and strongly pointed out that preserving data records of Earth science historical mission is mandatory. Particularly, the scientific community welcomes the LTDP European initiative to cooperate and optimize efforts to preserve data set heritage for future generation.

One of the outcomes of the FIRST study is that the criticality of preserving the context information is not a static attribute of the context information itself but a dynamic outcome of past commitments of the consumer community, information curators and holding institution. In the frame of FIRST a preliminary attempt has been made to rank context information criticality for a generic Earth Science mission and for the nine Earth science sensors types. This preliminary ranking will be tuned following the results of the pilot implementation projects initiated by the Agency to preserve ESA historical data set and their context information using the checklists as reference

In the frame of ESA survey, the user community has been also pointed out that non only data records but the latter context requires preservation too, as context information might often be:

- hidden or implicit: well understood in their respective designated user communities and data producer experts at the time the data records stream is acquired and processed;
- evolving the technological context (computer platforms, programming languages, applications, file format etc..) surrounding any piece of information will inevitably change over time until information is no longer usable; and the communities context (data producer, data consumer, designated communities i.e. communities and organization involved in the information‟s creation and initial use) may change over time and give different value to the data information over time of cease to exists.

The combination of the context surrounding earth science data information being both implicit and evolving requires that for the information to remain meaningful and useful after a long time span either the data records information must continuously evolve with the context, or the context must be captured and preserved along with the preservation of the data records, preferably at the time of the information creation.

The context surrounding earth science data records is particularly complex as stated in the introduction. For example use of remote sensing imagery requires detailed knowledge of sensor and platform characteristics, which due to its size and complexity is not usually bundled with data objects in the same way that the descriptive metadata is. Furthermore geospatial data may require deep analysis to remain usable over time.

As a major example, to support the long term climate change variables measurement, historical data records must be periodically reprocessed to conform to the most recent revisions of scientific understanding and modeling. This in turn requires access to and understanding of the original processing, including scientific papers, algorithm documentation, processing sources code, calibration tables and databases and ancillary datasets.

Whereas preservation of bits requires that the bits stay unchanged over time, preservation of context must accommodated and even embrace change to the context. File formats will need to be migrated over time, new access services will need to be developed and will require new kinds of support and information will inevitably be re-organized and re-contextualized.

Thus a key consideration in the design of an archive system is that it be able to capture and preserve complex contextual data records information objects; maintain persistent associations between data

records information objects and contextual objects; and support modification of the context over time.

2.3.1 Preservation Principles

The principles stated in the previous paragraph have been applied in the definition of the preserved data set content:

- minimum time reference for long term preservation is usually defined as the period of time exceeding the lifetime of the people, application and platforms that originally created the information;
- preservation of the data records (the raw data bits acquired from the mission instrument) is mandatory
- the data record context surrounding the information (hidden or implicit information) shall be addressed too when defining the preserved data set content;
- the context must be captured and preserved along with the data records, preferably at the time of the information creation and taking into account the evolving characteristics of context information, particularly for long term data series.
- the criticality of preserving the data set content is dynamic, an outcome of past commitments on the consumer community, information curators and holding institution

To analyse what shall be preserved, the approach is based on four main dimensions:

- Time dimension: How long? How long shall the data set content be preserved at minimum?
- The Content dimension referred to as What? What data set content shall be preserved?
- The Stage during which the dataset is generated When? When shall the information be captured and preserved?
- The past, current and future perceived importance ("persistency") referred to as Rank: How critical is that the each information content object is preserved?

2.3.2 European EO Space Data Set

The European EO Space Data Set consists of all EO space data from missions or instruments owned by public or private organisations from European Member States and Canada and of all EO space data over Europe from non-European Member States missions or instruments available through agreements with European entities (e.g. Third Party Missions managed by the European Space Agency). The space missions or sensors whose data constitutes the European EO Space Data Set are subdivided in the following main categories:

- C1: High and Very High resolution SAR imaging missions/sensors (different Radar bands).
- C2: High and Very High resolution multi-spectral imaging missions/sensors.
- C3: Medium resolution Land and Ocean monitoring missions/sensors (e.g. wide swath ocean colour and surface temperature sensors, altimeter, etc.).
- C4: Atmospheric missions/sensors.
- C5: Other Scientific missions/sensors.

All missions and instruments comprising the European EO Space Data Set are described in a document, which is updated every six months.

3. ROADMAP VISION

ESA has developed the Standard Archive Format for Europe (SAFE) [6] an extension of the XFDU standard [7]. SAFE has been designed to act as a common format for archiving and conveying data within ESA Earth Observation archiving facilities. Many of the most important datasets have been converted to this format.

The important point is that XFDU, and therefore SAFE, is designed to implement the OAIS Archival Information Package [8], which in principle has everything needed for long term preservation of a piece of digitally encoded information. Some of the other components under consideration for the ESA LTDP technical implementation are the hardware needed to store the large volumes expected.

4. LTDP ARCHITECTURE

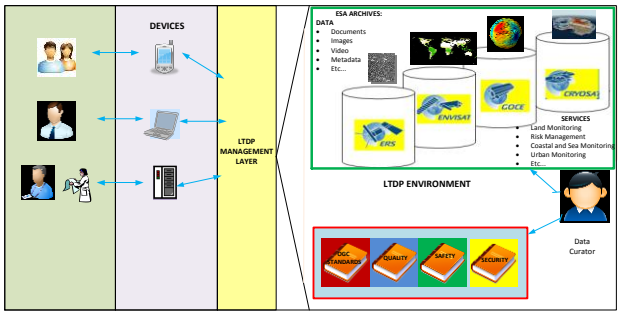

Figure 1 LTDP Architecture

Figure 1 shows the LTDP Architecture.

We distinguish different type of user, simple user or scientist, that with their devises access to LTDP Management Layer to obtain what they want. On the other side we have a Data Curator that is a crucial figure in LTDP Architecture. He preserves different mission data, documents, images, metadata, but particularly services with their technologies (i.e. Land Monitoring, Risk Management, etc...).

5. LTDP USE CASES

A use case expresses the functional, informational and qualitative requirements of a user (i.e. an actor or a stakeholder) whereby the functional requirements are represented by the „sequence of action", and the informational requirements cover the content of the „observable result". The qualitative needs encompass all the non-functional aspects of how the result is produced and the quality of the result which is important for the decision if the result is „of value" to the user. Therefore, the degree of abstraction and formalism, and the language, should be such that it is adequate for the domain of expertise of the stakeholders. To serve as an agreement, it should be understandable to the stakeholders but also precise enough.

In this work, the concept of use cases is applied in order to describe the high-level functional requirements. We use the

Unified Modelling Language (UML) for this purpose, but by extending the UML use case notation with references to major information objects that are required to implement the use case.

Figure 2 shows the basic template that is used to present the LTDP use cases .Two major types of actors are distinguished: first, human and users that use a client application by means of its user interface; and second , „Software Component" which represent a pieces of software that invokes an LTDP service by means of its service interface.

Use cases need information objects as inputs. These are indicated in the upper part of the diagram together with the required access method, i.e. create, read, write, delete. Results of use cases are listed as information objects in the lower part of the diagram. Information objects may be related to each other. Furthermore, use cases may have relationships to other use cases.

One use case may invoke another use case (which represents a dependency between use cases), or one use case may be a sub-variant of another.

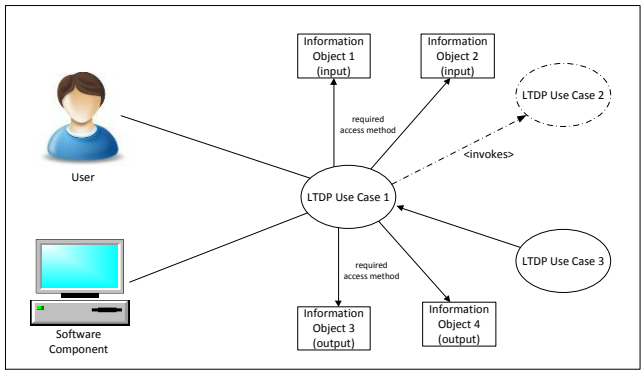

Figure 2 Main LTDP Use Case

5.1 DISCOVERY USE CASE

The Discovery Use Case deals with the question of how to find the EO resources (e.g. dataset, dataset series, services or sensors) of interest to a user. As in other application domain, such EO resources need to be described by some additional information, usually called metadata or metadata information. Metadata informs about the major characteristics of a resource .Metadata elements are stored in metadata stores (e.g. realised by relational databases) and accessed through interfaces of dedicated services.

The goal for the end user is to access those products that fulfil specific requirements according to his tasks. Essential requirements are, for instance:

- Region of interest
- Time series
- Usage of a specific satellite or sensor
- Corresponding ground station
- Additional attributes depending on the sensor type, e.g. cloud coverage.

As illustrated in figure 3, such requirements are entered as parameters in search queries. The access process delivers result sets that are specific to the resource types at which the search request has been targeted, i.e. delivers descriptions (metadata elements) of dataset series, sensors and /or services. The user may then browse through these metadata records and select those with which he wants to continue the interaction with other use cases.

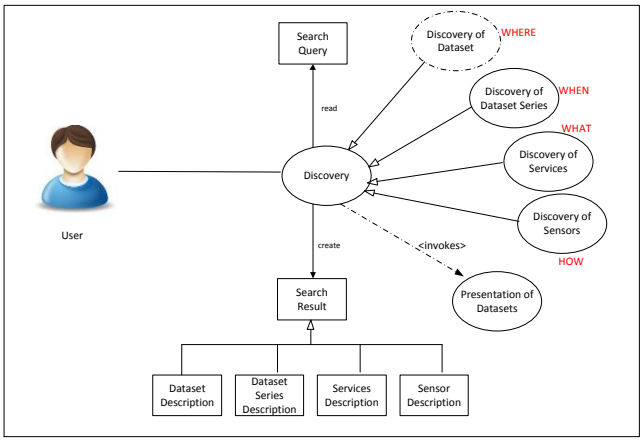

Figure 3 „Discovery" Use Case

6. CONCLUSIONS AND FUTURE DEVELOPMENTS

The Future LTDP ESA Program targets the preservation of scientific data and associated knowledge from ESA and ESA-managed Third Party Missions in all fields of Space Science and in particular scientific data generated by payloads and instruments on-board space platforms (e.g. spacecraft, International Space Station). These activities have the following main objectives: ensure and secure the preservation of archived data and associated knowledge for an unlimited time span knowing that they represent a unique, valuable, independent and strategic resource owned by ESA Member States and ensure, enhance and facilitate archived data and associated knowledge accessibility through state of the art technology and exploitability by users, including reprocessing, for all the ESA and ESA-managed Third Party Missions in all fields of Space Science covered under the LTDP activities.

In cooperation with other space science data owners of Member States establish a cooperative, harmonized and shared approach to preserve and maintain accessibility to European Space Science Data for the long term.

7. REFERENCES

[1] European Space Agency LTDP area on GSCB Website http://earth.esa.int/gscb/ltdp/objectivesLTDP.html

[2] European Space Agency LTDP Framework http://earth.esa.int/gscb/ltdp/EuropeanLTDPFramework.pdf

[3] European LTDP Common Guidelines, Draft

Version 2, http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue1.1.pdf

[4] First LTDP Workshop, May 2008, http://earth.esa.int/gscb/ltdp/LTDP_Agenda.html

[5] European Space Agency –PDSC http://earth.esa.int/gscb/ltdp/EuropeanDataSetIssue1.0.pdf

[6] SAFE web site http://earth.esa.int/SAFE/

[7] XFDU standard available from http://public.ccsds.org/publications/archive/661x0b1.pdf

[8] Reference Model for an Open Archival System (ISO14721:2002),http://public.ccsds.org/publications/archive/650x0b1.pdf or later version. At the time of writing the revised version is available at http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf or elsewhere on the CCSDS web site http://www.ccsds.org