Long Term Data Preservation
Analysis and contribution to the definition of the
Preserved Data Set Composition (PDSC)

Raffaele Guarino [(1)], Rosemarie Leone [(2)], Mirko Albani [(2)]

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

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

ABSTRACT

In this article, the results of the Long Term Data Preservation (LTDP) project FIRST (De**fi**nition of LTDP User **R**equirements and Preserved Data **Se**t Composition) are presented.

The European LTDP initiative aims at assuring the long term preservation of Earth Observation data from all ESA and Third Parties missions managed by ESA, as part of a joint and cooperative approach in Europe.

Through a coordinated work with Canadian and European Operators, Agencies and Organizations in Earth Observation, ESA published the Long Term Preservation of Earth Observation Space Data, European LTDP Common Guidelines [1].

The application of the LTDP guidelines by European EO space data owners and archive holders is fundamental in order to preserve the European EO space data and to create a European LTDP Common Framework.

The LTDP common guidelines recommend that a consistent and complete set of data is archived and preserved to enable current and possible future utilization and to guarantee the exploitability of the archived data and associated information content.

FIRST project objective is to capture user requirements and propose long term data preservation implementation roadmaps in order to meet the rising tide of scientific users data needs for long times series.

In the frame of FIRST a classification of Earth Science (ES) areas that have common needs of scientific data sets, preservation requirements, and time series demands is proposed. For each data category the data set to be preserved is identified from the scientific user perspective point of view.

Keywords: LTDP, User Requirements, System Requirements, Architectures, Long Term Archiving

INTRODUCTION

Earth Science communities represent a wide variety of disciplines, which utilize a multitude of different data and products. Timescales and time series continuity of data needed to analyze phenomena can be totally different depending on the specific scientific domain. For example, monitoring of global change processes is requiring even more for long-term time series of Earth Observation data spanning more than 20 years. These data are necessary to support international activities derived from the United Nations Framework Convention on Climate Change (UNFCCC).

The necessity to preserve and let available all scientific data is the cornerstone for any future use.

ESA has proposed and is leading the Long Term Data Preservation (LTDP) project aimed at guarantee the preservation of owned and partners' missions. In this frame preservation of data of all Earth Observation missions in the main objective.

Through a coordinated work defined with Canadian and European Operators, Agencies and Organizations having interests in Earth Observation, ESA has published the ***Long Term Preservation of Earth Observation Space Data, European LTDP Common Guidelines*** [1]. The document is organized in nine themes and guidelines specific to the theme. Those guidelines have been promoted through workshops and are in a consolidation process with other partners and organizations (e.g. CEOSS, GEO, NASA, etc.).

In this context the project **FIRST** (De**fi**nition of LTDP User **R**equirements and Preserved Data **S**e**t** Composition) is aimed at capture and analyze user requirements and needs having impacts in terms of long term preservation.

During the first part of the project, earth science data long term preservation requirements and needs were captured from the Earth Science user community in different fields and application domains and from international initiatives and programmes (e.g. Climate Change Initiative and GMES).

An assessment was done of the composition of the Earth Science data set that should be preserved in the long term in order to guarantee the satisfaction of the Earth Science user community's needs in the different domains with a particular focus on Earth Observation satellite data.

USER REQUIREMENTS COLLECTION

The method used to collect and analyze requirements is based on the main assumption that users of scientific data cannot be always distinguished from data producer and that a given individual or system may act in the role of both a Consumer and a Producer.

The project has focused three main aspects:

a. Identification of earth science data producers and data consumers ***field of interest***.
b. Identification of data set types
c. Identification of requirements from the user's perspective.

A taxonomic description of Earth Science areas and activities performed in each area is proposed below.

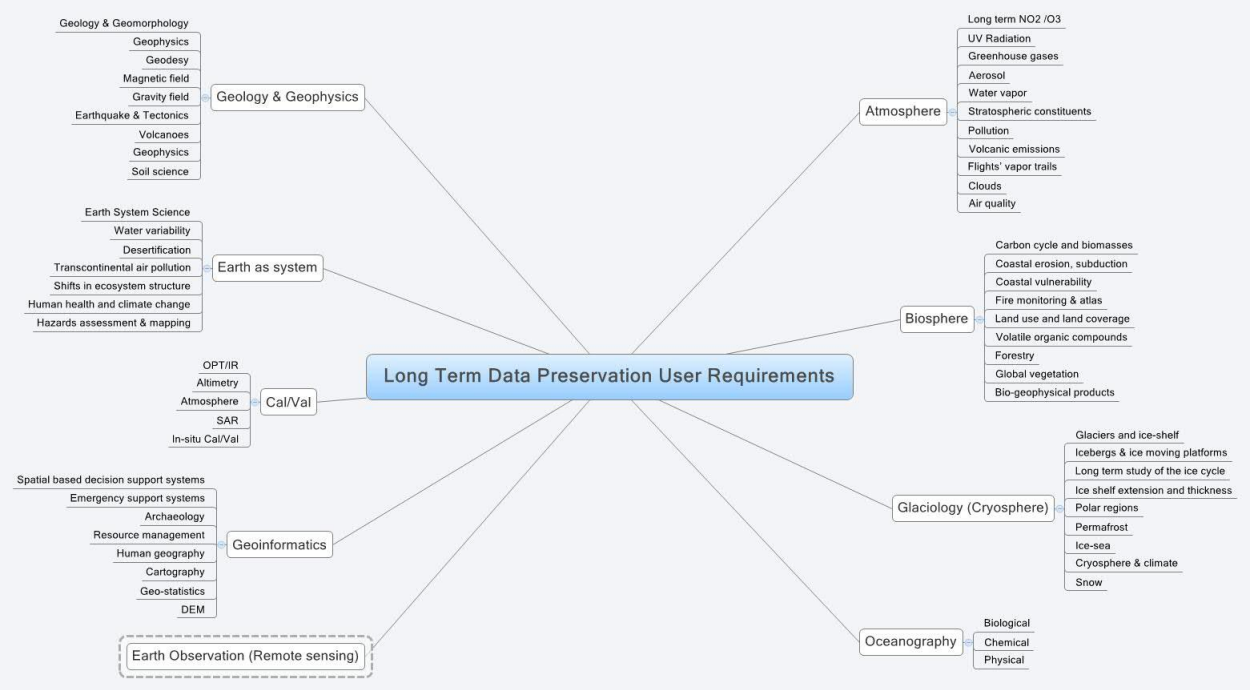

Figure 1 - LTDP User Requirements, Earth Science domains classification

For each area the preserved data sets composition is grouped as below:

1. **Primary Data**, identified as the key data produced by a sensor or instrument. Primary data can be raw or processed data.
2. **Secondary data**, identified as the set of data necessary to support the processing chain of primary data. This group includes documents, packaging information, and data necessary for corrective action, for calibration, to adapt the algorithm to specific cases, etcetera.
3. **Metadata**, being necessary to support storage, search and retrieval of data.
4. **Browse** images when generated.

User's requirements have been collected by:

1. Identifying scientific domains (Earth Science) having the need of Earth Observation data.
2. Identifying the kind of data used and necessary (data set composition, data elements).
3. Participating to events, seminars and workshops having relation with Earth Science data requirements and utilization (e.g. GMES, CCI, LIMES, Living Planet Symposium, etc.). The focus was to understand and register what kind of data and information are used (or would be necessary to use) for the purpose of scientist, service providers and value adders.
4. Performing a survey through analysis of questionnaires (one generic and one specific). Questionnaires were sent to scientific communities of practice, public organizations (e.g. Universities and Research Centers, civil protections) and value adders (e.g. commercial entities).
5. Analyzing specific needs of programmes or frameworks like: Global Monitoring Environment and Security (GMES) and its projects, Climate Change Initiative (CCI), European Strategy Forum on Research Infrastructure (ESFRI), others.
6. Interviewing researchers and investigators of different fields (e.g. forestry, oceanography, atmosphere, climatology, etc.).
7. Analyzing the needs and expectations of different communities to identify commonalities (e.g. CEOS, GEOSS, GOSS, I/Rinascimento Digitale, NASA/Planetary Data System, USA/Library of Congress, InterPARES, CERN, and others)
8. Analyzing projects having common points (e.g. CASPAR, APARSEN, PARSE-Insight, etc).

9. Analyzing existing legislation, regulations or rules of organizations having as an objective long-term data preservation or retention (e.g. OECD rules, ICSU/WDCC, ARIADNE, INSPIRE, others).

User needs and best practices were gathered, collected, analyzed, compared and elicited in order to achieve a set of user requirements.

FIRST User Requirement document was released and has been submitted for comments and ESA approval. A specific workshop has been performed in October 2010 (ref. [10] User Needs Workshop 2010). A second revision of the document incorporating comments received from the workshop participant is under review.

RESULTS

The complexity and the many different aspects concerned with preservation of digital data that must be considered have been confirmed.

Preservation involves the concept of continuity of maintenance and availability of captured data overcoming issues concerned with aging of technologies (hardware and software), of differences concerned with culture and knowledge, and assurance of trustability along the time. In some other cultures, e.g. web based archives, the process is called *curation*.

In short, preservation must grant future exploitability of today data, this implies:

1) **Identification** of all data, documents and information allowing future use and concerned links/relationships.
2) **Coherence** among all constituents/elements of the dataset.
3) **Traceability** granting back identification about sources (e.g. context, provenance).
4) **Trustability** about content (e.g. quality indicators).
5) **Useful** information, removing duplications, elements of misunderstanding or confusion.

Preserved data set composition structure (primary, secondary, metadata) has been confirmed and more details have been added.

The most important results of the study have been collected in two documents:

1. FIRST project **User Requirements Document** (**FURD**, [8]) is a document collecting all changes and requirements as follows:
    a. **Data categories** originally identified in version 1.0 of [1] (C1: high resolution SAR, C2: high resolution multispectral optic, C3: medium resolution colour optic, C4: atmospheric focused sensors, C5: other types of scientific missions and C6: other space related missions) have been refined and three new categories (i.e. C7,C8,C9) have been added:

        i. **C1 SAR S**ynthetic **A**perture **R**adar imaging missions or sensors, having high and very high resolution. Different radars and bands are considered in this category.
        ii. **C2 Optical** multi-spectral imaging missions and/or sensors, having high and very high resolutions. Examples are Landsat, SPOT, the coming Sentinel 2.
        iii. **C3 Medium** resolution Land and Ocean monitoring missions/sensors (e.g. wide swath ocean colour and surface temperature sensors, altimeter, etc). Examples are ENVISAT RA2, MERIS, AATSR.
        iv. **C4 Atmospheric** chemistry missions or sensors. Examples are the Canadian ACE (Atmospheric Chemistry Experiment), or the NASA/CNES CALIPSO.
        v. **C5 Other scientific** missions or sensors. Examples are: 3-X gravity gradiometers, GPS precision positioning, laser reflector, MIRAS (microware imaging radiometer with aperture synthesis), accelerometers, absolute scalar

magnetometer, electric field meters, vector field magnetometer, temperature and water vapour measurements, Flux gate magnetometer.

vi. **C6 Airborne** generated (e.g. digital cameras single/multiple, digital line scanners, radar, laser topographic/bathymetric, etc). Helicopter Observation Platforms (HOPs) are considered in this category.

vii. **C7: Balloon** caring of data captured by a payload of instruments carried by a balloon in a path or route along the atmosphere. Different kind of instruments can be part of the payload (e.g. geomagnetic instruments, wind, temperature, radiation, radio propagation, particles, optical properties, chemistry, etc).

viii. **C8: Ground** caring of data captured by instruments based on ground and then fixed in a specific position. Examples are: seismography, temperature, humidity, wind, pressure, radiation, radiance, pollution factors, rain, chromatography, soil property, and etcetera. Instruments often are organized as networks of similar instruments (e.g. seismographs).

ix. **C9: Hydro** caring of data captured by instruments specifically designed to capture water related information (e.g. temperature, salinity, pollution factors, wind, pressure, water flow/flux/level, etc). Here are included data coming from buoys as well as from ships or other means to capture local data.

b. **Requirements** have been organized for each data category (i.e. C1-C9) and each requirement contains its own traceability towards documents and sources of needs. There are ten sets of requirements; one is common the other nine are focused to each data category. One example is provided in the table in Figure 2 FURD [8], Example of requirement description.

c. **Instruments' classification**, the document contains a draft classification of instruments typically used in the frame of Earth Observation (see Figure 3 FURD[8], Instruments classification (draft).). This classification has the scope to support data set taxonomy and understanding of requirements.

2. **Preserved Data Set Document** (**PDSC**), the document contains the list of all elements that should be preserved. It is focused mainly on Earth Observation instruments and missions however can be easily extended to other contexts. The preservation concept and the lists of elements described in the document are based on following concepts:

   a. The lifecycle of a scientific mission is to be carefully considered. Preservation involves different elements like knowledge of the context, documents, engineering design etc... A useful preservation should start since the maiden go-ahead of the mission in order to preserve the entire knowledge gained. In other words since the conceptualization of the experiment or mission. And must be considered along the commissioning and implementation phases, during the operational phase and after the operational phase completion. PDSC considers the following stages:

      i. **Stage 1 Pre-mission**, it is the initial step when the mission is conceived and preparatory documents are created. Mainly addresses objectives of the mission, scientific layer, mission requirements, and etcetera.

      ii. **Stage 2 Mission implementation**, it represents the commissioning and implementation phase. Typically in this phase engineering problems and solutions are considered, data models are designed and qualification processes are considered.

      iii. **Stage 3 Mission operations**, it is the operations period of the mission where data are captured, used, stored and distributed. However during this phase corrective loops are in place. Those loops are concerned with maintenance of processing algorithms, maintenance of the end-to-end quality of data, with corrective actions to recover failures and the natural decay of instruments and sensors.

      iv. **Stage 4 Post mission**, it is the period when operations have been completed and utilization of data continues. Reuse of data and improvements of algorithms and processing modes are the main activities on scientific side, while data mangers should grant preservation and accessibility.

b. For each identified stages and for each defined data category a list of elements to be preserved is provided. Those lists are organized in separate list.

ID	Description	Source
FURD-CO-0010	Access to data and information including documentation for scientific purposes should be free and open.	Interviews, SOR-1 (*)
FURD- CO -0020	The owners or providers of Earth Science data and information including documentation should guarantee their preservation without limits (all forever).	Interviews, SOR-1 (*)
FURD-CO-0030	Documents must be aligned (**) with and matching to models, algorithms, procedures and data versioning (coherence of information principle).	Interviews, SOR-1, 3, 4, 5
FURD-CO-0040	Access mechanism to data should be simple, easily available, easily deployable, and economical for the user.	Interviews, SOR-1, 3, 4, 5
FURD-CO-0050	Data and information integrity, quality and reliability should be guaranteed and documented by the owner or provider.	Interviews, SOR-1, 2, 3, 4, 5, 6
FURD-CO-0060	Data, products and information should be made available on request at any time.	Interviews, SOR-1, 2, 3, 4, 5, 6
FURD-CO-0070	LTDP must define homogeneous conditions (***) of preservation. Homogeneous conditions of preservation shall be guaranteed by data providers.	Interviews, SOR-1,3
FURD-CO-0080	Information concerned with reference, provenance (****), context, fixity and access rights or conditions should be provided to the user by the data provider.	Interviews, SOR-3, 4

Figure 2 FURD [8], Example of requirement description

CAT.		Sub-cat.	Description Sub-cat.	Spatial Resolution (Range)	Bands
C1	VHR	C1_1	Very high resolution	< 1m ÷ 3m	X
		C1_2	High resolution	3m ÷ 25m	X,C,L
	MLR	C1_3	Medium resolution	25m ÷ 50m	C,L
		C1_4	Low resolution	50m ÷ >70m	C,L
C2		C2_1	Very high resolution multispectral	< 1m ÷ 8m	VIS,NIR
		C2_2	High resolution multispectral	8m ÷ 30m	VIS,NIR
		C2_3	Hyperspectral	20m ÷ 30m	VIS,NIR
		C2_4	Multispectral	30m ÷ 90m	VIS,NIR,TIR
		C2_5	Panchromatic	<1m ÷ 15m	Pan
		C2_6	SWIR	>20m ÷ 70m	SWIR
		C2_7	Vis/IR	50m ÷ 2700m	VIS,NIR,TIR
		C2_8	MW (passive)	23km ÷ 32km	K,Ka
C3	Land/Ocean	C3_1	Radiometers MultiSpectr.Imaging (MW)	1km ÷ 100km	K,Ka,Ku
		C3_2	Radiometers MultiSpectr.Imaging (Vis/IR)	100m ÷ 40km	VIS,NIR,SWIR,TIR,MWIR
		C3_3	Radiometers MultiSpectr.Sounding	48km	K-->E
		C3_4	MultiDirection/MultiPolarization	6km ÷ 47km	VIS;NIR
	Land	C3_5	Scatterometer	12km ÷ 100km	C,Ku,L
	Ocean	C3_6	Ocean colour Instrument	236m ÷ 825m	VIS,NIR
	Land/Ocean	C3_7	Altimetry	0.45m ÷ 10km	Ku,Ka
C4		C4_1	Wind scatterometers	50km	C
		C4_2	Imaging multispectral radiometers (Vis/IR)	V: 5km H: 40÷320km	UV,VIS,NIR,SWIR,TIR
		C4_3	Atmospheric chemistry	V: 1÷132km H: 32÷215km	UV,VIS,NIR,SWIR
		C4_4	Atmospheric temperature & humidity	V: 150m÷3km H: 3÷300km	TIR,SWIR, EHF
		C4_5	Multiple direction/polarization radiometers	5.5km	VIS,NIR
		C4_6	Imaging multispectral radiometers (MW)	20÷40km	K,Ka,W
		C4_7	Earth radiation budget radiometers	10km, 40km	UV,VIS,SWIR,FIR,TIR
		C4_8	LIDAR	V: 1÷2km H: 300m	UV
		C4_9	Could profile & rain radar	H: 500m	W
		C4_10	Radio Occultation Sounder for the Atmosphere	res<1K	L
C5		C5_1	3-X gravity gradiometers		
		C5_2	GPS precision positioning		
		C5_3	Laser reflector (precise orbit)		
		C5_4	Sat-to-Sat tracking		
		C5_5	MIRAS		
		C5_6	Accelerometers		
		C5_7	Absolute scalar magnetometer		
		C5_8	Electric field		
		C5_9	Vector field magnetometer		
		C5_10	Temperature and water vapour		
		C5_11	Flux gate magnetometer		

Figure 3 FURD[8], Instruments classification (draft).

CONCLUSION

The prime objective of the project is intended to help data producers and data users to set up the necessary activities to preserve data and documents.

The project has provided lists of elements that should be preserved according to the type of mission and stages of the mission evolution. This can be used twofold:

a. Like a checklist to support activities aimed at verify consistency and completeness of archived missions.
b. To support design and implementation activities in order to preserve the required elements.

The project has achieved its main goal providing requirements and preservation lists useful for practical application like assessment of the status of past missions and support to the design process of future one.

FIRST project has contributed to highlight the need for LTDP guidelines improvement. Particularly analysis related to data set composition trustability, context, and provenance, fixity and quality information, preliminarily addressed in the project is being improved and a new version of the Preserved data set composition document is under review.

AKNOWLEDGMENTS

FIRST user requirements finalization and data set composition definition was possible thanks to the participation and involvement of many experts in the fields of Earth Science.

The authors wish to thank:

- All people who were so kind replying to our questionnaires.
- All scientists who were so kind answering our questions, and providing their opinions and comments.
- All Canadian and European agencies staffs were so kind in supporting our interviews and questions.

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] - 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] - FIRST project User Requirements Document (FURD), Issue 1.0, November 2010

[9] – Preserved Data Set Composition (PDSC), LTDP-GSEG-EOPG-RD-11-0003, Issue 3.0, (under review)

[10] – European Space Agency, Long Term Data Preservation web site http://earth.esa.int/gscb/ltdp/

[11] – CCSDS Reference Model for an Open Archival Information System (OAIS) CCSDS PINK BOOK (http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/CCSDSAgency.aspx , http://nssdc.gsfc.nasa.gov/nost/isoas/ref_model.html)

[12] – OAIS ISO 14721 (http://www.iso.org/iso/)