Achievements in support of ESA's data preservation strategy
Focus on feedback from Earth Observation users

Veronica Guidetti [(1)]

[(1)] ***ESA-ESRIN***
PoBox 64, 00044 Frascati, Rome, Italy
EMail: Veronica.Guidetti@esa.int

ABSTRACT

This paper reports on the topic of long term availability of environmental data as perceived by the Earth Science community. In the context of the European strategy for preserving Earth Observation (EO) data and as partner of the EU FP7 PARSE.Insight project, ESA issued a survey-based initiative targeting its EO data user base. The timely and active participation confirmed the high interest in the addressed topic. Primary target of this action is to provide ESA teams dedicated to environmental data access, archiving and re-processing with the first insight from the Earth Science community on the preservation of space data in the long-term. ESA's Climate Change Initiative requires activities like long-term preservation, recalibration and re-processing of data records: by knowing the scientific community's standpoint about the stewardship of environmental data and the appropriateness of infrastructures for digital preservation, ESA teams working for the Climate Change Initiative and the EO User Services can take the most from actions like the one presented in this paper.

Keywords: Earth science, climate change, space data, environmental data, preservation, long term data provenance

INTRODUCTION

The objective of this paper is to report on the topic of long term availability of environmental data as perceived by the Earth Science community.
The initiative – survey based – has been advertised on the ESA Earthnet web portal [1], and comes under the umbrella of the ESA partnership in the EU FP7 project PARSE.Insight [2] and of the European strategy for preserving Earth Observation (EO) data [3].
The following is one of the many comments received by the community:

> *"I consider long-term preservation of earth satellite data to be critical and must be done at nearly any cost. I actually think it should be literally against the law to not preserve data. I believe the costs for long-term storage and continued accessibility should be coming down. ESA should step up to the plate and take care of this very important issue."*

Preservation of environmental data refers to its continued availability, accessibility and comprehensibility in the future.
This work's outcome is an overview on the scientific community's standpoint over topics like environmental data stewardship and infrastructures for digital preservation.

This initiative is hopefully the first of a series of actions aimed at closely cooperate with the scientific community, keeping it interlinked with ESA activities/achievements enabling the preservation of environmental and space data.

BACKGROUND

As exceptional source of information for protecting and preserving our environment, Earth observations are of unique value to understanding the climate change being furthermore usually impossible to resample them.
ESA's Climate Change Initiative relies on activities like long term preservation, recalibration and reprocessing of data records, algorithm development, products generation and validation.

> *"Satellite data are critical in providing the basic information for modelling and predicting climate change, The new initiative will ensure that ESA's potential in this area is fully realised"* [4]

In this context a European strategy for Long-Term EO Data Preservation (LTDP) and access has been approved by ESA Member States.
Among LTDP objectives there is that of understanding which knowledge about environmental data should be preserved, task that can be better achieved via a direct interaction with the scientific community. Also environmental non-space data are fundamental to the success of the scientific research and a further action planned by LTDP is to promote technological achievements enabling the interaction among heterogeneous information sources and procedures to assess the impact of preserving environmental data. As of the time of writing, LTDP defines a set of guidelines [5] to enable current and future EO data archiving systems to preserve data together with data processing and manipulation capabilities. This working framework acts as background to the initiative in point.

ESA's Climate Change Initiative

The ESA initiative in support of climate change [6] aims at ensuring the delivery of appropriate climate information as a direct response to the requirements emanating from the United Nations Framework Convention on Climate Change signatories. It requires activities like long-term preservation, recalibration and re-processing of data records to ensure the systematic generation of relevant Essential Climate Variables (ECVs) and their regular updating following the needs of the international climate change community.

Terrestrial EVCs are of critical importance to develop an appropriate response to global and regional climate changes. ECVs that are currently feasible for global implementation are listed in the table below [7]; others like soil-moisture, ocean-salinity, upper-air winds, and specific cloud properties are to be observed in the next years.

Domain	**Essential Climate Variables**	
Atmospheric, over land, sea and ice	**Surface**	Air temperature, Precipitation, Air pressure, Surface radiation budget, Wind speed and direction, Water vapour
	Upper air	Earth radiation budget (including solar irradiance), Upper-air temperature (including MSU radiances), Wind speed and direction, Water vapour, Cloud properties
	Composition	Carbon dioxide, Methane, Ozone, Other long-lived greenhouse gases, Aerosol properties
Oceanic	**Surface**	Sea-surface temperature, Sea-surface salinity, Sea level, Sea state, Sea ice, Current, Ocean colour (for biological activity), Carbon dioxide partial pressure
	Sub-surface	Temperature, Salinity, Current, Nutrients, Carbon, Ocean tracers, Phytoplankton
Terrestrial	River discharge, Water use, Ground water, Lake levels, Snow cover, Glaciers and ice caps, Permafrost and seasonally-frozen ground, Albedo, Land cover (including vegetation type), Fraction of absorbed photosynthetically active radiation (FAPAR), Leaf area index (LAI), Biomass, Fire disturbance	

Table 1: ECVs currently feasible for global implementation

ECVs values are derived by processing global historical time-series, while re-analysis of the archived data via models and data assimilation systems is regularly required. Furthermore, periodic re-processing of missions' basic datasets, application of newly available algorithms and CAL/VAL corrections remain critical requirements. The needs in terms of space data products are internationally defined and scientifically agreed: the challenge is then to provide consistent sets of ECVs.

ESA's Earth Science archives in brief

Describing the current state of the Earth and its environment or its variability over time requires a large number of observations. Observations from the fleet of ESA's EO satellites will be massively increasing with the operation of upcoming missions, obliging to deal with a continuous exponential growth in the volume of EO data archives [8]: if around 150TB were archived in early '90s and about 3PB are archived today, over 30PB are expected in the next 10 years.

Furthermore, a multi-mission data retrieval and distribution approach has been applied to ESA's EO data archives: the availability of data from different sources in addition to non-space data allows generating the best possible products, thus improving the development of ECVs and their values.

ESA's Earth Science archives date back to 1975, time-spanning from a few years to decades and their content represents valuable scientific time-series for a large number of Earth Science applications.

Requirement for an enhanced Infrastructure

Regardless of an increasing capacity, resources are limited, i.e. there are restrictions on the volume and quality of data that can be archived. On the other hand, the relentless growth of EO archives' volume imposes infrastructure upgrades on more performing storage media, data management and cataloguing, robotics, data security and computing systems, including adaptation to new standards.

Furthermore, an adequate infrastructure technology, based on very high performance computing and distributed architecture, would eliminate resource conflicts arising between the operations of existing missions and the required systematic re-processing of entire archives for study on climate change and provision of consistent ECVs' sets. It would enable full re-processing of global data on timescales of weeks instead of years, dramatically reducing the time for production and delivery of EO data products, while it should allow easy plug in of newly developed algorithms and easy access to auxiliary data and information.

Actions to secure easy access to global time-series, and to assure their long-term preservation are essential for generating consistent ECVs, as do related scientific and technical knowledge, improved algorithms, auxiliary data, and technical facilities for recovering, re-processing and calibrating the archived data.

INSIGHT IN THE EARTH SCIENCE COMMUNITY

Foundation activity of PARSE.Insight is to provide an insight on the topic of long-term availability of research data in user communities of selected disciplines like Earth Science.

In the context of PARSE.Insight, ESA opened a public consultation in May 2009 with the primary target to get an insight on the current and envisaged exploitation of historical environmental data streams by the scientific community, basing on the principle that the current value of environmental data streams can hardly be estimated while its potential future uses hardly foreseen, i.e. that technology advances or changes in user needs may turn data considered to be of little use valuable.

Such initiative was intended to provide useful input to LTDP developments and to depict a scientific community's standpoint over topics like environmental data stewardship and infrastructures enabling long-term data provenance/community access to distributed data.

Participants were asked to answer to ten multiple-choice questions with the possibility to add their own free text comments.

Who attended the initiative

The majority of participants belong to the ESA EO data user base, which is mainly composed of scientists and researchers working in research institutes, academia, national authorities, and international (non-)governmental organisations.

Representatives from industry and the commercial sector also actively participated, including students and the general public, who is being showing an increased interest in EO applications to day life and environment protection.

Respondents work in 64 different countries, among which EU (50%), United States (9%), Canada (5%), China (4%), Africa (3%), South America (2%), Australia, India and Russian Federation (1% each) and South East Asia.

Participants mainly have scientific background and work on research tasks for national and international programmes. Computer scientists, people working on operations and/or in the industry/commercial sector (20%), responsible for education and capacity building are present too.

Attendees' profile examples include technical directors of national data centres / remote sensing centres, representatives of international programmes, European Commission representatives, representatives from international (non-)governmental organisations, and computer scientists. The graph below illustrates the participants' spread among Earth Science areas.

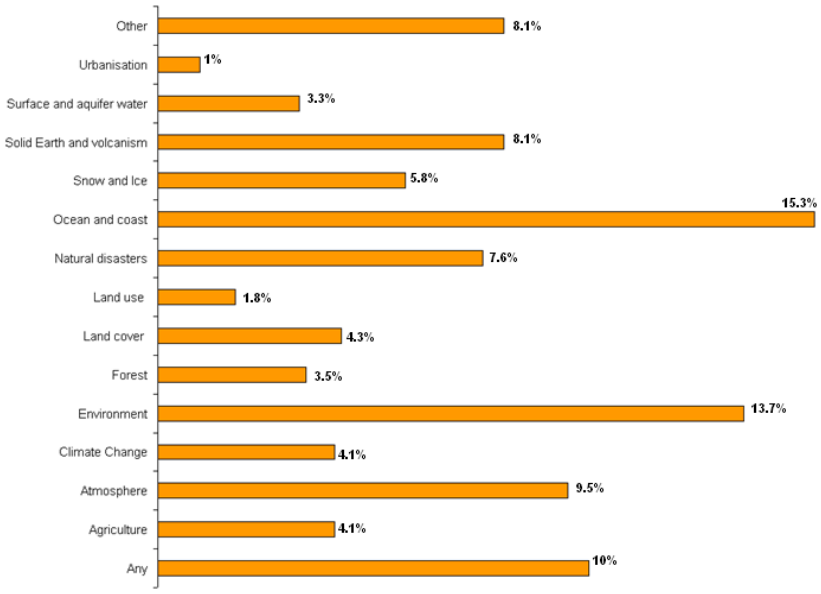

Figure 1 - Respondents' spread among Earth Science applications

Respondents have detailed knowledge of the data they use; they critically evaluate data, integrating them from different sources. They tend to focus on one particular application or type of data though they have extensive knowledge of the different data types available.

More than 20% are CAL/VAL users and around the 35% declared to work with time critical applications. Microwave remote sensing with focus on SAR data resulted to be the most common working area. Quoted radar instruments come both from ESA missions and ESA Third Party missions.

The 13% of respondents stated to be new to the topic of data preservation and some asked for good examples to show the importance of archiving EO data in the long-term.

Instead, those aware of it left a number of comments on its importance, like:

> *"[…] I often use data at large time intervals for change detection and trend analysis purposes: long-term preservation is a must, otherwise long-term variability is not detectable!"*
>
> *"In research we are often required to describe development or trends. Then we need to be able to 'go back'."*
>
> *"I deem data preservation also important for BOTH time-critical (near-real-time) and non-real-time applications"*

The high majority declared to constantly need to access historical environmental data while, anyway, more than the 40% declared to access them from time to time.

As per the graphic below, the lack of sustainable hardware, software or support of computer environment and the eventual loss of the data custodian are considered *very important* threats to data preservation. Users' inability to understand or use the data and the uncertainty of data's origin and authenticity are considered as *important* threats by the majority.

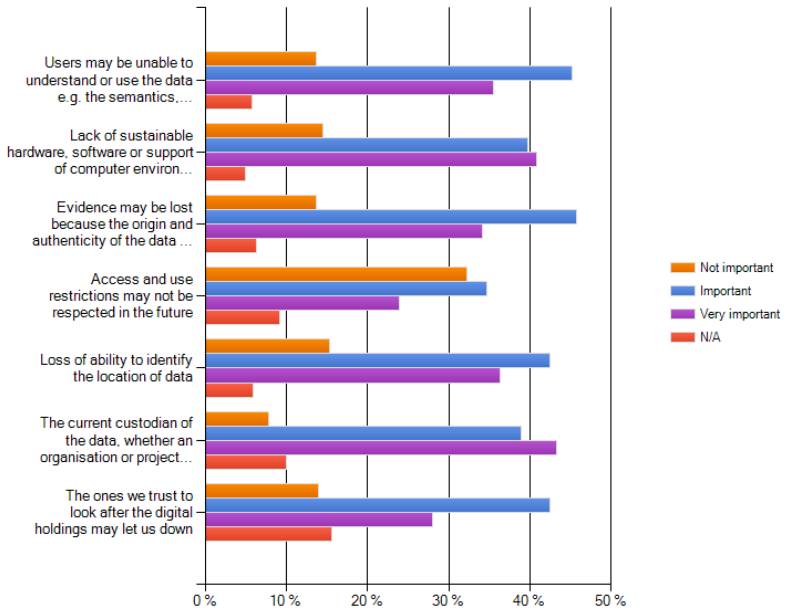

Figure 3 - Responses to "How important do you regard the following threats to environmental data preservation over the next 10 years"

"Historical experiences"

Around the 84% of respondents answered positively to the question "Have you ever required access to historical environmental data?" and detailed their own experiences.

Analysing their feedback, access to historical data was required for the missions/sensors and data types reported in the table below together with the applications. Some direct experiences:

> *"I have used altimetry data from 1992 and tide gauge data from 1950. Long time series are the basis for understanding climate change."*

"In building a climate data record of sea level I am constantly faced with unearthing altimeter data and auxiliary data from missions long deceased. Old FTP archives / web pages, etc. disappear, and basic knowledge about the available data withers."

Mission/Sensor	**Applications**
ENVISAT/AATSR,ASAR, GOME,MERIS, SCHIAMACHY/L0-L1 data	Measurement of slow deformation rates around earthquake faults; Analysis of tectonic loading of faults or the understanding of landslide history; Mapping of historical land use/land cover; Vegetation phenology; Forest health development; Tracing gases for data assimilation, time series analysis, climate model evaluation, trend analysis and for ascertaining seasonal changes and background levels of atmospheric pollutants; Coast line monitoring over decades; Variability of oceanographic conditions and productivity; Use of time series of water quality, plankton development, seasonal aspects; Study of solar variability and stratospheric ozone; CAL/VAL processing; Interferometry
ERS1/ERS2, SLC	
MSG	
PROBA/CHRIS	
Landsat(s)/MSS from 70s	
Terra-Aqua/MODIS	
NOAA/AVHRR	
SEASAT 1978	
RADARSAT Constellation	
QuickBird	
SPOT(s)	
Atmospheric, ocean colour, SST, LAI, NDVI time series data, in-situ data, LR, MR and HR data	

Table 2: Missions/Sensors data and applications for which access to historical data series has been reported

Occasions of data losses are reported by the 25% of respondents and mainly concern ERS1/ERS2, Landsat, Geosat, ENVISAT/ASAR-MERIS, ATSR-1/2, SeaStar/SeaWiFS, and SEASAT 1978 missions and sensor data, including in-situ data. Found reasons behind data losses include:

- Aging and degraded/damaged media, hard disks failures and inadequate backups, failures in storage and data transfer;
- Lost media, equipment to read data files no longer existing, disappearance of old cassette reader devices, files corruption;
- Change of data carrier formats, data format not readable anymore and metadata on data format lost; lack of metadata, no re-use of data archive due to change in processing software;
- Responsible staff not involved anymore.

According to some participant's report, it is not needed to go back a long time span (decades) to not be able to find the data: even a period of 5 to 10 years back may be enough to find it impossible to access them. Further comments on data losses are reported below.

"We currently handle data from a number of institutions and agencies. In some cases, there have been hard disk failures at their archives, leading to delays in the delivery of the requested data until it was possible to restore it. But other times, the data received was missing or unreadable, and it was determined that the original file at the archive was corrupted. […] data couldn't be retrieved and we ended with "gaps" in the generated results."
"SAR data from the Shuttle Imaging Radar mission C (SIR-C) that flew on the Shuttle twice in 1994 is effectively lost because the computer hardware to read and process the original data is obsolete and no new system has been created to read the data."

What to do with historical environmental data

According to respondents, scientific research, climate change study, and disaster evaluation can profit at most from the availability of historical environmental data. The majority of respondents further comment their choice illustrating a large spectrum of envisaged exploitation of those data.

Stress is on applications like ocean monitoring, oceanography and habitat mapping of marine species (including commercial fishes), coastal management, maritime safety, monitoring of natural waters quality, hydrology and water reserves, and oceans acidification.

Others refer to land subsidence prediction, urban management, land deformation such as tectonic, measurements of displacements to enable risk assessment, mapping of inter-seismic surface deformation along major faults systems, support to earthquake and volcano research.

Assessment of renewable energies availability, atmospheric pollution trends, policy making and civil security are also addressed.

> *"The added values of historical data compiled in standard formats and publicly available is much higher than at this time known."*
> *"Historical EO data is/will be of priceless value to most/lots of research and application fields."*
> *"Historical data are required for algorithm development, environmental monitoring, and climate change studies. Historical data underpins near-real environmental monitoring"*

Which environmental data to preserve

Participants have been given the following definition of 'environmental data': all types of environmental Earth observations (physical samples, in-situ and remotely sensed data), i.e. model output and synthesized products derived from these data, products developed through interpretation of original or synthesized products, and products whose quality is based on experimental capabilities/algorithms.

The users have been requested to select among environmental archives of interest according to thematic category, information typology and products typology.

The majority of the respondents expressed their interest for historical archives containing original data and optical/multispectral radiometry products, related to the land theme. Archives about oceans, air and cryosphere, containing synthesized, interpreted or experimental products and concerning radar imagery, atmospheric data and radar altimetry follow in priority.

Participants also remarked on the essentiality of having liberal data distribution policies and long term commitment for each mission, to enable operational data exploitation. This comment is in-line with the argumentation on data sharing principles on-going at EC, GEO/GEOSS and (inter-)national space agencies levels.

About Metadata

The preservation of metadata and data is closely related to the adoption of metadata standards.

Discipline dependent, metadata are essential for data discovery, access and integration. Their expansion and/or mappings among discipline-specific standards are critical to address multidisciplinary problems. The above experiences on data losses confirm the relevance and lack at the same time of proper metadata and metadata standards.

Around the 24% of respondents declare to not be familiar with any metadata standard or guideline. Among the rest, more than 54% is familiar with HDF followed by one third for netCDF.

The 20% is familiar with the ISO 19xxx series for spatial metadata definition and the INSPIRE directive [9]; the 14% knows the specifications for spatial data services from OGC [10]. Out of this set, all the other choices are represented between 1% and 6%.

"We need absolutely to align earth science data to the scientific standards like netcdf, hdf, or better gml and stop wasting time and money for developing proprietary format not very well exportable, neither portable."

How to access historical environmental data

The consultation included one question regarding the opportunity of having at disposal infrastructures enabling preservation of environmental data for the long-term.

Respondents are aware that making available and maintaining historical environmental data series means to reliably store them, to enable easy access both to data and knowledge, including applied processing, metadata, auxiliary data, and traceability and to make their analysis and re-processing possible. The community is highly aware about the need of adequate infrastructures for production and delivery of EO data products, based on high performance/on-demand/distributed computing technology and highly reliable storage systems.

Participants agree that an infrastructure for data preservation would impact their daily work (6% do not think so/do not know) and motivate their choices. They also provide examples of running infrastructures. Some of the users' comments are:

"[…] large reprocessing in a GRID environment. The objective is to give users an operational access to the needed data for its own application with no failure."
"Although GOME-1 is continuing to provide data after June2003, resources are limited or do not exist for producing and in particular recalibrating the data. In general the calibration of the data both pre flight and in orbit, in particular dealing with degradation, has a high cost. The agencies in general do not have a sufficient budget for such matters."
"Back in the 1990s when high volume data storage was a difficulty we lost about 2 years of 1980-1981 data through deterioration of reel-to-reel magnetic tapes which became unreadable and shredded when mounted in the drive. It was just a case of the substrate tape deteriorating with age."

CONCLUSIONS

The rate of attendance and the active participation to the presented initiative were unpredictable and confirmed a high interest by the Earth Science community in the addressed topic. The 65% of the respondents left their names to be contacted back about developments and follow-up actions in the field. Furthermore, it is worth to consider the worldwide provenance of the answers, far beyond ESA Member States.

The main achievement reached is the establishment of a link with the worldwide Earth Science community on the topic of environmental and space data preservation, then the following points emerge:

- The scientific community needs and wants to access historical environmental data and historical time series of Earth observations, for the most disparate applications across Earth Science;
- The community wants to enhance its experiences on historical data exploitation aiming at a more active involvement in the process, e.g. by reporting examples and suggestions to foster data availability and accessibility;
- Users are aware and informed about current infrastructures' limitations to enable data availability and accessibility, and ask for timely and effective solutions.

ESA teams working for the EO User Services and the Climate Change Initiative can take the most from a closer cooperation with the scientific community basing on actions like the one presented. The result depicted in this paper will under go the analysis of ESA teams dedicated to environmental data access, archiving and re-processing.

The presented initiative contributes to the recommendations by international fora like GEOSS [11] and GCOS [12] to promote the long term custody of satellite data records and quality metadata and the open access to them granted to a wider and aware public. Finally, it contributes to reflect the policy of ESA to develop, utilise, innovate and expand technologies, systems and scientific understanding.

ACKNOWLEDGEMENTS

Many thanks go to the Earth Science community, who actively and timely responded to the initiative in point, providing valuable contributions. Special acknowledgments go to the ESA EO User Services team that made this initiative possible and to the author's colleagues Maria Longo and Vincenzo Beruti for reviewing this paper.

REFERENCES

[1] - http://earth.esa.int/

[2] - http://www.parse-insight.eu/

[3] - http://earth.esa.int/gscb/ltdp/

[4] - http://www.esa.int/esaEO/SEMUX6NKRGF_index_0.html

[5] - http://earth.esa.int/gscb/ltdp/scopeLTDP.html

[6] - http://earth.esa.int/workshops/esa_cci/ESA_CCI_Description.pdf

[7] - http://www.wmo.int/pages/prog/gcos/index.php?name=EssentialClimateVariables

[8] - http://earth.esa.int/missions/

[9] - http://inspire.jrc.ec.europa.eu/

[10] - http://www.opengeospatial.org/

[11] - http://earthobservations.org/geoss.shtml

[12] - http://www.wmo.ch/pages/prog/gcos/index.php?name=AboutGCOS