Characterisation of metadata to enable high quality climate applications and services

Deliverable D600.2

Alignment of CHARMe with SAFE and LTDP

Partners providing input: Infoterra, STFC, CGI

version number	date	comments
0.1	04/10/2013	First draft
0.2	09/10/2013	Second draft, merged sections on SAFE and LTDP
0.3	05/12/2013	Third draft including STFC and CGI inputs
0.4	06/12/2013	Final draft prior to review
1.0	18/12/2013	First release

CHARMe is funded by the EC under its FP7 Research Programme

Executive Summary

The Standard Archive Format for Europe (SAFE) was developed by the European Space Agency (ESA) in the wider context of the Long Term Data Preservation Programme (LTDP) for Earth Observation (EO) products. Following a description of the development and characteristics of SAFE, the inclusion of commentary metadata in SAFE is considered from the viewpoints of its desirability, technical compatibility and governance.

In terms of desirability, SAFE is intended as an archive information package for storing individual products at low processing levels. It is likely that a significant amount of commentary from non-EO specialists and users will be related to whole data types (not individual products) and / or to aspects reflecting the higher level processing and algorithms applied. Governance is also problematic for SAFE as preliminary discussion with the LTDP Working Group indicate serious reservations about letting third parties write information directly/physically to SAFE packages. From the technical standpoint, if SAFE were to be integrated with CHARMe then it should be via the inclusion of references to C-metadata annotations from the metadata section of the SAFE Manifest file.

Whilst the current and intended use of SAFE is a poor match for the purpose of storing C-metadata or linking to CHARMe, this is not the case in the wider LTDP context. The EO LTDP Framework is intended to promote the preservation of all European EO datasets. In this context, long-term data preservation does not mean solely preserving the data as bytes on tape or disk, using the latest technologies. It is also fundamentally concerned with preserving the knowledge about this data. C-metadata clearly complements the aims of preserving knowledge about data and, by providing context on its 'fitness-for-purpose', encouraging appropriate exploitation.

There are two possible approaches for adding the C-metadata to the existing metadata to be preserved for the long-term:

1. A copy of the C-metadata relevant to the datasets in the long-term archive could be harvested and stored within the long-term archive in addition to the existing metadata.
2. The long-term archive could maintain a link to a CHARMe node storing the relevant C-metadata.

The second approach is preferred in that all that is required is to maintain the linkage from the LTDP archive to a CHARMe node (C-node). However, this also implies a need to maintain the availability of the C-node and ideally duplicated C-nodes to address LTDP Guideline 5.4 which recommends maintenance of back-up copies of all archived information.

LTDP Guideline 6.1 concerns the discovery of preserved information. In this context a CHARMe node would constitute the metadata catalogue for the C-metadata and the CHARME plugin would provide the required data access search capabilities. Thus the CHARMe implementation is already consistent with the LTDP guideline so long as the CHARMe plugin can be integrated with LTDP archive discovery services.

While C-metadata is very important to give context to data in Long Term Data Preservation, directly connecting C-metadata to individual data products via SAFE is not considered an appropriate mechanism. It is recommended that the metadata detailed in the existing European LTDP Common Guidelines be extended to include C-metadata from CHARMe. The best forum for discussing the possible inclusion of CHARMe within the LTDP framework is the LTDP Working Group, which currently consists of representatives from various space agencies, the UK Space Agency being represented by CHARMe partner STFC. A plan for the promotion of CHARMe within LTDP should be drawn up for 2014.

Contents

1. Introduction

1.1. Document scope

This document was conceived to examine the inclusion of CHARMe commentary metadata in Standard Archive Format for Europe (SAFE) developed by the European Space Agency (ESA) in the wider context of the Long Term Data Preservation Programme (LTDP). An implicit assumption was that such an inclusion of commentary metadata (C-metadata) in SAFE products was both appropriate and desirable. The scope of the document has been widened slightly to challenge these assumptions and to consider the broader relationship between, and alignment of, LTDP and CHARMe by means other than through physical integration of commentary metadata into SAFE products and product specifications.

It should also be noted that consideration of SAFE as an appropriate package for long term archiving and preservation of C-metadata is outside the scope of this document.

1.2. Document structure

The document structure is as follows:

- Chapter 1 provides a brief introduction to the document scope, structure, relevant references, acronyms and a glossary of terms.
- Chapter 2 introduces SAFE and provides an overview of the SAFE architecture, before considering the potential use of CHARMe within SAFE.
- Chapter 3 discusses the LTDP programme and analyses the implications of CHARMe for the LTDP framework and guidelines.
- Chapter 4 summarises the conclusions of the analysis and considers the way ahead for the alignment of CHARMe with SAFE and LTDP.

1.3. References

RD-01 Long Term Preservation of Earth Observation Space Data, European Common LTDP Guidelines, (v2.0), http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue2.0.pdf, 30th June 2012

RD-02 European Earth Observation Data Set, http://earth.esa.int/gscb/ltdp/European-Data-Set-Issue2.0.pdf

RD-03 Producer-Archive Interface Methodology Abstract Standard (PAIMAS), http://public.ccsds.org/publications/archive/651x0m1.pdf

RD-04 Producer-Archive Interface Specification (PAIS), http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206511R1/Attachments/651x1r1.pdf

1.4. Acronyms and abbreviations

AIP	Archival Information Package
API	Application Program Interface
ASI	Agenzia_Spaziale_Italiana (Italian Space Agency)
CCSDS	Consultative Committee for Space Data Systems
CHARMe	CHARacterisation of Metadata to enable high quality climate applications and services
CNES	Centre National d'Études Spatiales (National Centre for Space Studies)
CSA	Canadian Space Agency
C-metadata	Commentary metadata
C-node	CHARMe node, a repository and server of C-metadata
DIP	Dissemination Information Package
DLR	Deutschen zentrums für Luft-und Raumfahrt (German Aeropace Centre)
DOSTAG	Data Operation Scientific and Technical Advisory Group
EO	Earth Observation
ESA	European Space Agency
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
FIRST	deFInition of LTDP user Requirements and preservation data SeT composition
HARM	Historical Archives Rationalization and Management
ISO	International Standards Organisation
L0	Level 0 EO data products are reconstructed, unprocessed instrument/payload data at full resolution; any with all communications artefacts (synchronization frames, duplicate data , communications headers, etc.) removed.
LAST	Long term data Archive Study on new Technologies
LTDP	Long Term Data Preservation programme of ESA
LTDP-DQ	Long Term Data Preservation – Data Quality
OAIS RM	Open Archival Information System Reference Model
PAIMAS	Producer-Archive Interface Methodology Abstract Standard
PAIS	Producer-Archive Interface Specification
PB-EO	Programme Board on Earth Observation
PDGS	Payload Data Ground Segment
PDI	Preservation Description Information
PDR	Preliminary Design Review
SAFE	Standard Archive Format for Europe
SCIDIP-ES	SCIence Data Infrastructure for Preservation with focus on Earth Science
SDF	Structured Data File
SIP	Submission Information Package
STFC	Science and Technology Facilities Council

UKSA	United Kingdom Space Agency
XFDU	XML Formatted Data Units
XML	eXtensible Mark-up Language

1.5. Glossary

Component	An object (i.e. file) that can be grouped together to be part of a Package.
Fixity Information	The information that documents the authentication mechanisms and provides authentication keys to ensure that the Content Information object has not been altered in an undocumented manner.
Package Exchange File	A collection of files that have been bundled together into a single container that also contains a Manifest describing the contained files and the relationships among those files.
Representation Information	The information that maps a Data Object into more meaningful concepts. An example is the ASCII definition that describes how a sequence of bits (i.e., a Data Object) is mapped into a symbol.
SAFE Product	Equivalent to an *XFDU Package*.
XFDU Package	A *Package Interchange File* that contains an *XFDU Manifest* and is conformant to the semantics specified in the XFDU Specifications.
XFDU Manifest	Is a document, conforming to the XML Schema specified in XFDU Specifications, that contains metadata about *Component*s and the relationships among them. The XFDU Manifest is itself a *Component*.

2. SAFE

2.1 Background

Since 1975 the European Space Agency has managed data downlinked from satellites in a number of formats which have been mission or even instrument specific. With new missions the heterogeneity, as well as the data volumes, of archived data has increased dramatically; data format technology has continuously improved and the very means of data dissemination has shifted from physical media to electronic transfer and hosted processing.

In recognition of these challenges, in 2004 the EO Ground Segment Department at ESA ESRIN set up the Historical Archives Rationalization and Management (HARM) project with the principal objective of defining how to convert the historical datasets into a new modern format. This format was to be based on the latest technologies and standards and able to ensure the long-term preservation of ESA data holdings. The format developed by the HARM project was named Standard Archive Format for Europe (SAFE).

safe Standard Archive Format for Europe

SAFE builds on the information package concept of the broader Open Archival Information System (OAIS) Reference Model. OAIS was developed under the auspices of the Consultative Committee on Space Data Systems (CCSDS) which is an organization of Space Agencies producing recommendations and standards mainly for ground systems and their interface to space systems. OAIS has been adopted by the International Standards Organisation (ISO) as the ISO 14721:2012 standard.

The OAIS Reference Model encompasses data preservation from producer to consumer through the movement, management and preservation of three types of packaged information (see also Figure 1):

- Submission Information Package (SIP) from producer to archive;
- Archive Information Package (AIP) within the archive;
- Dissemination Information Package (DIP) for dissemination of information to the users.

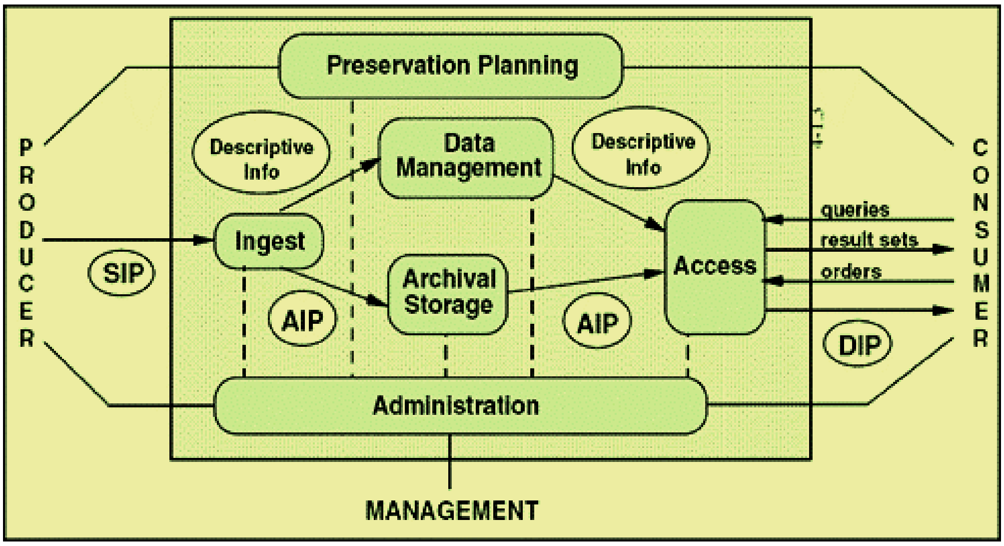

Figure 1 – the OAIS Reference Model for long term preservation of digital information

It should be noted that SAFE was primarily developed as an Archival Information Package enabling the long term data preservations of Level 0 (L0) EO data. In the situation where data are distributed directly from the archive, it is normal practice that they are processed to a higher level and converted into a format more oriented to the end-user utilization. That said, SAFE can be used to bundle higher level EO products for dissemination to users and is the adopted delivery format for Level 1 data from the Sentinel-1 mission.

CCSDS has also defined an implementation recommendation for OAIS information packages in the form of the XFDU, XML Formatted Data Units, specification. This specification (available at http://sindbad.gsfc.nasa.gov/xfdu) provides a namespace schema definition for XFDU, Application Program Interface (API) tool libraries and structure rules. Physically, XFDU enables the encapsulation of data and metadata, including software, into a single package (e.g. file or message) to facilitate information transfer and archiving.

2.2 Architecture

The physical model for a SAFE Product is illustrated in Figure 2.

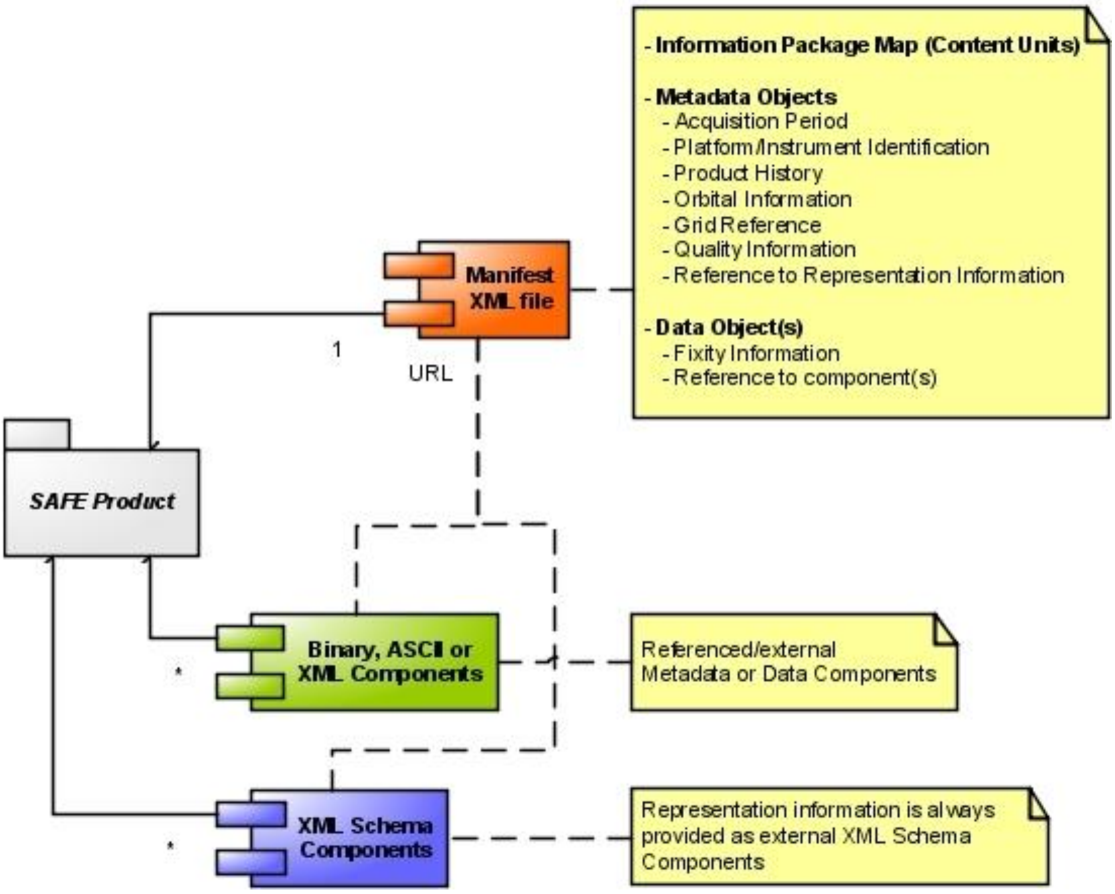

Figure 2 – the physical model for SAFE.

A ***Component*** is an object (i.e. file) that can be grouped together to be part of a Package.

The three forms of component are:

- ***Binary, ASCII or XML Components***: includes the data objects stored as discrete files. Each Data Object should be described by one or more Metadata Objects. The Metadata Objects may be distinct components, files in ASCII or XML format, or may be wrapped in the *Manifest Component*.

- ***Manifest Component***: an XML document conforming to the XFDU Manifest file specification. It contains the Information Package Map, Metadata Section and Data Object Section.
- ***XML Schema Components****:* discrete files containing the representation information for the data held by a SAFE product. XML Scheme files are considered to be Metadata Objects and are referenced from the *Manifest Component*.

The informationPackageMap element of the Manifest Component contains the top-level hierarchical view of the SAFE Product. The metadataSection element of the Manifest contains all of the metadataObject elements. A metadataObject element may either wrap a Metadata Object directly or simply include a reference to the Metadata Object, which is stored as a separate Component in the package. The dataObject elements of the dataObjectSection normally include a reference to a Data Object, which is stored as a separate Component in the package (i.e. Data Objects are rarely embedded in the SAFE Manifest file).

The example provided below shows the top-level elements of the Manifest Component of a Sentinel-1 SAFE file:

```
<xfdu:XFDU version="esa/safe/1.2/sentinel-1/sar/level-1/grd/standard/smdp"
           xmlns:xfdu="urn:ccsds:schema:xfdu:1">
   <informationPackageMap>
   <metadataSection>
   <dataObjectSection>
</xfdu:XFDU>
```

Note the use of the XFDU namespace to define the types for the elements used in the SAFE Manifest Component.

In the context of storing Commentary metadata (C-metadata) in SAFE, it is worth exploring the metadata section of the Manifest Component in more detail using the two metadataObject elements shown below:

```
   <metadataSection>
      <metadataObject ID="mapoverlayAnnotation"
                      classification="DESCRIPTION"
                      category="DMD">
         <dataObjectPointer dataObjectID="mapoverlay"/>
      </metadataObject>
      …….
      <metadataObject ID="acquisitionPeriod"
                      classification="DESCRIPTION"
                      category="DMD">
         <metadataWrap mimeType="text/xml"
                       vocabularyName="SAFE"
                       textInfo="Acquisition Period">
            <xmlData>
               <acquisitionPeriod xmlns="http://www.esa.int/safe/1.2">
                  <startTime>2012-01-01T04:37:24.406908</startTime>
                  <stopTime>2012-01-01T04:37:33.556671</stopTime>
                  <startTimeANX>261214.512000</startTimeANX>
                  <stopTimeANX>270364.275000</stopTimeANX>
               </acquisitionPeriod>
            </xmlData>
         </metadataWrap>
      </metadataObject>
      …….
      …….
   </metadataSection>
```

Both of these metadataObject elements are categorised as Descriptive MetaData (DMD) and classified as "DESCRIPTION". They differ in how the description metadata is presented. The `acquisitionPeriod` metadataObject encloses the metadata directly in a `metadataWrap` element. Note the use of the ‘SAFE’ vocabulary. The `mapoverlayAnnotation` metadataObject, by way of contrast, uses a `dataObjectPointer` element to point to (reference) a component (file) in the SAFE product so that the descriptive metadata lies outside the Manifest Component.

SAFE directly supports the description of provenance information under the category of Preservation Description Information (PDI) as illustrated in the following example of a metadataObject:

```
<metadataObject ID="processing"
                classification="PROVENANCE"
                category="PDI">
   <metadataWrap mimeType="text/xml"
                 vocabularyName="SAFE"
                 textInfo="Processing">
      <xmlData>
         <processing name="Post Processing"
                     start="2012-08-09T16:06:49.999864"
                     stop="2012-08-09T16:09:23.000000"
                     xmlns="http://www.esa.int/safe/1.2">
            <facility country="Italy"
                      name="ESRIN headquarters"
                      organisation="ESRIN"
                      site="Rome">
               <software name="Sentinel-1 IPF" version="001.10"/>
            </facility>
            <resource name="S1A_S6_SL1__1_DV_20120101T043724.SAFE"
                      role="Level-1 Intermediate Product">
               <processing name="SLC Processing"
                           start="2012-08-09T16:07:10.000000"
                           stop="2012-08-09T16:07:45.000000">
                     …….
                  <resource
                        name="S1A_S6_RAW__0SDV_20120101T043724.SAFE"
                        role="Level-0 Product">
                     <processing name="Generation of Test Sentinel-1
                                       SAFE SAR Standard L0 Product
                                       for SM Swath 6 (dataset
                                       S1S-PTRG-S6-1)"
                                 start="2011-05-15T12:30:00.000000Z"
                                 stop="2011-05-15T14:12:09.000000Z">
                           …….
                     </processing>
                  </resource>
               </processing>
            </resource>
         </processing>
      </xmlData>
   </metadataWrap>
</metadataObject>
```

Note how the processing steps are nested with the final (Post-Processing) step utilising the output of the previous processing step (SLC-1 Processing) as its resource name. As an aside, this approach contrasts to the ISO 19115 (2009) standard for geospatial data. ISO 19115 records provenance as a series of processing steps (`LE_ProcessStep` elements), each of which encapsulates a link to the data source description and a description of the processing applied.

2.3 Use of CHARMe in SAFE

Any discussion of how C-metadata or the CHARMe architecture might be linked to SAFE products needs to first examine if such a linkage is desirable or appropriate.

From a usage point of view, the motivations for, and purposes of, SAFE and CHARMe are poorly aligned:

- SAFE was primarily developed as an Archival Information Package enabling the long term data preservations of Level 0 (L0) EO data. Many issues worthy of comment will clearly trace back to the Level 0 data but equally, there will undoubtedly be commentary that relate to the algorithms and processing of higher level (Level 2-4) products. These higher level products are generally of more immediate interest to the wide range of non-EO scientists interested in, and make, commentary.
- In an EO context, most (though not all) commentary metadata is likely to be annotated to instruments, product types and dataset collection targets rather than to individual EO products. SAFE is not intended to, nor designed to, support the aggregation of datasets at the collection (dataset series) level.

From a technical perspective there are a number of ways in which C-metadata might be linked to a SAFE Product. The C-metadata could be:

- Encapsulated directly in the Manifest Component.
- Stored in the SAFE Product as a separate XML Component (file), referenced from the metadata section of the Manifest.
- Referenced (for example via a URI or DOI) from the metadata section of the Manifest.

In terms of governance, the intention to use SAFE for production and archiving by data producers is somewhat at odds with the use of CHARMe to enable third party and *ad hoc* C-metadata annotation. From preliminary discussion with the LTDP Working Group there would be serious reservations about letting third parties write information directly/physically to SAFE packages. In addition, current LTDP activities are heavily biased towards supporting highly formalised deposits to the archive. This would only be compatible with the CHARMe metadata being physically stored within the SAFE package by the archive operator by periodically aggregating C-metadata and updating the relevant SAF products.

In summary:

1. the current and intended use of SAFE is a poor match for the purpose of storing C-metadata or linking to CHARMe;
2. if SAFE were to be integrated with CHARMe then it should be via the inclusion of references to C-metadata annotations from the metadata section of the Manifest.

3. Wider implications of CHARMe for LTDP

3.1 History

In 2001, the ESA Programme Board on Earth Observation (PB-EO) endorsed a strategy for the "Management of Historical Archives" (ref. PB-EO (2001)4). The ESA technical advisory board for PB-EO, the Data Operation Scientific and Technical Advisory Group (DOSTAG), subsequently produced document ESA/PB-EO/DOSTAG (2003)6 promoting access to across-missions products and services and the exploitation of historical archives. The profile of historic archive management was further raised in 2003 via the Oxygen implementation plan (ESA/PB-EO (2003)51). This reinforced and enhanced, in a wider context, the overall issue of data archives and improved user access to, and the availability of, source data. Finally, in 2004 the *ad-hoc* nominated PB-EO Ground Segment Task Force concluded its report with a set of recommendations (ESA/PB-EO (2004)53) for the facilities and archives infrastructure aiming at:

- A maximized competitive approach;
- Enhanced infrastructure exploitation;
- Facilities and operations rationalization;
- Technology exploitation;
- Cost reduction;
- Possible standardization and re-utilization at European level of Agency investment.

The Earth Observation Department of ESA has since recognized that the main process to be undertaken is the standardization and harmonization of its ground segment architectures to reach economies of scale during development, operations and maintenance. This includes the need to achieve the following goals:

- Archive maintenance in order to ensure data integrity;
- Archive and data management rationalization;
- Data conversion to new technologies in order to reduce cost of operations;
- Enhancement of data access;
- Standardization of the format in which datasets are preserved.

Achieving these goals will also pave the way for simplifying data exchange and interoperability between ESA and external operators.

It is well known that special attention is required for data to be preserved over the long term and this can be reflected in costly operations for maintenance and exploitation. The identified challenges include:

- The datasets have to be regularly transcribed to new media technology, to prevent the problems created by their obsolescence;
- Since the long term archive is normally based on datasets archived usually only up to a very low processing level, normally called Level 0, higher level products have to be generated by processing systems on demand;
- In addition, in the case of distribution of the data holdings directly in the archived format, it is normal practice that they are converted or reformatted into a format more oriented to the end-user utilization;
- It is a common requirement to have to extract from the long term archive a portion of the single data file (subsetting) to create a "child" product, optionally to be pre-processed;

- Finally, more and more data are distributed to end-users and exchanged among data holders in electronic format over network infrastructure (private Intranets, public Internet, academic networks, etc.).

3.2 European LTDP Common Guidelines

In 2006, ESA initiated a programme among all the European, as well as Canadian, stakeholders to develop a common approach for the long-term preservation of Earth Observation (EO) space data. The Agency consulted with the Member States in 2007 to present an EO Long Term Data Preservation (LTDP) strategy for the:

> "preservation of all European (including Canadian) 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 among the EO space data owners."

An LTDP Working Group was formed at the end of 2007 to liaise with representatives from space agencies (ASI, CNES, CSA, DLR, EUMETSAT and ESA) as well as other European EO space data archive holders in order to help define and promote a cooperative and harmonised collective approach among data owners to LTDP. The following eight main themes were identified:

1. Preserved data set content definition and appraisal
2. Archive operation and organization
3. Archive security
4. Data ingestion
5. Archive maintenance
6. Data access and interoperability
7. Data exploitation and re-processing
8. Data purge prevention

For each of these themes, guiding principles and key guidelines were produced, which are detailed in the European LTDP Common Guidelines (RD-01). The LTDP guidelines provide an important reference for the long term preservation of EO space data. European EO space data archive holders are strongly encouraged to adopt these LTDP guidelines in order to preserve their EO space datasets and to help create a European LTDP Framework.

There are three levels of adherence (A, B and C) amongst the key LTDP guidelines, with A being the basic level and C the most advanced. Each key guideline is associated with a particular level of adherence. European EO space data archive holders should aim to increase compliance with the various LTDP guidelines, moving from the adoption of the basic Level A guidelines towards the adoption of the more stringent Level C guidelines where possible:

Table 1 Levels of adherence to the LTDP Common Guidelines

Adherence Level	Description	Condition for adherence
Level A	Basic data security, integrity and access.	Implementation of all high priority key guidelines.

Adherence Level	Description	Condition for adherence
Level B	Medium data security, integrity, access and interoperability.	Implementation of all high and medium priority key guidelines.
Level C	High data security, integrity, access and interoperability.	Implementation of all key guidelines.

The ESA LTDP strategy and the European LTDP Common Guidelines were further developed through two concurrent projects in 2010:

- FIRST (deFInition of LTDP user Requirements and preservation data SeT composition) provided a detailed analysis of Earth Science user requirements relating to the long-term preservation of Earth Science data, products and related information;
- LAST (Long Term Data Archive Study on New Technologies) focussed on defining the technologies and strategies required for the long-term preservation of Earth Science data.

More recently, ESA has commissioned an LTDP-Data Quality (DQ) study, entitled "Evaluation of requirements on data quality information in relation to the Long Term Data Preservation (LTDP) guidelines", which is being performed by a team comprising CGI, NPL and GMV. The objective of the Quality Assurance for Earth Observation (QA4EO) framework developed by the CEOS Working Group on Calibration and Validation is to provide an approach for assessing the quality of EO data to enable its full exploitation.

The LTDP-DQ study is undertaking a review of both frameworks in order to create concrete recommendations for the secondary information to be saved and the procedures to be used for ensuring they are saved. The strong focus is on ensuring that all required quality information is saved as part of the secondary information of the LTDP. There is a strong coupling between the LTDP and QA4EO frameworks, since it is necessary to save sufficient secondary (quality) information to allow later assessment of data quality and uncertainties. With regard to the procedures, the study is looking at an implementation strategy, focussing on knowledge management rather than on a purely technical solution. That is, it defines strategies for ensuring that the correct knowledge, primarily calibration, validation and quality information, is identified, stored and made traceable.

3.3 Linking CHARMe to LTDP

The EO LTDP Framework is intended to promote the preservation of all European EO datasets. However, in this context, long-term data preservation does not mean solely preserving the data as bytes on tape or disk, using the latest technologies. It is also fundamentally concerned with preserving the knowledge about this data. This means preserving the relevant metadata, documentation about the data format(s) and the software required to reprocess this data, for example, to recreate the Level-1 and Level-2 data from the Level-0 data.

C-metadata clearly complements the aims of preserving knowledge about data and, by providing context on its 'fitness-for-purpose', encouraging appropriate exploitation. There is therefore scope for integrating C-metadata within the European LTDP Common Guidelines and the LTDP Framework. This needs to be discussed with ESA and the LTDP Working Group. The CHARMe project and its relevance to LTDP was introduced at the PV2013 meeting, "Ensuring the Long-Term preservation and value Adding to

Scientific and Technical Data", in ESRIN in November 2013. Further discussions are scheduled for early December concerning LTDP collaboration. The LTDP programme is now focussing less on technologies and more on the harmonisation of archival processes. In this context, CHARMe could be the mechanism for adding commentary metadata to LTDP archives.

In terms of the existing European LTDP Common Guidelines, all that refer to the preservation of metadata are high-priority, Level A guidelines. Firstly, the preserved data set content must include not only data but also metadata (Theme 1):

GUIDELINE 1.1 – Preserved Data Set Content - (Level A)

Identify and archive the following set of data content for each Earth Observation space mission or instrument according to the "Earth Observation Preserved Data Set Content" document [R. 6]:

- Data Records: these include Raw data, Level 0 data and higher-level products, browses, auxiliary and ancillary data, calibration and validation data sets, and **metadata**;
- Processing Software: this includes all the software used in the product generation, quality control, and the product visualization and value adding tools;
- Mission Documentation: this includes among others mission architecture, products specifications, instruments characteristics, algorithms description, Cal/Val procedures, mission/instruments performances reports, quality related information [R. 12], etc.

The C-metadata from CHARMe logically fits within the "mission documentation" metadata to be preserved for the long-term in the context of the preserved data set content (Guideline 1.1).

The European LTDP Common Guidelines also include metadata generation as part of data ingestion (Theme 4):

GUIDELINE 4.2 – Metadata generation - (Level A)

Define a descriptive set of metadata for archived data and generate them as part of the ingest process or when archived data content is updated. The resulting metadata should be formatted according to relevant standards.

List of recommended standards:

- ✓ EO collection metadata: ISO 19115 Geographic Information - Metadata [R. 13]
- ✓ EO product metadata: OGC's GML 3.2.1 Application Schema for EO Products (OGC-07-036) [R. 13]
- ✓ ISO 19100 Series
- ✓ OAIS Information Model for Descriptive Information (e.g. Quality Information) [R. 7]
- ✓ Others TBD

C-metadata could be included with the data to be ingested into an archive or the existing archived data content could be updated with the C-metadata. Indeed, since the C-metadata is typically created post-fact, it is likely that the C-metadata would supplement existing metadata created at the same time as the associated dataset. There are two possible approaches for adding the C-metadata to the existing metadata to be preserved for the long-term:

1. A copy of the C-metadata relevant to the datasets in the long-term archive could be harvested and stored within the long-term archive in addition to the existing metadata.
2. The long-term archive could maintain a link to a CHARMe node storing the relevant C-metadata.

The first approach has the advantage of allowing a copy of the C-metadata to be made and stored in a separate location. This conforms to LTDP Guideline 5.4 that recommends maintaining identical copies of all archived data (Level A), preferably at different locations (Level B) and using alternative technologies (Level C). However, this approach might be technically demanding and expensive to implement operationally. Furthermore, the C-metadata harvesting would need to be repeated at regular intervals in order to maintain an identical copy of the latest C-metadata.

The second approach is preferred in that all that is required is to maintain the linkage from the LTDP archive to the CHARMe node (C-node). However, this also implies a need to maintain the availability of the C-node and ideally duplicated C-nodes to address the LTDP Guideline 5.4 recommending identical copies of all archived data.

The final LTDP guidline relevant to CHARMe covers the provision of metadata search capabilities to allow the discovery of archived data records as part of the data access and interoperability requirement (Theme 6):

GUIDELINE 6.1 – Preserved data set content discovery - (Level A)

Ensure continuous EO missions' preserved data set content discoverability through the following activities:

a) Provide and maintain mechanisms to search and discover archived data records.

b) Provide and maintain a searchable metadata and browse image catalogue of archived data records.

c) Provide and maintain mechanisms to search and discover Mission Documentation and value adding/visualization tools relevant for the designated user community.

The CHARMe node would constitute the metadata catalogue for the C-metadata and the CHARME plugin would provide the required data access search capabilities. Thus the CHARMe implementation is already consistent with the LTDP Guideline 6.1 so long as the CHARMe plugin can be integrated with LTDP archive discovery services.

As a final point, CHARMe could potentially solve some of the issues identified during the ESA LTDP-DQ study, such as the need to identify which secondary data is applicable to a given primary data set and to understand the linkages between secondary data, for example, tracing the flow of information to an end-to-end budget. Many of the other issues identified during the study have also been discussed during CHARMe development, such as the need to cope with different versions of documents and data sets.

4. Summary and future direction

While C-metadata is very important to give context to data in Long Term Data Preservation, connecting it directly to individual data products via SAFE is not considered an appropriate mechanism.

The wider implications of the CHARMe project for the ESA LTDP programme have been assessed and it is clear that a C-metadata service would be highly complementary, especially in terms of the programme aim to ensure support the access to and exploitation of EO datasets. This leads to a recommendation that the metadata detailed in the existing European LTDP Common Guidelines be extended to include the C-metadata from CHARMe. Two possible approaches have been considered for implementing C-metadata in the LTDP context: the favoured one being to maintain links from long-term archives to one or more C-nodes. A further recommendation is to provide duplicated C-nodes to maintain identical copies of the C-metadata.

A final point is that the European LTDP Common Guidelines have been developed by the ESA LTDP Working Group with the intention that all European EO space data archive holders, not only ESA, should endeavour to implement a common strategy with respect to long-term data preservation. Participants in the LTDP programme include STFC (a CHARMe project partner) and the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT). Since the CHARMe project is concerned with the "characterization of metadata to enable high-quality climate applications and services", EUMETSAT is clearly an important stakeholder to support the adoption and integration of CHARMe into the LTDP Framework.

Future direction for 2014

LTPD Working Group

The best forum for discussing the possible inclusion of CHARMe within the LTDP framework is the LTDP Working Group, which currently consists of representatives from the following institutions:

- European Space Agency (ESA)
- Centre National d'études Spatiales (CNES)
- Canadian Space Agency (CSA)
- Canada Centre for Remote Sensing (CCRS)
- The German Aerospace Center (DLR)
- Science and Technology Facilities Council (STFC), representing the UK Space Agency (UKSA)
- EUMETSAT
- Agenzia Spaziale Italiana (ASI)

The remit of the working group extends not only the basic preservation of the European Earth Observation Data Set, but also extends to stewardship access and future exploitation. This European EO Data Set consists of a number of different categories of EO data:

- High and Very High resolution SAR imaging missions/sensors (different radar bands)
- High and Very High resolution multi-spectral imaging missions/sensors
- Medium resolution Land and Ocean-monitoring missions/sensors (e.g. wide-swath ocean colour and surface temperature sensors, altimeter, etc)
- Atmospheric missions/sensors
- Satellite based, other scientific missions/sensors

The association of commentary metadata with EO data sets could be seen as an important aspect of the evolution, enhancing the potential for exploitation. The FIRST and LAST studies and LTDP surveys of standards, procedures and initiatives do not propose any viable alternative mechanism to the CHARMe approach. The LTDP programme is entering a new phase developing archival process models based upon the PAIMAS standard (RD-03) and the PAIS specification (RD-04). There is the potential to extend the EO Data Sets by inclusion of commentary metadata.

The next LTDP Working Group meeting is due to be held at Instituto Nacional de Técnica Aeroespacial, Maspalomas Canary Islands Spain on 20th and 21st of March 2014. There will be representation at this meeting from STFC, which will provide an excellent opportunity to introduce CHARMe for consideration by the working group.

LTDP collaborative data access forum

In addition to involvement with the LTDP working group, STFC is also actively participating in the review of collaborative aspects of the LTDP Architecture project. The current LTDP Architecture project aims to define and consolidate the architecture of European Long-Term Archiving EO systems, especially in the context of the ESA Payload Data Ground Segment (PDGS). To do this, it is foreseen to perform a complete assessment of the impacts in terms of Cost Benefit Analysis of implementing the guidelines from the LTDP initiative, taking into account all the systems involved, their data flows, data sets and their related operational policies for LTDP implementation.

The LTDP guidelines, and the standards adopted as a result, will have various impacts on the legacy and future long term archiving systems, the systems used to access them and on the overall operational concept. This will particularly apply to the ESA PDGS environment. Those impacts will have an influence on the final definition of the architecture(s) for the LTDP implementation and the evolution of individual systems towards LTDP compliance. Since the LTDP guidelines allow different architectural options, the resulting architecture and policies defined in this project must be done in such a way as to minimise the impacts on current systems, allowing them to evolve easily towards the consolidated architecture.

The CHARMe project is again presented with the opportunity to identify the best possible collaborative route for integrating CHARMe with LTDP at the Preliminary Design Review (PDR) review meeting due to be held at ESA/ESRIN on 12th December 2013. This review meeting will cover collaborative aspects and user interaction with the LTDP archive and metadata.

SCIDIP-ES project

A beneficial collaboration could potentially be established between CHARMe and the SCIence Data Infrastructure for Preservation with focus on Earth Science (SCIDIP-ES) project. SCIDIP-ES (http://www.scidip-es.eu) is a European 7th Framework project, which aims to promote state of the art digital data preservation technologies within Earth Science repositories.