IMPROVING THE EXCHANGE OF EO ARCHIVED DATA
THE CEOS PROPOSAL FOR A STANDARD DATA ARCHIVE INTERCHANGE FORMAT

Gian Maria Pinna
European Space Agency - ESRIN
Earth Observation Applications Department
Via G. Galilei, 00044 Frascati (Rome), Italy
Tel. +39 6 94180644, Fax. +39 6 94180632
E-mail GianMaria.Pinna@esa.int

ABSTRACT

The CEOS Baseband Data Archive InterChange Format (ICF) is a new standard developed by CEOS via the Archive Task Team (ATT) of WGISS. ESA/ESRIN is presently chairing the ATT.
The work aims at developing a new standard format for the exchange of data among archiving centers. The lack of such a standard has been in the past one of the main problems in exchanging level 0 data with other agencies in the world.

Most Agencies are currently converting, and almost all acquiring, the data directly into computer compatible format, unfortunately adopting different formats. As a result, each Agency uses its own format for the EO data archived and the exchange of these datasets among cooperating Agencies is difficult. In addition, each Agency uses its own definition for the data archived and some ambiguity exists about the processing level applied to them. It is also clear that there's no driving force in the industry to adopt a universal and common format for the data archiving systems they develop for the data holders. A role for CEOS ATT has been clearly identified to propose and define, in collaboration with industry, an effective way to translate the internal archive format in a new InterChange Format (ICF) to be ideally supported by all systems.

1 INTRODUCTION

The CEOS Baseband Data Archive InterChange Format (ICF) work aims at developing a new standard format for the exchange of data among archiving centers. The lack of such a standard has been in the past one of the main problems in exchanging level 0 data with other agencies in the world. Now most agencies are converting, and almost all acquiring, directly into computer compatible format. Unfortunately almost all companies providing such systems are adopting a different format and the problem of cross-compatibility is appearing again.

As a result, presently each Agency uses its own format for the EO data archived and the exchange of these datasets among cooperating Agencies is difficult. In addition, each Agency uses its own definition for the data archived and some ambiguity exists about the processing level applied to them.
It is also clear that there's no driving force in the industry to adopt a universal and common format for the data archiving systems they develop for the data holders. ATT's proposal has been to establish a common format to be used only for data exchange. Ideally this InterChange Format (ICF) will be adopted by all companies and embedded in their archiving systems as a known and widely accepted ‘export’ format. Agencies would then exchange the archive data by mean of the ICF.

2 CEOS BASEBAND DATA CONCEPT

The Baseband Data concept intends to address what processing should or should not be performed and what data and metadata should be retained in the archives in order to keep the loss and destruction of the original sensor data to a minimum.

Until recently, most high data rate earth observing satellite sensors have kept their raw signal data in its rawest forms on high density data tapes (HDDT). Lately, some ground stations have begun direct data capture to computer files and with it the ability to perform some initial processing.

With the advent of direct capture to computer files options have come to pre-process the data:

- to remove bad data and redundant data
- to remove artifacts of transmission
- to repair for transmission errors
- to separate the multiplexed information
- to add auxiliary information from external sources
- to transform the data into some other format

The reformatting of the data is often motivated for the convenience of the archives, for the convenience of the next level processing, for data distribution, or for a more structured format of the raw data. The motivation to preserve the original sensor information is sometimes lost in the attempt to have one format addressing all data system needs.
This evolution in the sensors signal storage also permits to have the data in a form with a much higher independence from the type of medium used. The so-called Computer Compatible Format (CCF) permits now this degree of freedom and is going to be adopted by the vast majority of data holders for their archives.

Although organizations such as Consultative Committee for Space Data Systems (CCSDS) have moved forward with very clear methods for transmission of data from spacecraft to the ground station, the standards for preserving and exchanging this raw data between data centers are less well formulated. For historical data the methods for exchange are archaic or non-existent or mission-specific. There is a strong need for a standardization of the data that is one step higher with respect to the raw signal archive. With the ability to transport large amount of data inexpensively and quickly and with the ability for the end user to process large quantities of data on fairly inexpensive systems, there is an increasing demand for ground stations to capture raw data and distribute this data to the end user. The exchange of data between archive facilities can also greatly benefit with more reliable media at lower costs as long as the formats have some commonality.

The definition in non ambiguous terms of the “level of processing” required by the EO data for long term archiving has been a critical step for any future job the ATT intended to carry out. To this end, the ATT has published a white paper titled "Baseband Data Concept", which aimed to define this minimum level. The Baseband Data concept has been the starting point for the definition of the ICF. The overall goal of the Baseband Data concept is to ensure cost effective archival without losing or destroying data, and at the same time ensure that the data can be exchanged cost efficiently from the various Baseband Data archives. This can be accomplished by assuring that all pre-processing that requires specific equipment and technology is performed before the data enters the Baseband Data archives. Any further processing and use of the data among the earth observing industry and users should be processed with commonly available technology.

3 ARCHIVE FORMATS COMPARISON

The aim of the comparison of some of the most popular archiving formats was to check the CEOS ICF feasibility. The basic requisite to be confirmed was that no significant information losses would occur when a dataset is translated from a source format into a target format. A basic way to achieve this objective is to assess exchange losses in concrete cases through a set of Existing Archive Formats (EAFs). While the limited analysis could not obviously ensure that any format can be translated into any other, most of the results can be extrapolated to almost all archive formats.

The analysis was performed by considering for what concerns the metadata structure that an EAF may be schematically thought as a table that contains cells that are filled with a data entity. The term data entity here is understood as the information that the EAF is meant to preserve. It can consist of the content of a single data field or the combination of several data fields.
A source EAF is a table that is completely filled and target EAF is another table, which is empty. Therefore, exchange shall consist in filling up the target EAF with information contained in the source EAF. In other words, the idea of the analysis was to check that all information required in target format is available in source format.

Information required by the target EAF might not be directly available. Several operations may be necessary to express it in the right formalism. The EAFs analysis methodology used during the ICF development was to tag each information as a function of the difficulty that the exchange of such information could present. In order to achieve maximum results with a reasonable effort, the analysis was performed by comparing each of the chosen EAFs against a pre-defined data entity dictionary. This data entity dictionary was generated by an in-depth classification of all entities belonging to all EAFs. The classification of data fields contained in each EAF naturally lead to the identification of five main information blocks (Figure 1). Some are divided into smaller information blocks.

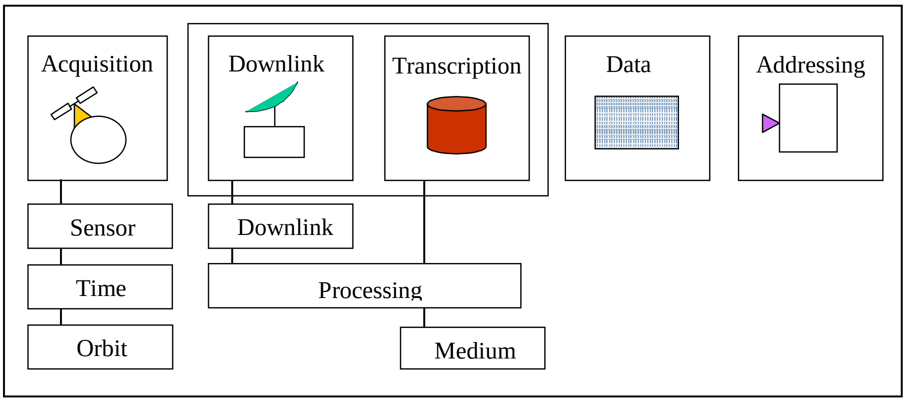

Figure 1

- the ‘Acquisition information block’ contains all information concerning when, where and how the satellite acquired the data set. It contains three areas: sensor, time and orbit.
- The ‘Downlink and Transcription information block’ contains all information concerning when, where and how the dataset was captured by the ground station, processed and recorded onto the current medium. It also contains three areas: downlink, processing and medium.
- The ‘Data information block’, in addition to the image data itself, contains all information about its characteristics and quality information.
- The ‘Addressing information block’ contains all information required to access data.

In reality, data fields are not necessary organized within the format exactly according to these identified blocks. Considerations other than information content force to scatter information at different places. These may be for example transcription convenience, location access ease or security redundancy. They are different for each analyzed EAFs.

The analysis performed on the chosen EAFs, according to the adopted investigation methodology, and the comparison of the missions/instruments analyzed shows that, despite some specific differences, all formats perform basic pre-processing: PN decoding is applied, downlinked channels are interleaved, play-back data are inverted, Reed-Solomon corrections are applied, DPCM is decompressed and data are almost always unpacked into bytes. Even for the more elaborated level 0 formats, changes are usually limited to a re-arrangement of the transmitted data and are reversible.

As a conclusion the EAFs analysis showed that:

- Information commonalties between existing archive formats are enough to allow a transfer of a consistent amount of data.
- Differences of structure and syntax do not prevent the information exchange. Obviously, some losses are possible but storing extra information in separated structures should overcome most of them.

Moreover, this analysis highlighted the fact that it is possible to complete efficiently a target format as long as not only source format is employed but also all necessary external sources of information are used. Indeed, the exchange does not concern solely archive formats but it involves the source and target whole information systems.

4 ICF MAIN CONCEPTS

The figure 2 summarizes the process of exchanging EO data between two co-operating agencies, using the CEOS ICF.

In brief, the figure represents the responsibilities of various actors in the process of acquiring, archiving and distributing the data. Each of these actors (which role may obviously overlap), have specific responsibility in defining the format in which the data are archived or transferred.

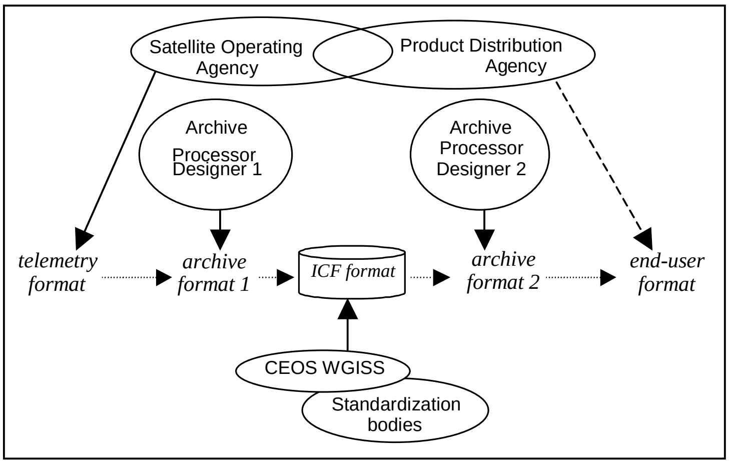
Satellite Operating Agency
Product Distribution Agency
Archive Processor Designer 1
Archive Processor Designer 2
telemetry format
archive format 1
ICF format
archive format 2
end-user format
CEOS WGISS
Standardization bodies

Figure 2

Telemetry format This format has been issued by the satellite operating agency (e.g. ESA for ERS, NASA for Landsat, CNES for SPOT) and distributed to the ground stations within a document often called “Satellite to Ground Station Interface Specifications”. This reference document defines the way data are transmitted from satellite and possibly the way auxiliary data useful for further processing are transmitted from Mission Operation Center. This document is unique and contains the exhaustive list of data entities that are handled by the mission.

Archive formats These formats are normally designed by the companies, sometimes in conjunction with the operating agency, that have developed complete systems able to receive, archive and produce end-user data for one or more missions, or by the operating agency.

End-user formats In the majority of the cases, end-user formats are defined by the satellite Operating Agency it-self. But, in the same way distribution is sub-contracted to subsidiaries or other organizations, end-user formats are sometimes defined by these third-part organizations.

The ICF is proposed as a simple way to transfer data between two data holders without the need to support the rapidly growing number of archive formats today present in the EO world. The ICF is controlled by CEOS WGISS as a third party agent, in collaboration with mission, archive formats and standards experts.

In order to achieve the proposed result the ICF should conform to the following main requirements:

1. **Portability** – The format should be suitable for a wide variety of storage and exchange technologies. Format should be simple and managed by standard software.
2. **Exhaustiveness** – All the elements should be kept that will allow post-processing until a reliable end-user product.
3. **Openness** – The format should be open for further introduction of new mission/instrument and new acquisition technologies.

The three general requirements seen before translate into a set of detailed requirements and norms for the CEOS ICF. The knowledge of these requirements is essential for the appropriate use of the CEOS ICF.

The CEOS ICF has the following main characteristics:

Interchange
The CEOS ICF allows interchanging archived data among several facilities. It is thus not only able to represent sensor measurements and auxiliary data but also information necessary to their exploitation (origin, format, bounds, etc.) This does not imply that all the information related to measurements can be stored but at least those essentials to their identification and processing.

Baseband Data Concept
The CEOS ICF conforms to the Baseband Data Concept as defined by the CEOS Archive Task Team. The main purpose of this concept is to keep the loss and destruction of original data to a minimum in order to preserve Earth Observation datasets for an indefinite period (long-term archive).

Baseband Interchange Unit
The "Baseband Interchange Unit" concept has been created for the CEOS ICF needs. A CEOS ICF Archive dataset contains a single "Baseband Interchange Unit". The BIU is an important concept in the CEOS ICF and its normative definition is the following:

> Definition: Baseband Interchange Unit (normative)
>
> A "Baseband Interchange Unit" is a contiguous sequence of measurements and related auxiliary data acquired during a single satellite pass and over a continuous period of time.

This concept greatly decreases the effort required for the development of transcription tools and increases their robustness. There is therefore no multi-volume representation and thus no ambiguity concerning the number of Metadata Files or Signal Data Files nor any redundancy of information. The CEOS ICF is always a coherent subset of measurements acquired in a continuous period of time.

Multi-missions and multi-sensors
The CEOS ICF does not assume any mission, sensor or specific parameter. Therefore the format specification does not define attributes or entities like for example specific low/high gain mode status. However, a maximum number of "generic" entities (even complex ones) have been defined and thus can be declared easily. For instance, sets of "ephemeris" vectors can be stored in common geodetic and time systems. Specific attributes/entities storage problems are solved by the requirement "Extendibility" defined here below.

Extendibility
The CEOS ICF enables the definition of new entities and the extension of predefined entity definitions. This does not mean that the CEOS ICF definition can be modified but only that it is possible to extend it in order to convey very specific values. This capability guarantees that CEOS ICF can remain an interchange format for a wide variety of missions and in particular for coming ones whose specific components are not already known. The extension capability is based on a typical class inheritance mechanism well known in Object Oriented Modeling. Extension definitions may be ensured by the organization responsible for the specific component. For example, the European Space Agency may be responsible for the ENVISAT specific attribute definition for CEOS ICF extension;

CNES may be responsible for those concerning SPOT; CCSDS may be responsible of the satellite identification number, etc.

Hardware/Software and Operating System independence

The CEOS ICF has been designed to be as independent as possible from hardware, software and operating systems.

In order to satisfy all these requirements the CEOS ICF has been primarily designed to be stored on hard disk or any other "file system" devices (e.g. CD-ROM). In actual fact, the CEOS ICF can be transferred on-line to the receiving facility, or stored on magnetic tape medium using tape archiving tools. Facilities participants in each interchange should agree on a common tool. This freedom is permitted in order to allow the CEOS ICF to be stored on the maximum number of media possible and accessed by most operating systems.

As shown in Figure 3, one CEOS ICF dataset is nominally composed of two parts: one "Metadata File" and one or more "Signal Data Files".

The CEOS ICF Metadata File is a **standard XML document**. The filename obeys to CEOS ICF defined rules and gets the standard extension "xml" or "XML". Instead, no specific rule is required for Signal Data Filename syntax except that it shall be identifiable by an URI reference in the Metadata File.

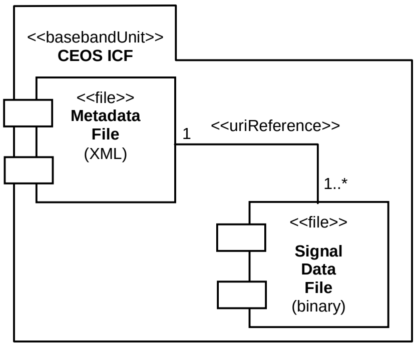

Figure 3

Other optional files can accompany the CEOS ICF archive. Such files may be for example the metadata validation scheme (XML Schema), some style-sheets for metadata display (XSLT) or auxiliary files defined in any CEOS ICF extension. The CEOS ICF recommends placing them in the same directory with the same base name. It is important to recall that these files are not part of the standard.

An important feature of the CEOS ICF Baseband Data Unit is that, in order to cope with possible file system size limitations, the CEOS ICF allows the partitioning of the Signal Data File into several subsets, each of them identified by a separate URI in the Metadata File.

The Metadata File of a CEOS ICF archive contains information related to the Baseband Interchange Unit. It describes the linked Signal Data File, identifies the physical components involved such as the sensor or platform and provides administrative information regarding the facilities involved.

XML has been chosen for the CEOS ICF metadata file for several reasons but in particular:

- XML is supported by a wide variety of applications
- It is easy to write programs which process XML documents
- XML documents are human-legible and reasonably clear
- XML documents terseness and structure can be described and validated.

All CEOS ICF markups are defined in a specific namespace. The goal of this namespace is to separate CEOS ICF “markup vocabulary” from any other defined elsewhere. The CEOS ICF namespace is identified by:

http://wgiss.ceos.org/ceos-icf/ceos-icf-v1.0-23042002

As for almost all others, the CEOS ICF namespace has a URI syntax, but does not assume the correspondence to any existing file or data source. Interest is only focused on uniqueness of this namespace identifier.

A CEOS ICF Metadata File can be validated using the ICF XML Schema. The availability of a CEOS ICF validator is one of the most important characteristics of the format. The CEOS ICF XML Schema is identified by the following URI:

http://wgiss.ceos.org/ceos-icf/ceos-icf-v1.0-23042002.xsd (TBC)

The targeted resource can be downloaded for validation of an occurrence of CEOS ICF Metadata Files. According to the XML Schema specifications, the xsi:schemaLocation attribute can be used in the Metadata File to permit validation tools or XML processors to automatically retrieve the schema. The example below illustrates the use of this attribute within the CEOS ICF Metadata File:

☞ Sample of use of namespace and schema location within CEOS ICF (informative)

```
<ceos_icf
    xmlns=”http://wgiss.ceos.org/ceos-icf/ceos-icf-v1.0-23042002“
    xmlns:xsi=”http://www.w3.org/2001/XMLSchema-instance”>
    xsi:schemaLocation=”http://wgiss.ceos.org/ceos-icf/ceos-icf-v1.0-23042002
            http://wgiss.ceos.org/ceos-icf/ceos-icf-v1.0-23042002.xsd”>
            …
</ceos_icf>
```

The XML capability to define additional namespaces in XML is used in the CEOS ICF to satisfy the requirements of extendibility, as defined previously in this paper.

As an example, consider the following **`sensor`** element, where an exporting agency has included two additional entities to the standard definition:

☞ Sample of correct sensor instance (informative)

```
<sensor>
    <family_name>HRV</family_name>
    <number>2</number>
    <spot:mirror_step value=”23”/>
    <spot:sensor_mode>XI</spot:sensor_mode>
</sensor>
```

The two additional entities belong to a separate namespace, in the example named **spot**. The standard CEOS ICF XML-Schema will still validate this modified `sensor` element and ignore the additional entities, but an importing agency may decide to use this additional information by using this separate namespace.

The CEOS ICF defines the units and reference systems in which the data entities are expressed in a specific instance of the Metadata file. The BIPM (Bureau International des Poids et Mesures) units system is assumed by the CEOS ICF standard definition and, unless specified differently, the World Geodetic System WGS 84 and the UTC are used for the coordinate systems and the times respectively.

Similarly, the following standard global data types are defined in the CEOS ICF: integer, positiveInteger, dateTime, string, double and uriReference, all as defined by W3C for the XML-Schema. In addition, the CEOS ICF recommends the following representation for the dateTime data type: CCYY-MM-DDThh:mm:ss.sss, fully compatible with the ISO-8601 standard from which the W3C definition of dateTime is taken.

The Signal Data File in a CEOS Baseband Data ICF is a binary file containing the telemetry data extracted from the input archive. The signal data file is broken down in signal data records of fixed length. Because CEOS ICF is designed to accommodate telemetry data acquired from different sensors and by different platforms, the signal data records have a dynamic layout. The precise content of the signal data records is described in the Metadata File through the `signal_data` element and in particular with the `block` element, by means of a precise mapping between the native telemetry and the signal data record. This mapping mechanism will be described in the next chapter.

5 ICF METADATA FILE ELEMENTS

As stated previously, the Metadata File of a standard CEOS ICF data representation is a standard XML document. The UML diagram in figure 4 shows the first level decomposition of the CEOS ICF Metadata File XML root element, together with its validating XML-Schema.

🕮 Fraction of schema for CEOS ICF root element (normative)

```
<xsd:element name=”ceos_icf” minOccurs=”1” maxOccurs=”1”>

   <xsd:complexType>
        <xsd:sequence>
           <xsd:element name=”platform”
                        type=”platform”
                        minOccurs=”1”
                        maxOccurs=”1”/>

           <xsd:element name=”processing_log”
                        type=”processing_log”
                        minOccurs=”0”
                        maxOccurs=”1”/>

           <xsd:element name=”signal_data”
                        type=”signal_data”
                        minOccurs=”1”
                        maxOccurs=”1”/>
        </xsd:sequence>
   </xsd:complexType>

</xsd:element>
```

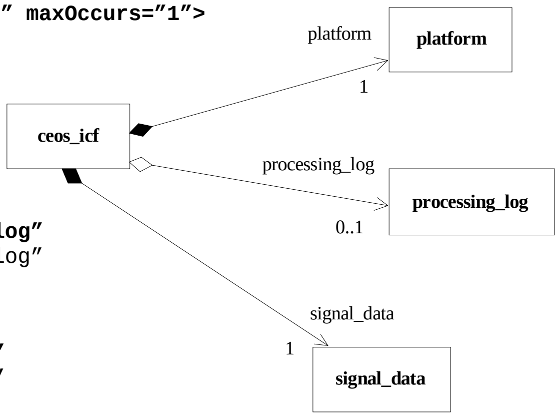

Figure 4

The CEOS ICF Metadata file must contain one and only one `ceos_icf` element. This element is the root node where all CEOS ICF metadata are stored. A `ceos_icf` element contains metadata relative to a single Baseband Interchange Unit.

The mandatory `platform` element identifies the system (satellite/aircraft) that acquired the data present in the Signal Data File. This element contains sub-elements that unequivocally identify the platform as well as those that identify the specific sensor that acquired the data and the precise position in space. In particular the `orbit` element describes the trajectory of the platform. Usually the described orbit matches exactly the acquisition period of the data present in the Signal Data File. However it can also describe a larger period but in any case it shall contain the acquisition period of the data present in the Signal Data File. The `orbit` element is defined using ephemeris element set and attitudes around the center of gravity of the platform.

The optional `processing_log` element (for the corresponding UML diagram refer to the ICF document) gathers history information enabling the maintenance of the traceability of the described archive data through its main activities: receiving, pre-processing, archiving and exporting phases. Even if logged processing is dated, no assumption is made about the chronological order of the processing list.

The main element of the CEOS ICF Metadata File is the `signal_data` element (UML diagram in figure 5). This element provides start time and end time of the acquired data embedded in the Signal Data File, in addition to its physical location through its `subset` element that contains the URI

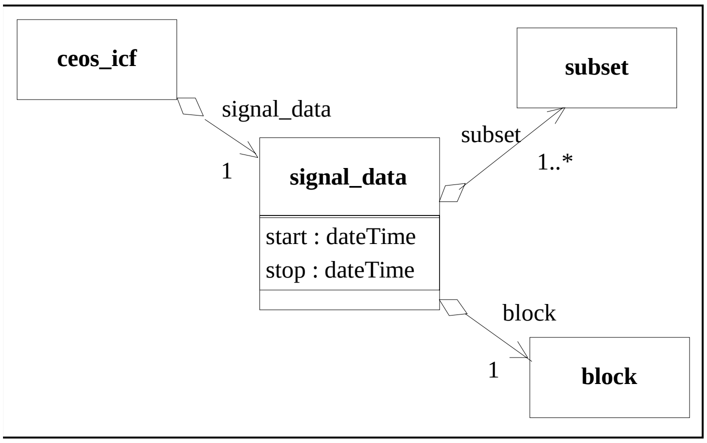

Figure 5

of the various subsets that compose the Signal Data File. In addition it provides the internal structure of the Signal Data Files, composed of Signal Data Records (SDRs) through its `block` element.

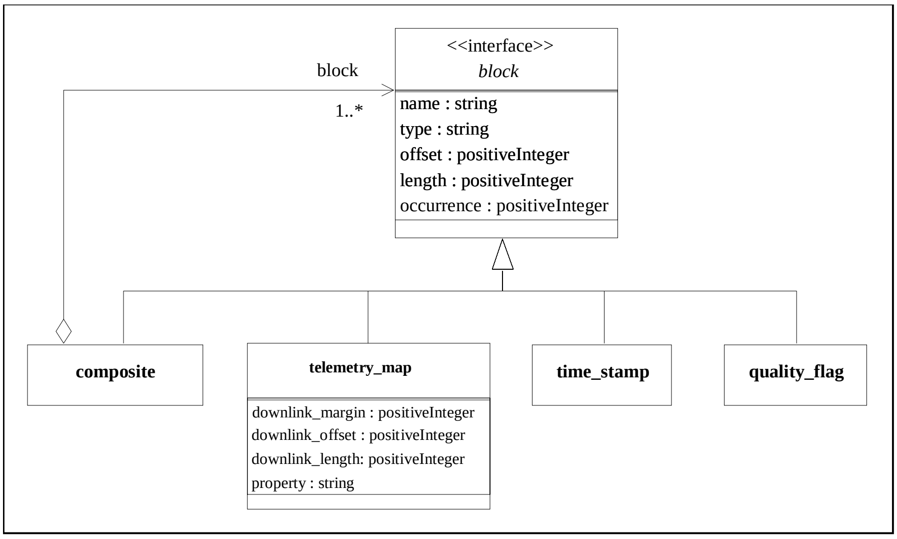

Figure 6

This element (see the UML diagram in figure 6) is the most important part of the CEOS ICF Metadata File and it completely defines the Signal Data Records and their contents. The SDRs are considered as an archive block that may be divided in sub-blocks. The `block` element may specialize in four different implementations:

- **composite block**: breaks down all or part of the Signal Data Record into one or more sub-blocks
- **telemetry map block**: provides a mapping from the telemetry format to the Signal Data Records of the present archive
- **time stamp block**: provides time information regarding a particular facility or system
- **quality flag block**: indicates if the Signal Data Record can be used or contains an error.

The reader should refer to the ICF document for a better understanding of all four elements in details. A description is provided here only for the `telemetry_map` block. A `telemetry_map` block element describes the mapping between a part of the Signal Data Record format and the transmission format structure. As already mentioned, this format is known exactly for each platform and stands as the unique reference to describe the contents of the Signal Data Record. This very versatile mechanism enables the description of many archive data regardless of the acquisition platform.

The Signal Data Record is a subset of the smallest repeated pattern within the overall transmitted telemetry. Generally it corresponds to a frame or a major frame. It must be noticed that the smallest repeated pattern is not the smallest pattern that can be retrieved in the telemetry, which may be the byte, the format fields or the minor frame. The smallest repeated pattern is a data subset that has no higher structure definition in the telemetry. For example, the smallest repeated pattern for the ERS SAR instrument includes the frame 0 and the frames 1 to 28. For the Landsat TM or ETM+ sensors it is the major frame. The fact that the Signal Data Record is a subset of this smallest repeated pattern reflects the possibility that some parts of the original telemetry may have been removed according to the general concepts of the CEOS ICF format (e.g. the synch code); other parts extracted from the input archive may have been slightly modified with respect to the Baseband concept that restricts the CEOS ICF format (e.g. decompression may have been applied, CRC code removed, byte aligned, etc.)

Each block is defined by an optional name and comment description, its offset and length inside the SDR, the number of consecutive occurrences of this block inside the SDR and by its offset and length inside the transmission format structure (parameters `downlink_offset` and `downlink_length`). An additional parameter `downlink_margin` describes if a portion of the telemetry format has been discarded at the beginning of the block (type: header) or at the end (type: footer). In addition one or more optional `property` elements can indicate if a particular processing has been applied to the block when extracted from the original telemetry format structure. With this method of representation, very easy to implement in a computer program although complicated to explain, each block of the SDR is uniquely mapped to the corresponding block in the original telemetry format structure. The various parameters entering into play in the SDR description are represented in figure 7.

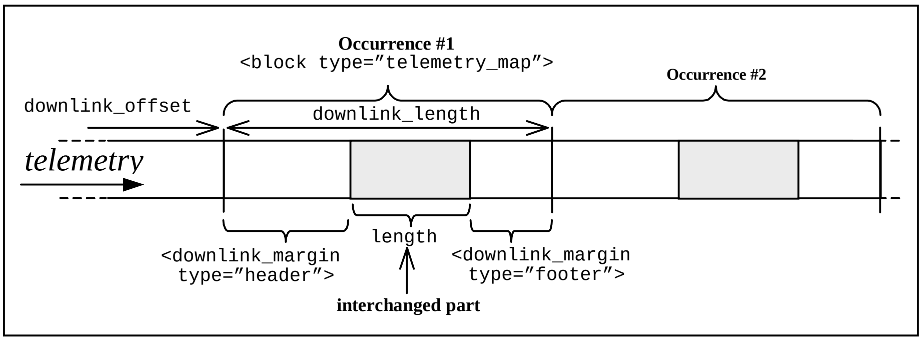
Occurrence #1
<block type="telemetry_map">
Occurrence #2
downlink_offset
downlink_length
telemetry
length
<downlink_margin type="header">
<downlink_margin type="footer">
interchanged part

Figure 7

6 ICF FUTURE USAGE

The development of the CEOS ICF has just been completed at the time of writing (July 2002). The standard will be proposed to the CEOS member agencies for endorsement and voluntary application. A great interest has been shown by both data holders ands industry in the standard, and the hope is that it will be implemented in the major archiving systems as import/export format to be used whenever the need arises.

A pilot project between ESA and USGS to test its operational applicability with some sample data (most probably Landsat 7) has been proposed at the last ATT meeting. The pilot project is open to other agencies that will want to participate with the same or different datasets.

In ESA the CEOS ICF is considered for adoption for the Cryosat Ground Segment, presently in its development phase, for the transfer of the data between the mission archive at the acquisition stations and the long-term final archive, which will be located at one of ESA's partners agencies premises. The adoption of the CEOS ICF by the Cryosat project opens good perspectives for the usage of the ICF in the future by other ESA Earth Observation missions.

7 CONCLUSION

The CEOS Baseband Data Archive InterChange Format represents the first attempt to standardize the process of exchange of Baseband archived data among co-operating data holders. Although the standard has not yet been published (at the time of writing), it has attracted much interest both by EO agencies and industry. Its success will be determined by the continuous and high-level co-operation achieved so far among CEOS members and industry during its development.