CAA Metadata Dictionary

	Name	Date	Signature
Prepared by:	Harri Laakso	1/4/2014	
Reviewed by:	Chris Perry		
Approved by:	Harri Laakso		

Table of Contents

List of Tables

List of Figures

Document Status Sheet

Document Title			Cluster Metadata Dictionary
Document Reference Number			CAA-CDPP-TN-0002
Issue	Revision	Date	Reason for Change
1	0	2004 Sep 21	First version released
	1	2004 Dec 9	
	3	2005 Feb 4	
	4	2005 Mar 7	
	5	2005 Mar 9	
2	0	2005 Mar 17	New Issue
	1	2005 May 18	This revision had limited circulation, and was never formally released. Changes are listed with those of Revision 2 below.
	2	2006 May 4	• Table 1, two instruments for each of CIS and STAFF • New Sect. 6 which describes time stamps • Sect. 7, page 10, addition of definition of the allowed character set for free text strings • Sect. 7.10, page 24, use of "DATA" explained • Sect. 7.10, page 25, numbers now associated with processing levels • New Sects. 8.5 and 8.6 concerning documentation and software products : both need to be completed • Page 42, note added to definition of CATDESC • Former Appendix D "Measurement Type Definitions" extended to a more complete Appendix B "Glossary of all CAA Keywords" which, however, is still incomplete • Appendix D, "Index", is more comprehensive • Addition of the enumerated keywords : EXPERIMENT = "WBD" INSTRUMENT NAME = "WBDn", "CIS-CODIFn" "CIS-HIAn", "STAFF-SCn, STAFF-SAn" for $1 \leq n \leq 4$ ENTITY = "Oxygen+" PROPERTY = "Component" PROPERTY = "Raw_Particle_Counts" PROPERTY = "Status" FLUCTUATIONS = "Mean_Square_Level" INSTRUMENT TYPE = "Particle_Correlator" INSTRUMENT TYPE = "Data_Processing_Unit" INSTRUMENT TYPE = "Auxiliary" MEASUREMENT TYPE = "Particle_Correlator" MEASUREMENT TYPE = "Status" instead of "Spacecraft_Status" MEASUREMENT TYPE = "Emitted_Current"

2	3	2008 Mar 5	• OJN-673787: More support for multi-parameter products, as described in the new Sect. 7.4.1, page 12. New enumerated keyword "Multiple" on pages 36, 37 and 38 • VUQ-772393: Two new types of coordinate system introduced in Sect. 7.7, page 18 • Definition of a new type of data entity “Coordinate rotation” in Sect. 7.8, page 20 • VDJ-971869: Clarification on when to use TIME_RESOLUTION, page 27 • LOG-701234: Warning concerning interpretation of minimum and maximum time resolution, page 27 • VHH-616637 Entity “Ions” replaced by “Ion” on page 43 • Difference between “Fluctuation_Level” and “Mean_Square_Level” clarified on page 58 of Appendix B
3	0	2013 Nov 1	• Renaming of the Cluster metadata dictionary • Several updates in enumerated values for the following keywords: o MISSION, MISSION_AGENCY, EXPERIMENT, OBSERVATORY, INSTRUMENT_NAME for the Double Star mission; the values are taken from appendix A (v1.1, dated on 16 Nov 2009) o MEASUREMENT_TYPE, PROPERTY, TENSOR_FRAME for all datasets; the values are taken from appendix B (v1.0, dated on 1 May 2010) o QUALITY: Modified definition o New coordinate systems: FAC, MFA o New instrument type: Solid_State_Detector o New PROPERTY: Photon_Flux • new keywords o TARGET_SYSTEM • Changes in appendices o Appendix D deleted o Enumerated lists are now given in Appendix B

Executive Summary

The Cluster mission constitutes, together with SOHO, the Solar Terrestrial Science Programme (STSP), the first cornerstone of the ESA's Horizon 2000 Programme. The four Cluster spacecraft were launched in pairs on two Soyuz rockets in July and August 2000. Originally planned for a two-year mission, after one year of successful operations, the mission was extended for an additional 35 months, up to December 2005. Cluster is fulfilling its promise as a revolutionary magnetospheric space mission so well that, in February 2005, the mission was extended for a further four years, until December 2009. The four-point measurements, made with identical instruments on closely spaced satellites have already yielded unparalleled views of space plasma processes in key regions of the magnetosphere and in the near-Earth, upstream solar wind. The data acquired over the past four years validate the Cluster mission concept and its scientific objectives. Magnetospheric phenomena are clearly seen to have a richness and complexity that could only be suspected. The Cluster observations have shown that, on the scales so far explored, 3D plasma dynamics clearly plays an essential role in shaping the larger scale structures of our space environment. The mission is successfully delivering summary and prime parameter data through the Cluster Science Data System, and raw data to the Principal Investigator teams.

The Executive is now seeking to augment the mission by implementing a Cluster Active Archive (CAA) that will contain processed and validated high-resolution scientific data, as well as raw data, processing software, calibration data, documentation and other value added products from all the Cluster instruments. Cluster Active Archive is expected to function until December 2010, by which time the data should have been migrated to a permanent archive. The scientific rationale underpinning this proposal is as follows :

- Maximise the scientific return from the mission by making all Cluster data available to the worldwide scientific community
- Ensure that the unique data collection returned by the Cluster mission is preserved in a stable, long-term archive for scientific analysis beyond the end of the mission
- Provide this archive as a major contribution by ESA and the Cluster science community to the International Living With a Star programme

Such an archive must contain a description of the archived data adequate for users :

- to use generic search tools to locate the required data (no search tool can produce results more exact than description of the product for which they are searching)
- to see a short description of the scientific parameters thus found
- to obtain references to a more complete description of the instrument and the derivation of the parameters
- to read the data files and extract the numerical values
- to understand what those values represent
- to obtain adequate information to implement applications programming interface(s) (API) for generic software
- to have this information available, world-wide
- All this must also be understandable by future generations of scientists, and, moreover
- the above requirements imply that the description be comprehensible by both machine and the human eye

For all these reasons a metadata dictionary is required.

In February 2004 the CAA Project established a Working Group with the mandate to produce a CAA Metadata Dictionary to satisfy the above requirements. The resulting metadata dictionary is a compromise between

- existing data formats, especially the Cluster Exchange Format developed by the Cluster Archiving Task Group, and already used within the Cluster community,
- the requirements of international collaboration as exemplified, for example, by SPASE (Ref. 4), and
- the objective of having the system rapidly in operation.

The second issue of the CAA Metadata Dictionary was released after detailed discussions with all the Cluster Investigator Teams and hands-on experience from test datasets. It was anticipated that some minor revision would be required. Subsequently,

revision 1 was never officially released,

revision 2 dated May 4 2006 contained minor changes agreed in the light of eight months of experience gained during the preparation for and pursuance of routine data ingestion, and

revision 3 is the present document. Appendix B is not yet complete, not all keyword definitions are yet provided. The document name was changed to CAA-MDD-0001 as the new releases of the MDD are authored by the CAA

1 Introduction

1.1 Background

Metadata is information which describes a dataset. It should be complete, that is, contain all the information required to read and interpret the bits (syntactic description), and to understand what the resulting numerical values (or bit strings) represent (semantic description), including how the data was obtained ; the latter information impacts upon the scientific significance of the data. The purpose of the CAA Metadata Dictionary is to describe fully the required CAA metadata information, and to explain how that information must be formatted so as to be exploitable by the generic software of Cluster Active Archive.

The way in which data is organised within the data centre may be expected to be closely correlated with the way in which it is acquired, and this is addressed in the present document. In view of the quantities of data involved, it is expected that most metadata delivered to CAA (unlike that supplied to the end user) will be detached, to facilitate update and avoid duplication.

The development of generic tools for managing and interpreting the metadata does not fall within the scope of this document.

For reasons of compatibility, many of the keywords used in this document are the same as those used for the CSDS (Cluster Science Data System, ref. 1) implementation of CDF (Common Data Format, see http://cdf.gsfc.nasa.gov/cdf_home.html). For data exchange within the Cluster community the Cluster Exchange Format (CEF) was developed by the Cluster Archiving Working Group : CEF-1 was compatible with CSDS. Nevertheless, for reasons of improved precision, clarity, or both, CAA has considered it advantageous to make a number of modifications to CEF, not all of which are backward compatible. The new version of CEF is designated CEF-2.

The CAA Metadata Dictionary and CEF-2 are intended to be fully compatible. Eventually the Metadata Dictionary will be a totally independent and free-standing document. Nevertheless, during the implementation phase of CAA and in view of the minor modifications which must be anticipated, it is thought better to avoid unnecessary duplication of information. This leads to numerous cross-references between this document and the corresponding CEF-2 document (Ref. 3).

1.2 History

In the early years of the CAA project two key working groups were formed, one for the data format and the other one for the metadata. The members of the metadata working group were C.C. Harvey (chair), A.J. Allen, F. Dériot, C. Huc, M. Nonon-Latapie, C.H. Perry, S.J. Schwartz, T. Eriksson and S. McCaffrey.

The first version of the CAA metadata dictionary was developed in years 2003-5, and the last version (v2.3) was released in 5 March 2008. Due to a variety of reasons, the dictionary was not updated although some annexes were released to keep track of new keywords and enumerated values of some keywords. Some keywords and values were added directly into the software. In order to keep a proper track of all keywords and their enumerated values, it was decided to release an updated version of the metadata dictionary that contains all additions made in years 2008-2013.

Since the maintenance of the metadata dictionary is now taken care of by the CAA project, the document name was changed. However, versioning was adopted from the past document. The first new release is v3.0.

2 The Information Associated with a Dataset

The metadata includes the following categories of information :

1. Essential scientific information to enable the scientist to identify the dataset(s) which interest him, and to understand the data once he has recovered it. This is the semantic description. It may also includes less essential, but exceedingly useful, practical information to help the scientist, or his application programmes, to use the data correctly and usefully: e.g., plot scales, and labels for the axes.

2. Information to enable the data files to be read and the parameters to be interpreted and recovered correctly. This is the syntactic description.

We add a third category,

3 Information to enable the data centre to organise the metadata from multiple datasets from the same instrument or the same mission in such a way as to provide complete metadata for any individual dataset without archiving multiple copies of redundant information. This is curation information.

The semantic description is useful both to search for data and to exploit it scientifically once it has been recovered. The syntactic information is essential to read and interpret the data (the bits) recovered, but is not used for searching. The syntactic description, and much of the semantic description, are specific to each particular dataset. On the other hand, curation information is more related to the centre in which the data is archived. Curation information is also closely related to the structure and/or the organisation of the mission which produce the datasets, and is therefore considered to be part of the Cluster metadata.

Although the metadata includes all the required information concerning the data which is archived, only part of this information is generally required at any one time. For example

- when a user is searching data within the archive he is not particularly concerned by the format in which the data is delivered, nor by the units unless he is using the physical value as a search criterion, but

- when he decides to recover some data to analyse it, this information becomes essential.

The approach adopted by CAA is that all information shall be held in a homogeneous way, and the various data centre applications, such as search, data extract and preparation for delivery, interface to CAA applications, etc., shall select the metadata parameters as required. In response to a request from a CAA scientific user all metadata concerning the requested dataset and its parameters is delivered. Higher level data not required by scientific applications may be requested separately (to be verified).

The CAA search engine, as well as the search criteria ordered to external data centres (for example, via Web services), thus order, in principle, the possibility of searching any or all of the CAA metadata. This does not imply that it is necessary to implement searching of all the information held ; merely that this possibility exists and may be implemented if one day it becomes desirable. Note too that, if the metadata is preserved in a well-ordered way, searching through the totality of the metadata will be relatively easy using generic software.

The metadata identified for Cluster Active Archive are listed in Appendix A. A metadatum consists of a text string which expresses a predicate of the form “CONCEPT=Value”. The left side is a keyword denoting some concept which is characteristic of the dataset, and this concept takes values or attributes which are of one of the following types:

- numerical value,
- logical name,
- free text string,
- a formatted text string, including the following possibilities:
    - a personnel reference (Section 7.10.6),
    - an ISO standard time code (Section 6),

  - an ISO time range (two successive ISO standard time codes separated by the sign "/"),
  - a units conversion factor (Section 7.10.9).
- enumerated text, that is, a text string selected from a pre-defined list.

The tables of Appendix A list for each the CAA concept keywords:

- its cardinality (occurrence of the keyword), which takes one of six possible values

	compulsory	conditional	optional
unique or	1..1	?..1	0..1
multi-valued	1..N	?..N	0..N

- its source, that is, the entity responsible for allocation of a value
- the type of attribute, a succinct description and, for enumerated keywords, the list of enumerated values,
- references to a fuller description.

Enumerated attributes are used for searching data by keyword, both within CAA, and in the context of data centre interoperability (Virtual Observatory).

Appendix D provides an alphabetical list of all the CAA concept keywords, with additional information and cross-references.

Finally, tools are required to exploit the machine readability of the metadata (in particular, for interoperability and for interfacing to application programs), and these tools must have a much wider field of application than Cluster, or even space plasma physics: otherwise they will rapidly become obsolete. To this end, the data description consists of two parts:

1. The data model, which provides a logical framework into which the metadata information is placed. This framework allows the development of generic metadata handling tools, dependently of the precise context ; thus, in principle, they can be applied to other missions and to other disciplines.

2. The data dictionary, which is the ensemble of words used to populate the data model. For another discipline, these words will be different or have a slightly (or even totally) different meanings. But the data model will remain unchanged.

3 The Data Model

To understand the data model, let us consider the requirements. To find an object, or to use it correctly, it is essential to be able to describe it accurately and in as much detail as required. A spoken language uses one type of word, nouns, to identify concepts, and another type, adjectives, to describe their properties. Any everyday material object can be described in terms of concepts such as its mass, composition, colour, length, width, height, age, owner, etc.. Each of these concepts has a value, which may be

- numeric as in the case of “mass”, “length”, “age”,
- a text string which is either
  - precise as for “owner”, or
  - less precise, as for “colour” (red, yellow, ...) or “composition” (metal, plastic, ...).

Less precise values can generally be qualified by other adjectives such as dark, bright, hard, soft, etc.. For the CAA space physics metadata model we follow the same ideas, extending them to non-material physical concepts. It may seem strange to assimilate a concept such as a “magnetic field” to a “material object”, but there is no logical (or ontological) impediment to doing this.

For the CAA data model each concept is a priori independent of the others: the model is flat. In fact, each concept has a numerical value or a (limited) set of adjectives which may be used to describe it, and ipso facto this imposes a degree of hierarchy. The more concepts for which numerical, enumerated or descriptive values are specified, the more precise is the description. But there is a practical limit : the description must be comprehensible, and this limits the number of concepts and their values. This limit is decided by common agreement of the users, and may differ markedly from one community of users to another, as illustrated by another everyday example. The English word “bell” has multiple French translations, each of slightly different meaning : carillon, clarine, cloche, clochette, sonnette, timbre. Physicists are generally more precise, but never in total agreement -especially when the concepts or their understanding is evolving rapidly.

These different concepts are mutually independent in the sense that the concepts, and the conditions which they satisfy (which are described in terms of the appropriate adjectives) may be searched in any order to obtain the same result. Hence, these concepts commute and, by analogy with quantum mechanics, they are termed orthogonal concepts. Orthogonality implies that the concepts are completely independent of each other. The orthogonal keywords take values, which may be either numeric or, more usually, free, formatted or enumerated text strings (see Appendix B). The concepts and their descriptions must be machine understandable : they must be selected from a list of recognised keywords. The lists of orthogonal concept keywords and their possible value keywords constitute the data dictionary. Value keywords may be hierarchical, as will be seen below.

The flat set of orthogonal concepts qualified by values is the data model. The same model may be applied to other discipline areas with different names for the concepts and different values for each concept ; that is, by changing the data dictionary which is used.

The bottom line is that the better the data is described the better the chance of finding exactly what one wants and, once found, the easier it will be to exploit it scientifically. Care should be taken to maximise the number of concepts, both to optimise the search criteria, and to avoid overloading the descriptive text strings. For example, magnetic field and fluxgate magnetometer are two different concepts. The magnetic field is the parameter which is measured : it may be measured by fluxgate magnetometer, vector helium magnetometer, search coil (on a spinning spacecraft), or even by determination of local plasma gyro-resonance frequency.

The Cluster Active Archive will hold in this way all information describing the archived numerical data, graphical data, software, or documentation.

4 The Data Dictionary

As explained above, the data dictionary is the list of the keywords used to populate the data model for CAA. It covers both the concept keywords, their allowed values, and the qualifiers.

As mentioned in Section 2, for reasons of information management it is important to avoid holding large quantities of redundant information. Therefore the different concepts have been organised hierarchically, as shown in Table 1. At any given level the metadata applies to all data at that, and all lower, levels. For example, all information at the dataset level applies to all parameters and all files of that dataset; and all information about the mission applies to all observatories and all experiments of that mission, plus the instruments, the datasets and the files. This categorisation is doubly convenient :

1. When data is recovered by the end user all the metadata pertaining to the data recovered should be delivered with the data ; every dataset must be associated with the totality of the metadata relevant to it. All data relating to an instrument is relevant to all datasets from that instrument, but this information should not be held at multiple locations, one for each dataset. Exactly how the data centre solves this problem is an internal data centre architectural problem, and some redundancy of information will probably be inevitable.

2. All the metadata held in the data centre must be gathered, and it is best to gather it only once.

Table 1: Hierarchical Levels for Concept Keywords

Level	Description
Mission	This level contains information relevant to the whole mission.
Observatory	The Cluster mission consists of 4 observatories : Cluster-1, Cluster-2, Cluster-3, and Cluster-4.
Experiment	The Cluster mission has 11 experiments, each identified by its Principal Investigator, plus the auxiliary data.
Instrument	The Cluster instruments are identified by Observatory and Experiment. Some experiments obtain data (on each spacecraft) from more than one physical instrument, each of which produces its own datasets ; for two of these, CIS and STAFF, it has been found convenient for the purpose of archiving to identify two instruments. Thus CAA contains datasets from 4 (11 + 2) = 52 instruments, as listed in Appendix B.
Dataset	Each instrument produces one or more datasets ; this level of metadata is common to the whole of each dataset. Parameter A dataset contains one or more parameters, each of which has its own metadata.
File	Each dataset is composed of files, the number of which will grow regularly with time during CAA.

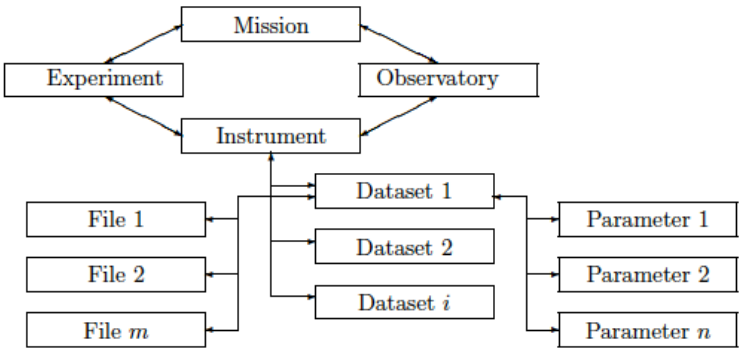

Figure 1: The CAA hierarchy of concepts.

The arrows indicate the ascending and descending links. The n parameters and m data files are indicated only for Dataset 1 ; a similar hierarchy will exist for all i datasets of each instrument.

The hierarchy of data levels is illustrated as in Figure 1; we consider that experiment and observatory are at the same hierarchical level ; the mission metadata block has two descending links, and the instrument metadata blocks have two ascending links. Note that the different parameters and data files are shown only for Dataset 1 ; there will be similar branches for all datasets.

For CAA, there will be:

- one block of metadata at the mission level (for the Cluster mission),
- four blocks at the observatory level (Cluster-1, Cluster-2, Cluster-3, Cluster-4)
- eleven blocks at the experiment level (one for each of the eleven instruments),
- sixty blocks of metadata (listed in Table 6) and the instrument level, plus
- a further six blocks of metadata for the various auxiliary data products.

To recover all the metadata relative to any one dataset it is necessary to know the relation between these blocks of metadata. For example, when looking at the metadata associated with the CIS-1 instrument (CIS instrument on Spacecraft 1) it is necessary to know that this is associated with metadata concerning the Experiment CIS and the Observatory Spacecraft-1, and that these are associated with the Mission Cluster.

Linkage between the different levels (illustrated by the arrows in Figure 1) is provided at each level by concept keywords included specially for this purpose. For example, at the experiment level the link to the mission level is defined by the predicate PARENT_MISSION="Cluster" (see Table 3). The same predicate exists at the observatory level (Table 4). In Table 6 (Instrument Level) two ascending links will be provided by predicates such as PARENT_EXPERIMENT="ASPOC" and PARENT_OBSERVATORY="Cluster-2" for the ASPOC2 instrument. Only ascending links are provided but, since all the links are reversible, all the required linkage information exists; its exploitation it is an implementation issue. There is no redundant information, which greatly reduces the likelihood of internal inconsistency. Furthermore, with only ascending links being documented, it becomes much easier to insert new datasets.

5 Data Type

CAA will archive numerous different types of data product, as listed in Table 2.

Table 2: Cluster Active Archive Data Product Types.
The first two columns show the product ID code and the Short-name as obtained from document Ref. 9. Col. 3 indicates the section of this document describing the product type and associated metadata.

ID	Short-name	Section	Description
CP	CAA_Parameter	7	CAA numerical science data.
PP	CSDS_Prime_Parameter		Spin-period resolution numerical science data, one dataset per experiment from each of the four spacecraft.
SP	CSDS_Summary_Parameter		One minute resolution numerical science data, one dataset per experiment from the reference spacecraft.
CG	CAA_Graphic	8.1	Pre-plotted graphics, with display type and plot scales chosen by the PI.
CT	CAA_Event_Data	8.2	These are files characterised by two times : begin and end. They are used for status data, calibration coefficients, etc., or for classification of scientific events which have identifiable start and stop times.
CQ	CAA_Quality/Caveats	8.3	Special event-type set containing information about the quality of the archived products. These sets will updated more frequently than the datasets, graphics sets, or other product to which they refer.
CC	CAA_Calibration_Data	8.4	Data returned from the instrument internal calibration cycle.
CD	CAA_Document	8.5	Every archived product must be documented.
EP	ESOC_Parameter		
JP	JSOC_Parameter		
CS	CAA_Software_Product	8.6	
CE	CAA_External_Product		Data not obtained from Cluster but whose presence is useful to CAA.
	Other		Data which the user may down load if believes that he can process it by himself

The first three items on the list are scientific data products, and the corresponding metadata is described in Section 7. The other products are described in Section 8 and, to avoid redundant information, whenever possible reference is made to metadata already described in Section 7.

6 Time Tags

There is no a priori reason for data records to be ordered in any particular way. It is quite possible for a data set to contain, for example, the solar wind density as a function of solar wind velocity (ordered or not), with no reference to time. Nevertheless, the parameter which is most natural, and which makes for easiest management of the data, particularly with respect to correlation between parameters from different instruments, is the time of observation (and, or course, the spacecraft and instrument which made the observation). Most, if not all, Cluster data sets are time ordered, and most data extraction, analysis and display tools depend upon time being the key indexing parameter. Time tags thus stand apart from other dataset parameters, whose description is explained in Sect. 7 below.

All times are expressed using ISO Time Code A (Month/Day of Month Calendar Variation), as described in the International Standard ISO 8601:2000 "Data Elements and Interchange Formats – Information Interchange – Representation of Dates and Times", (Ref. 12). The format for ASCII Time Code A is as follows:

YYYY-MM-DDThh:mm:ss.d→dZ

where each character is a one octet ASCII character with the following meaning :

YYYY	= Year in four-character subfield with values 0001 – 9999
MM	= Month in two-character subfield with values 01 – 12
DD	= Day of month in two-character subfield with values 01 – 28, – 29, –30, or –31
"T"	= Calendar-Time separator
hh	= Hour in two-character subfield with values 00 – 23
mm	= Minute in two-character subfield with values 00 – 59
ss	= Second in two-character subfield with values 00 – 59 (– 58 or – 60 during leap seconds)
d→d	= Decimal fraction of second in one-to n-character subfield where each d has values 0 – 9
"Z"	= time code terminator (optional)

Note that the hyphen (–), colon (:), letter "T" and period (.) are used as specific subfield separators, and that all subfields must include leading zeros. The time code may terminate before either of the colons (:), or the period (.) character ; alternatively, as many "d" characters to the right of the period as required may be used to obtain the required precision. An optional terminator consisting of the ASCII character "Z" may be placed at the end of the time code. For Cluster, the telemetered data has a timing precision of the order of 2 ms, so no more than three significant figures after the decimal point are usually required. The possibility to enhance the timing precision of Cluster data is described in Section 6.2.

Some types of data product, such as catalogues, event tables (see Section 8.2) or caveat files, have records which correspond to time intervals of variable duration rather than specific times as in a dataset. For these products the time intervals are expressed as an ISO time range, that is, a pair of ISO times codes separated by a "/" character.

6.1 Empty Data Files

Cluster data is delivered in files containing data from a predetermined time interval, which is often 24 hours. It may happen that no valid data is available for the entire duration of a file. In this case an "empty" file should be delivered. This file must contain the standard header (as when data is present), indicating the correct start and stop times, but need no data records. Its purpose is to help distinguish between intervals for which there is no data, and those for which data is not yet delivered.

6.2 Enhanced Timing Precision

The precision of the Spacecraft Event timestamps of science packets in files delivered to the PI teams by ESOC is 2 ms. This precision is inadequate for inter-spacecraft cross-correlation of electromagnetic field data at frequencies greater than about 5 Hz (1/100th of the inverse of the precision). Fortunately the DWP team has been able to enhance the timing precision of the Cluster RDM ; the correction comes in the form of small offsets to be added to the RDM timestamps, offsets which are to be found in DWP TCOR dataset. Details can be found in the documentation for this dataset.

The TCOR offsets improve the datation of each RDM data packets by increasing its precision to a few microseconds. CEF files containing such enhanced precision data should be written using ISO time codes with six digits after the decimal point.

NOTE THAT files used to produce such high resolution data must be handled very carefully even before the correction is applied: due to the inherent internal precision of software, some format transformations, for example from CEF to CDF then back to CEF, may change the value of the datation by a millisecond or so. This is less than the probable error of the uncorrected data, but orders of magnitude greater than the corrected value obtainable by using TCOR offset.

7 Metadata Keywords for Numerical Data Products

Attached to this document there are appendices containing :

Appendix A All concept keywords and their descriptions.

Appendix B A alphabetic list of all enumerated concept keywords, with a list of allowed value keywords

The metadata description is entirely case insensitive for anything not in double quotes. By convention, for ease of comprehension (by the human eye),

- concept keywords are written in CAPITAL letters, while
- value keywords which are text strings are enclosed within quotation marks. They are of two types :

  Enumerated text strings which are case sensitive. They are generally written in lower case with a leading capital letter, and underscores instead of spaces.

  - Free text strings may contain any of the 36 alphanumeric characters, plus the following 9 characters : white space .,:; +-or #. Characters with accents are NOT allowed. Some allowed characters are converted "on-the-fly" during metadata ingestion : in particular, the character & (amperes and) will be converted to #038, and consequently "illegal" characters cannot be replaced by their XML or HTML representation. Quotation marks are not allowed in text strings : the ploy of using two consecutive quotation marks to represent a quotation mark within a text string has not been implemented.

The case sensitivity of both enumerated and free text strings is preserved during archiving. However, when searching the metadata for keywords within text strings, case sensitivity will be inhibited.

7.1 Concept keywords

CAA metadata concept keywords are listed in Tables 3 through 14 which are grouped together in Appendix A. They are organised by concept, and these are grouped by affinity (what is measured, where it is measured, how it is represented, etc.).

Note that the concept keywords are unique : that is, the same keyword cannot appear at more than one level of the hierarchy. The only exception concerns keywords of the type "PARENT ...".

7.2 Value keywords

The complete list of enumerated keyword values for Cluster can also be found Appendix B. Values for enumerated keywords must be selected from these allowed values.

7.3 The Physical Parameter

Table 8 indicates how the physical parameter is described in terms of three orthogonal concepts :

ENTITY	This is the object which is studied, the word “object” being used somewhat loosely.
PROPERTY	This is the property of that object which is measured.
FLUCTUATIONS	Often it is the variations of this property which are studied, such as the power spectrum of the fluctuations.

The concept of “Fluctuations” is optional, and the default value (when the concept keyword is absent) is FLUCTUATIONS="Waveform". The vast majority of space plasma datasets consist of time series of measurements of a physical parameter which can be described by ENTITY and PROPERTY alone. This is waveform data. But sometimes the sheer quantity of data requires that some statistical property of this waveform data be extracted to characterise the fluctuations of the physical parameter. The most commonly used statistical property is the Fourier spectral power density, although other properties are also used, notably the wavelet transform.

Note that the concept FLUCTUATIONS refers to the real fluctuations of the physical property, which are measured by the sensor and characterised by some statistically significant parameter which is the dataset parameter. FLUCTUATIONS is not intended to described the random variations due to the experimental error always associated with any physical measurement, although of course this will be what is measured in the absence of any significant physical fluctuations. If desired, the estimated experimental error of each datum can be specified

- in the global metadata if it remains roughly constant over the entire dataset,
- in the caveat file if it varies during the course of the mission, or
- as another parameter in each record if it varies from record to record of from magnetospheric region to magnetospheric region.

It may be helpful to consider the following example. Field parameters, such as the electric or magnetic field, are measured by sensors which are fixed with respect to the spacecraft. On a spinning spacecraft the sensor output is modulated at the spacecraft spin frequency; to avoid aliasing, this signal is sampled at a frequency at least twice the spacecraft spin frequency, and generally much higher. The spin modulation of the resulting waveform data is not described by the keyword FLUCTUATIONS unless it is the level of these fluctuations which is determined. Thus,

- the spin modulated data is simply the waveform of the field in the stated frame of reference. However,
- if this waveform signal is Fourier analysed, there will be a large peak at the spin frequency. The magnitude (and possibly the phase) of this peak (or equivalent information derived by some other method, such as least squares fitting of a sine wave at the spin frequency) will, if determined, be described by the concept keyword FLUCTUATIONS and one of the value keywords enumerated in Appendix B.

The above description of the physical parameter is complemented by

MEASUREMENT_TYPE	which is a more general description of the measurement which is widely used, especially by legacy data centres, and is recommended by SPASE.

This general description (see Appendix B) of the data acquired is associated with the instrument more than with any particular dataset or parameter from that instrument.

7.4 Compound Parameters

Sometimes a physical parameter is derived from more than one physical observable, for example:

- The plasma β, the ratio of the particle pressure to the magnetic pressure, involves the plasma density, the temperatures of the ions and the electrons, and the intensity of the magnetic field. It is a quantity which certainly originates from more than one experiment.
- The Alfvén velocity involves both the density and the magnetic field.

These compound parameters may therefore involve information, such as instrument description, instrument PI and archive scientist, from more than one instrument; these parameters require special attention.

For quantities derived using data from multiple sources, ENTITY="compound" is used to indicate that reference must be made to another concept keyword, COMPOUND. The predicates

COMPOUND="value,parameter_ID1,parameter_ID2,...,parameter_IDn "
COMPOUND_DEF="text string"

define the compound quantity. The keyword “value” defines the compound parameter selected from Appendix B, and “parameter_IDi ” ($1 \leq i \leq n$) identifies the n parameters used to create it. The text string is used to explain precisely how these n parameters have been used, including the algebraic formula employed expressed, for example, in LATEX notation. In principle, this description is even applicable in the case where one or more of the parameters “parameter_IDi” is not archived at CAA, provided the remote archive(s) provide adequately precise “parameter_IDi”s.

7.4.1 Higher-level description

A derived parameter can be described completely only by reference to each and every individual parameter used for its derivation. Nevertheless, it is necessary to provide metadata descriptors for the parameter at all levels of the hierarchy shown in Figure 1.

This information is provided by including the special value ”Multiple” in the enumerated lists of each of the concepts keywords

- EXPERIMENT,
- OBSERVATORY, and
- INSTRUMENT.

The value "Multiple" must be followed by the enumerated values corresponding to each of the experiments, observatories or instruments used. It must be realised that not all possible combinations of the indicated experiments, observatories and instruments are necessarily used to generate the parameter. For example, if two input parameters are used to derive a compound parameter, one could come from an instrument on one spacecraft, the other from a different instrument on a different spacecraft, with no way of knowing which parameter is used from which spacecraft. The information enumerated under ”Multiple” is insufficient to define completely a compound parameter.

At present no mention is made in the CAA metadata of data originating from sources other than “MISSION="Cluster"”. Although such data is perfectly legitimate, its metadata description is beyond the scope of CAA alone. (Work is in progress to resolve this problem.)

7.5 Dataset and Parameter IDs

It is convenient to identify datasets, graphical products, event tables (and possibly other entities, TBD), as well as individual parameters, by a unique identifier (ID).

The CAA Product Naming Convention is described in Ref. 9. The product ID consists of three parts:

1. The dataset ID which is unique to a collection of files with identical characteristics that make up a dataset.
2. The instance ID which is used to identify a particular file within the dataset. It is unique within the given dataset and consists of a date field and a file version number.
3. The file extension that identifies the type of formatting to which the file conforms.

For data objects, CAA supports :

1. The CSDS/ISTP standard (see Ref. 10)

```
Mission_DataType_Source_Date_Version.ext
|------Dataset ID-----||-Instance -||type|
```

Example : C1_PP_ASP_20010201_V01.cef

2. An extended form

```
Mission_DataType_Source_ExtendedID_Date_ExtendedInst_Version.ext
|----------Dataset ID -----------| |--------Instance ------||type|
```

Example : C1_PP_ASP_ION_CURRENT_20010201_0000_0100_V01.cef

The second form is easily distinguished from the first by the use of the double underscore. The first three parts (Mission, DataType and Source) are to be specified by the control authority (in this case the CAA, in conformity with Ref. 9). The data producer, in conjunction with the CAA, will be given some flexibility in supplying meaningful information for the extended fields.

When the date field is not applicable it should be set to the value 00000000.

(It may be noted that identification of the Cluster spacecraft (1, 2, 3 or 4) from which the data has been obtained occurs at mission level in the ID, but at instrument level in the metadata. This difference is largely historical.)

Particular parameters within datasets are identified by their parameter ID and the dataset ID separated by a double underscore, thus

```
ParameterID_DatasetID
```

where it is recommended that ParameterID be of the form

```
Entity_Property_[Fluctuations/Time res]_[Representation]_[Ref frame]
```

Here the different text strings are abbreviations of the keywords enumerated in Appendix B, those enclosed in "[ ]" being optional. For example, the vector magnetic field in Cartesian GSE coordinates obtained from the FGM spin resolution dataset would be

```
B_Vec_xyz_gse_C1_CP_FGM_SPIN
```

where B_Vec_xyz_gse (the part before the double underscore) is the parameter, and C1_CP_FGM_SPIN the dataset. (In this particular case, the time resolution is defined at the dataset level.)

7.6 Vectors and Tensors

None of the concepts presented in this section apply to scalar data.

Some physical parameters can be represented by an array of numbers. A true physical parameter is independent of the coordinate system in which it is measured. Consequently, any array which represents a physical parameter must transform in accordance with strict rules, for example, under rotation of the coordinate system ; these rules ensure that the physical property itself is invariant, and only its representation changes. Arrays which satisfy these rules are called vectors (tensors of order 1) or tensors (tensors of order 2); higher order tensors are also encountered, for example, in

fluid mechanics. In general, arrays of data do not have any such transformation rules. It is therefore important to identify vector and tensor arrays, so that their generic properties may be exploited by application software.

To meet this objective, substantial changes have been made to Cluster Exchange Format, as explained in Ref. 3. These changes are reflected in the Metadata Dictionary. Every effort has been made to maintain compatibility with the CDF implementation for CSDS (see Ref. 2), even at the expense of redundancy of information ; but this has not always been possible. This section explains the changes related to data arrays. Other changes are described in Sect. 7.10

The CAA Metadata Dictionary will eventually exist as a "stand-alone" document, an essential long-term requirement for any metadata dictionary. But during the development phase of CAA the latest version of the document DS-QMW-TN-0010 (Ref. 3) will take precedence in any matter concerning the organisation of multi-dimensional arrays.

7.6.1 TENSOR_ORDER

This is the order in the mathematical sense:

- 1 for vectors such as magnetic field or velocity,
- 2 for tensors such as a plasma pressure, and
- 0 for scalars (the default value).

Note that the rank of the tensor is something different: normally a tensor of order 2 which represents a physical quantity (in a space of three dimensions) has rank 3, but the rank may be reduced under certain conditions. For example, the volumetric tensor which describes the Cluster tetrahedral configuration has rank 2 when the spacecraft are coplanar, and rank 1 when they are collinear.

Often a tensor of order 1 or 2 is incomplete: for example, the Cluster EFW experiment measures the electric field components in only two directions, in the spacecraft spin-plane. Therefore the EFW electric field vector has only 2 components, and the STAFF electric field cross-spectral tensor has only 2×2 components. A vector or tensor with one spatial component missing can still be rotated, but only about the axis parallel to the missing component ; this is useful information for applications. The concept REPRESENTATION (section 7.6.2) enables incomplete vectors and tensors to be described. Note that a tensor of order 2 with missing components is not the same as a tensor of reduced rank (2 in this case). The latter is complete, but one of its eigenvalues is zero (in a direction which can be derived from the corresponding eigenvector) ; this is good scientific information. But the former is incomplete and the information in one spatial direction simply does not exist.

Finally, note that TENSOR_ORDER **must not** be confused with the way in which the elements are ordered in the data records: C-ordering (rather than, for example, FORTRAN-ordering) is always used (see CEF-2 specification, Ref. 3).

7.6.2 REPRESENTATION

This concept must be used to provide essential information describing vectors and tensors :

- the type of representation (Cartesian, polar) used,
- whether the vector or tensor is complete or, if not, which component is missing,
- subscripts for labelling the axes.

Type of Representation

Vectors can be expressed in several ways

xyz orthogonal Cartesian coordinates
rtp (r, θ, φ) spherical polar, θ = colatitude
rlp (r, λ, φ) spherical polar, λ = latitude
rpz (r, φ, z) cylindrical polar

The representations

xy (x, y) plane Cartesian
rp (r, φ) plane polar

are considered to be incomplete examples of first two possibilities and should be described as such (see below).

Other representations may be defined by the user. For example, sensor coordinates, which are not even orthogonal. The CAA Metadata Dictionary will eventually be extended to cover this possibility.

Nature and Representation of the Parameter

Data is stored in multi-dimensional arrays, of which some dimensions may correspond to the tensor dimensionality : vector, tensor (of order 2 or higher). Other dimensions of the data array correspond to variables associated with other independent variables of the data product, such and a particle energy or a spectral frequency channel. The identification and description of these different dimensions of the data array is provided by the concept keywords DEPEND_*i* and LABEL_*i*. The value *i* = 0 corresponds to the most slowly varying variable which, for the vast majority (all) of CAA datasets, corresponds to time of measurement, which increases from the beginning to the end of the dataset. Nevertheless, other variables could be used as principal key.

Each dimension *i* of a data array must be identified by

- either DEPEND_*i*,
- or LABEL_*i*,

but **never both**. When present, LABEL_*i* provides partial (subscript) information for labelling graphic representations of the data (the "root" of the label is provided by LABLAXIS, see Table 13). Furthermore,

- DEPEND_*i* is used (instead of LABEL_*i*) if the dimension of the data array represents something like an energy, or frequency, etc., rather than a spatial dimension.
- REPRESENTATION_*i* must be used, generally with LABEL_*i* (but never with DEPEND_*i*), if the i^{th} dimension of the array represents a dimension of a vector or tensor quantity.

The use of REPRESENTATION_*i* ensures that the vector or tensor nature of the dimension is described with adequate information to permit generic software to be used to perform vector or tensor operations, plot hodograms, etc..

In addition, REPRESENTATION_*i* indicates how the vector or tensor is represented. For example, a vector in Cartesian coordinates would be identified by a predicate of the form

REPRESENTATION_*i* = "x", "y", "z"

where *i* is the independent variable of the data array which represents the vector components. For a tensor of order 2 there would be two predicates,

REPRESENTATION_*m* = "x", "y", "z"
REPRESENTATION_*n* = "x", "y", "z"

where *m* and *n* are respectively the independent variables of the data array which represents the tensor elements.

For an incomplete vector or tensor with one component missing, for example the z-component, the above predicates would become respectively

REPRESENTATION_*i* = "x", "y"

where *i* is the independent variable of the data array which represents the vector components. For a tensor of order 2 there would be predicates such as

REPRESENTATION_*m* = "x", "y"
REPRESENTATION_*n* = "x", "y"

Polar Coordinates

Polar coordinates are handled likewise. For example, a two-dimensional (incomplete) vector expressed in plane polar coordinates r, φ would be described by the predicate

REPRESENTATION_*i* = "r", "p"

The two or three elements of the representation of a vector in polar coordinates do not have homogeneous units. Therefore the predicates for the concepts UNITS and SI_CONVERSION (Table 9) have their syntax extended, thus :

UNITS = "km", "degree", "degree" in spherical polar coordinates, or
UNITS = "km", "radian", "km" in cylindrical polar coordinates,

with similar extensions of syntax for SI_CONVERSION.

The Diagonal Elements of a Tensor

The diagonal elements of a symmetric or Hermitian tensor have special significance. They do not form a vector because they cannot be rotated like a vector ; the complete tensor is required for that. But their sum, which is called the "trace" of the tensor and is equal to the sum of the eigenvalues, is invariant under rotation of coordinates. The sum of the diagonal elements of a Hermitian cross-spectral matrix of vector field fluctuations gives the total power spectral density of the fluctuations; and the three individual elements, if different, indicate that the field fluctuations are not isotropic (although if they are the same the field is not necessarily isotropic).

The diagonal elements of a tensor of order 2 are identified by

REPRESENTATION_*i* = "xx", "yy", "zz"

where i is the independent variable of the data array which represents the diagonal components.

7.7 Reference Frame

It is clearly essential to specify the reference frame in which vector or tensor data is supplied. A reference frame is specified by :

- the coordinate system used to specify the orientation of the axes, and
- the origin of the reference frame, including its vector velocity. The latter is particularly important for measurements of the plasma convection velocity or the electric field (due to the induced $\mathbf{v} \times \mathbf{B}$ field).

7.7.1 COORDINATE_SYSTEM

Allowed coordinate systems are listed in Appendix B. GSE is the preferred CAA coordinate system for representing vectors. The following systems used by Cluster merit further explanation:

SC The "SpaceCraft" coordinate system of this document is called the Attitude System in the Data Delivery Interface Document (DDID, Ref. 11), Section I.1.3.2 (page 136). It has the same axes as the Body-Build System of the Experiment Interface Document (EID), but they are permuted cyclically so that the z-axis of the SC system lies along the nominal spin axis, as is the case for the vast majority of spin-stabilised spacecraft.

SR The "Spin Reference" system is defined in Section I.1.3.2 (page 136) of the DDID (Ref. 11) ; its z-axis is the maximum principal inertia axis of the spacecraft. It is therefore close to, but not identical with, the SC system ; the transformation matrix, which may vary slowly with time (e.g., as fuel is used), is defined on page 137 of the DDID ; the matrices for each spacecraft are archived in CAA as part of the TBD dataset. Although this coordinate system is an essential step in the transformation between SC and SR2 coordinates, it is unsuitable for archiving (not fixed with respect to the SC system) and not listed in Table 10.

SR2 This coordinate system is the despun SR coordinate system : it has its z-axis aligned with the SR z-axis, and its x-axis in the meridian containing the direction of the Sun.

ISR2 The Inverted SR2 coordinate system. Since the Cluster spin axes are near (≈ 6º) the direction of the southern ecliptic pole, rotation of the SR2 coordinate system through 180º about its x-axis brings it close (within ≤ 6º) to the GSE system.

It should be noted that as the satellite spin axes are not exactly parallel, the SR2 and ISR2 systems are slightly different for each spacecraft.

In addition there are two coordinate system related to the magnetic field:

FAC The Field-Aligned Coordinate system. The **Z** axis is parallel to the magnetic field vector **B**. The **Y** axis is defined with **B** and the direction of the Earth from the satellite position, **R**, being parallel to **B** × **R**. The **X** axis is parallel to **Y** × **Z**.

MFA The Magnetic Field Aligned system. The **Z** axis is parallel to the magnetic field vector **B**. The **Y** axis is defined with **B** and the direction of the Sun from the satellite position, **S**, being parallel to **B** × **S**, and **X** is parallel to **Y** × **Z**)

The above coordinate systems are used by all instrument teams. There are also a number of coordinate systems which are either instrument-specific or data-specific.

Instrument Instrument specific coordinate systems may be defined, for example, by the axes $\mathbf{I}^i$ of three sensors aligned in different directions. The Instrument system may be described by the projection of each of the axes $\mathbf{I}^i$ onto each of the axes $\mathbf{S}^j$ of the orthogonal SC system. The matrix $T_{ij} = \mathbf{I}^i \cdot \mathbf{S}^j$ transforms any vector x_i^{SC} expressed in the SC system to its representation x_j^{Inst} in the Instrument system, thus $x_i^{\mathrm{Inst}} = \sum_{j=1}^{3} T_{ij} x_j^{\mathrm{SC}}$. Provided that the three sensors axes $\mathbf{I}^i$ are not coplanar $|T_{ij}| \neq 0$ and this equation can be inverted, $x_k^{\mathrm{SC}} = \sum_{i=1}^{3} T_{ki}^{-1} x_i^{\mathrm{Inst}}$ where T_{ki}^{-1} is the inverse of T_{ik}, that is $\sum_{i=1}^{3} T_{ki}^{-1} T_{ij} = \delta_{kj}$.

The matrix $R_{ki} = T_{ki}^{-1}$ may be included in the metadata as a support parameter which completely describes the instrument specific coordinate system in terms of some standard system, such as SC. Furthermore, the transformation from the Instrument coordinate system to this standard system is given by

$$x_k^{\mathrm{SC}} = \sum_{i=1}^{3} R_{ki} x_i^{\mathrm{Inst}} \qquad (1)$$

If the Instrument system is orthogonal then T_{ij} is unitary, $|T_{ij}| = 1$ and $R_{ki} = T_{ki}^{-1} = T_{ij}$, and the matrix R_{ki} represents a pure rotation. If the Instrument system is not orthogonal Eq. 1 is still correct. Only in the rather unlikely event of wanting to transform data from a standard system to a non-orthogonal instrument system is care required to ensure that the condition $|T_{ij}| \neq 1$ is handled correctly.

Many instruments would like to make data available in an instrument-related coordinate system. Two cases can be distinguished:

1. The instrument axes are fixed with respect to the SC coordinate system, as for magnetic and electric sensors rigidly attached to the spacecraft structure. The matrix R_{ki} for each spacecraft remains unchanged throughout the entire mission, and can be included once and for all in the instrument global metadata.
2. The data is pre-processed aboard the spacecraft. In this case the coordinate system can change with telemetry mode, especially for instruments such as particle detectors whose operation is synchronised with the satellite Sun-sensor. In this case the above description of the Instrument coordinate system can still be used but, depending upon how the data is organised into datasets, the matrix R_{ki} may or may not be a global attribute of the dataset(s).

The description of coordinate transformations is discussed in Section 7.8. In all cases there should be a textual description of the "Instrument" specific coordinate system in one of the higher level metadata elements. So for example it could be specified in either the DATASET_DESCRIPTION, or in the EXPERIMENT_DESCRIPTION if the same coordinate system is used for many products.

Data (not to be confused with DATA used in a totally different context in Section 7.10.2) Data specific coordinate systems are derived by applying algorithms to observations of an anisotropic scientific event, such as an individual boundary crossing, or an interval of plasma turbulence. The resulting coordinate system is specified in terms of a rotation with respect to some standard coordinate system (plus possibly a translation, as in the case of the deHomann-Teller frame). Rotations to a "Data" system will be used to transform vector data of arbitrary origin from a standard coordinate system to one related to the specific event. Therefore it is useful to define the transformation in the sense

$$x_k^{\mathrm{Data}} = \sum_{i=1}^{3} R_{ki} x_i^{\mathrm{GSE}} \; ; \qquad (2)$$

so that transformation software does not need to invert the matrix. There are two significant differences with respect to a rotation describing an "Instrument" system:

1. The transformation of Eq. 1 converts from specific to standard coordinates, while that of Eq. 2 transforms from standard to specific coordinates. Both transform is the sense in which they are expected to be used.
2. There is a clear possibility of the simultaneous existence of more than one version of data which is nominally the identical, yet differs due to being derived using different algorithms. The metadata as described in Sect. 7.8 would probably (TBD) need to be enriched to take account of this possibility. For this reason a more complete description of the metadata for "Data" related coordinate systems will be provided in a future version of the CAA Metadata Dictionary.

GSE is the preferred CAA coordinate system for representing vectors. As the name implies, strictly speaking Geocentric Solar Ecliptic coordinates have their origin at the centre of the Earth while, for Cluster, the measurement frame is the satellite-centred. Therefore it may be necessary to specify:

7.7.2 FRAME_ORIGIN

FRAME_ORIGIN, which may be used when the origin of coordinates **is different** from the origin implicit in the name of the coordinate system used. This does not apply to field vectors, whose magnitude indicates the magnitude of the physical parameter, not a distance from the origin of coordinates; but it does apply, for example, to spacecraft position. For Cluster, FRAME_ORIGIN is used only to indicate satellite-centred GSE coordinates ; for SC, SR2 and ISR2 coordinates, the spacecraft-centred origin of coordinates is implicit.

Note that the x-axis as determined by the solar sensors is in a plane containing the satellite-Sun line, not the Earth-Sun line. However, as Cluster orbits the Earth with an apogee of only ~20RE, the angular difference between the determined x-axis and the true GSE x-axis is not measurable. (This is not the case for interplanetary spacecraft at large distances from the Earth.)

7.7.3 FRAME_VELOCITY

The value of some physical parameters, for example the electric field or the plasma flow velocity, depend upon the motion of the coordinate system in which they are measured. So it is often useful to transform (by translation, not rotation) the measured quantity to a reference frame moving at some other velocity. FRAME VELOCITY defines the motion of the origin of this coordinate system. It takes the values :

- Observatory — when no correction has been applied
- Inertial — when the field (expressed in GSE or any other convenient coordinate system) has been corrected for the effects of the spacecraft velocity **v** with respect to inertial (GEI) coordinates
- Earth_Corotating — which is useful for ionospheric studies.

7.7.4 FRAME

FRAME is partially redundant with the more powerful description provided by the three concepts TENSOR_ORDER, REPRESENTATION, AND REFERENCE_FRAME. It is conserved for backward compatibility: that is, to allow CAA files to be read (but not fully exploited) by systems capable of interpreting only the information contained in files of Version 1 of Cluster Exchange Format. See Ref. 2, page 57 for further details.

7.8 Coordinate rotations

A vector quantity x is represented by its projections xi onto a system of coordinate axes. When the coordinates change, the quantity itself does not change, but its representation does. In the case of transformation by rotation of the coordinate system, the representation x_i^b in terms of axes b may be derived from the representation x_j^a in terms of the orthonormal axes a

$$x_i^b = \sum_{j=1}^{3} R_{ij}^{ba} x_j^a \qquad (3)$$

where $R_{ij}^{ba} = \mathbf{b}^{(i)} \cdot \mathbf{a}^{(j)}$; that is, the projections of the three axes $b^{(i)}$ of the new system onto the three axes $a^{(j)}$ of the old system. If the system b is also orthogonal, then the transformation is unitary, $R_{ij}^{ab} = R_{ji}^{ba} = (R_{ij}^{ab})^{-1}$ *and* $|R_{ij}| = 1$. Expressions similar to eq. 3 exist for tensor quantities: for a tensor x_{ij} of order 2 the rotation is

$$x_{ij}^b = \sum_{k=1}^{3}\sum_{\ell=1}^{3} R_{ik}^{ba} R_{j\ell}^{ba} x_{k\ell}^a$$

and, quite generally, for a tensor of order n the matrix R^{ba} is applied n times.

Such a transformation matrix may be considered to be a datum (or data object) which happens to represent a rotation of coordinates. In general these coordinate transformation data objects vary with time.

- Some vary rapidly, such as the transformation from a rotating spacecraft frame of reference to an inertial frame of reference. The corresponding matrices, which are not tensors because they "span" coordinates systems, may nevertheless be handled as a time series in a way similar to that of a second order tensor observable, such as the pressure tensor.
- Others vary regularly, but slowly, such as the rotation from the inertial GEI system to GSM coordinates (variations with a period of one day) or from GEI to GSE (period one year). (See Appendix B for definitions of the coordinate systems.)
- Others vary irregularly, but are valid during a certain interval of time. These include the rotation from SR2 to GEI coordinates, which changes each time the spacecraft performs a spin axis correction manoeuvre.

Two types of coordinate rotation have already been used in Section 7.7 to define new coordinate systems:

- Coordinate systems which are an essential part of the support data. The transformation from experiment coordinates to spacecraft coordinates described by Eq. 1 is stable either for the duration of an instrument mode of operation, or for an entire dataset, thus allowing the description of the rotation to be part of the dataset global metadata. This transformation is interesting because the sensor coordinate system can be significantly non-orthogonal: the transformation is still given by Eq. 3 (c.f. Eq. 1), but with a matrix R_{ij}^{ba} which is not unitary.
- Coordinate systems which are science data dependent, such as the LMN (boundary normal) coordinates, useful when studying a shock or discontinuity. The transformation from GSE to LMN is assumed to be stable for the duration of the boundary crossing for which it is defined, and rotation matrices for successive (multiple) shock crossings could usefully be assembled into an event file.

Rotations are data objects which must be described like any other non-scalar parameter. A rotation can be represented in many ways: by its Euler angles φ, θ, ψ, by a quaternion q_0,q_1,q_2,q_3, or by a rotation matrix R_{ij} , $1 \le i, j \le 3$. The Euler angles express the essential information: the value of the rotation about an axis which is specified by two angles measured with respect to a known coordinate system. Euler angles are easy to visualise but difficult to use. Quaternions are the opposite: the rotation is expressed in terms of components along generalisations of $\sqrt{(-1)}$ in a four-dimensional complex hyperspace. Quaternions the preferred representation for intensive computational applications (computer graphics, inertial navigation systems) for two reasons: they are easier to interpolate and, because the same information is contained in only 4 numbers (instead of 9 for a rotation matrix), they yield greater computational speed and precision. Nevertheless, CAA considers only the rotation matrix, the representation with which the majority of space physicists feel most at ease.

Like all other parameters, coordinate rotations are described by ENTITY and PROPERTY :

```
ENTITY      =     "Transformation"
PROPERTY    =     "Coordinate_rotation"
```

The first line indicates that the datum represents some sort of coordinate transformation (rotation, translation, or both), while the second adds some specificity. But much more information must be available and understandable to both the potential user and generic applications software:

- A rotation matrix is a datum, and as such it needs a parameter ID to identify it as clearly as possible, and preferably uniquely. The ID should indicate the input and the target coordinate systems, its time interval of validity and, if appropriate, the physical instrument (e.g. STAFF-2) or the type of event (it e.g. bow shock crossings) to which it is relevant. Possible extensions of the naming conventions of Sect. 7.5 are under study.
- The name of the input coordinate system a (on the right side of Eq. 3).
- The name of the output coordinate system b (left side of Eq. 3), that is, the coordinate system into which vector or tensor input data is converted.
- For data-dependent transformations such as boundary normal, the data set and analysis method used to derive the transformation (see section 7.7).

The data itself consists of records containing the elements of the matrix R_{ij}^{ba} (c.f. eq. 3), arranged in the order of the C programming language. For rotations which form part of the global metadata the elements will be entered by means of a DATA statement (see section 7.10.2).

All rotations transform vectors and tensors from a system identified by COORDINATE_SYSTEM="Keyword", towards a system identified by TARGET_SYSTEM="Keyword". The allowed “Keyword” values for COORDINATE_SYSTEM are enumerated in the lists in section 7.7.1 and Appendix B. Coordinate rotations can be divided into three types: generic, and two types identified in Sect. 7.7.1 in connexion with the definition of particular coordinate systems:

- Generic coordinate rotations transform between two systems, both of which are on the enumerated list for COORDINATE_SYSTEM, but with "Keyword"="Instrument". These rotations may well be implemented already as part of the user’s data analysis system.

- Instrument-specific coordinate rotations transform from an instrument-specific system to a standard system. They can be recognised by COORDINATE_SYSTEM="Instrument". Unlike other keywords on the COORDINATE_SYSTEM enumerated list, the keyword value "Instrument" does not define a recognised standard coordinate system, it serves:
  - to indicate that the rotation is “Instrument-type”, and
  - to ensure that XML metadata description validates against the appropriate

  XML schema file. Instrument-specific coordinate rotations serve as support data for instrument-related applications, and it is these application which know precisely how to use the rotation matrix. The correspondence between the coordinate rotation and the physical parameter to which it applies it established by means of the

  START_VARIABLE = Name of rotation

  which is part of the CEF syntax (see examples below, this has not yet been described in the MDD).

- Data-specific coordinate rotations transform from the system COORDINATE_SYSTEM="Keyword" where "Keyword"="Instrument" to a system identified by TARGET_SYSTEM="Keyword".

The "Keyword" values for TARGET are enumerated; the list includes all keywords on the list for COORDINATE_SYSTEM, plus agreed keywords describing "recognised" coordinate systems, plus the keyword "Other" for use when the coordinate system is not on the enumerated list; it should then be adequately described by free text via DATASET_DESCRIPTION (Table 7).

The following descriptions are examples of the above rules applied to the two types of coordinate rotation.

7.8.1 Example 1 : Instrument rotation from GSE coordinates to GSM coordinates

```
START VARIABLE = GSE2GSM
       ENTITY               ="Transformation"
       PROPERTY             ="Coordinate_rotation"
       SIZES                = 3, 3
       VALUE_TYPE           = DOUBLE
       COORDINATE_SYSTEM    = "GSE>Geocentric Solar Ecliptic"
       TARGET_SYSTEM        = "GSM>Geocentric Solar Magnetic"
       FIELDNAM             = "R Matrix GSE to GSM"
       REPRESENTATION_1     = "x","y","z"
       REPRESENTATION_2     = "x","y","z"
       TENSOR_ORDER         = 2
       SI_CONVERSION        = "1>unitless"
       UNITS                = "unitless"
       PARAMETER_TYPE       = "Support Data"
       CATDESC              = "Transformation GSE to GSM coordinates"
END VARIABLE = GSE2GSM
```

7.8.2 Example 2 : Instrument rotation from STAFF SSW6RF coordinates to spacecraft coordinates

```
START VARIABLE = STAFF_SSW6RF2SC
       ENTITY               ="Transformation"
       PROPERTY             ="Coordinate_rotation"
       SIZES                = 3, 3
       VALUE_TYPE           = DOUBLE
       COORDINATE_SYSTEM    = "Instrument"
       TARGET_SYSTEM        = "SC>Spacecraft"
       FIELDNAM             = "R Matrix from STAFF SSW6RF"
       REPRESENTATION_1     = "x","y","z"
       REPRESENTATION_2     = "x","y","z"
       TENSOR_ORDER         = 2
       SI_CONVERSION        = "1>unitless"
       UNITS                = "unitless"
       PARAMETER_TYPE       = "Support Data"
       CATDESC              = "Transformation from STAFF search coil to SC coords"
END VARIABLE = STAFF_SSW6RF2SC
```

7.8.3 Example 3 : Data rotation from GSE coordinates to LMN boundary normal coordinates

```
START VARIABLE = gse2lmn
       ENTITY               ="Transformation"
       PROPERTY             ="Coordinate_rotation"
       SIZES                = 3, 3
       VALUE_TYPE           = DOUBLE
       COORDINATE_SYSTEM    = "GSE>Geocentric Solar Ecliptic"
       TARGET_SYSTEM        = "lmn xyz"
       FIELDNAM             = "R Matrix to lmn"
```

```
        REPRESENTATION_1    = "x","y","z"
        REPRESENTATION_2    = "x","y","z"
        TENSOR ORDER        = 2
        SI_CONVERSION       = "1>unitless"
        UNITS               = "unitless"
        PARAMETER_TYPE      = "Support Data"
        CATDESC             = "Transformation from GSE to LNM>Boundary Normal coordinates"
END VARIABLE = gse2lmn
```

7.9 Arrays

Other data can be presented in arrays which have nothing to do with any vector or tensor nature of the data : particle energy spectra, wave dynamic frequency spectra, etc. see Sect. 7.6). This is done in conformity with the prescription of the Cluster Exchange Format, Version 2 (CEF-2) given in Ref. 3. As stated in Sect. 1, this version of the Metadata Dictionary tries to avoid unnecessary duplication of information ; this leads to numerous cross-references between this document and Ref. 3.

7.9.1 DEPEND_0 and DEPEND_i

DEPEND_i is used to specify bin dimensions for arrays other than those due to the vector or tensor nature of the data. In particular, DEPEND_0 does the same for the most important parameter used for ordering the data. This is generally, but not necessarily, the time. For further details, see Ref. 3.

7.10 Additional Information

Whilst use of the keywords is self-explanatory in the majority of cases, there are nevertheless certain keywords which merit additional explanation.

7.10.1 Complex Data

Complex data values are represented by an extra dimension taking a LABEL_i indicating either "Re" or "Im". If this is the last index then the real and imaginary parts of a value are successive entries in the record.

7.10.2 DATA

This is used to provide the values of variables which are fixed for all records of a dataset. It is especially useful for support parameters (see PARAMETER_TYPE, section 7.10.8), such as the energy channels, or frequency channels, used for spectral measurements. These “non-record-varying” variables should be supplied within the header variable metadata segment, and no entry is then allowed in the data records. The presence of a parameter “DATA” will be taken to indicate that this is a non-record-varying variable, and the value(s) associated with this parameter are the data for that variable for the entire dataset. They are comma separated. Elements of arrays will be written with the natural C ordering (last index varies fastest), and data lines for arrays may be continued using “\” as a continuation marker following one of the commas separating the list of values.

7.10.3 ERROR_MINUS and ERROR_PLUS

All experimental measurements have an associated error which, ideally, should be provided to potential users.

The error may often be estimated from the characteristics of the instrument. For example, a spectrum analyser observing with a bandwidth B with integration time t has an associated fractional statistical error of $1/(B\Delta t)$ which is physically unavoidable (the “uncertainty principal”) and constant in time. The relevant information will be provided in the global metadata.

Some instruments may perform, for example, averaging (in flight or on the ground) to find a mean value which is the archived parameter. For example, measurements of a particle flux count rate may be averaged, and the fluctuations of the count rate around this mean value may have one of two possible interpretations :

1. Indicate the level of the natural fluctuations of the count rate, which is a useful physical observation in its own right. In this case the level of fluctuations should be treated as an extra physical parameter in the data record.
2. Arise only from the experimental uncertainty in the measurements of individual data points, and hence (by dividing by √N) of the archived quantity itself ; in this case use of ERROR_MINUS and ERROR_PLUS is indicated.

7.10.4 INSTRUMENT_TYPE

This concept supplies information about the nature of the instrument used to obtain the data, including sensor and other hardware or software features as appropriate. Often the same physical parameter can be obtained from more than one type of instrument. For example, electron density can be derive from particle counter data, Langmuir probe data, and wave spectral data, and active resonance sounder data. (Not to be confused with MEASUREMENT_TYPE).

7.10.5 MEASUREMENT_TYPE

As explained in Sect. 7.3, this concept is somewhat redundant with ENTITY, PROPERTY and FLUCTUATIONS. It is included here mainly for compatibility with legacy data systems operating within the SPASE initiative; these systems may not have the detailed information provided by ENTITY and PROPERTY. The values of MEASUREMENT_TYPE are listed in Appendix B, and they are more naturally associated with the description at the instrument level, in Table 6. (Not to be confused with INSTRUMENT_TYPE).

7.10.6 Personnel Coordinates

At the mission level (MISSION_KEY_PERSONNEL), experiment level (INVESTIGATOR_COORDINATES, EXPERIMENT_KEY_PERSONNEL) and dataset level (CONTACT_COORDINATES) information is required concerning the identity, role and e-mail address of persons with specific responsibility within the Cluster mission. This information is to be provided in the standard format,

Name > role > e-mail address

7.10.7 PROCESSING_LEVEL

Four processing levels are defined for Cluster Active Archive:

Raw (also known as L0). This is data which has not been reorganised in any way, except for demuliplexage of the science data packets from the different instruments of the same spacecraft.

Uncalibrated (or L1). Data which has been reorganised into a format more adapted to the ensuing scientific processing, but without any processing which is in any way irreversible. This is the highest level of "processed" data product which is expected never to improve with time.

Calibrated (or L2). This is data which has been converted into physical units ; it is the data which the instrument was conceived to supply, and which CAA will supply, to the scientific community. The conversion procedure may vary in time, as knowledge of instrument performance including possible drift of the calibration coefficients is better understood. Reverse processing is generally impossible: if required, any new dataset would have to be produced from "Uncalibrated" data.

Derived (or L3). "Calibrated" science data is frequently used to produce further science data product(s) of high added value, but whose production involves the application of scientific interpretation or substantial numerical calculation (or both). Examples of two such Cluster datasets are the EFW electric field after removal of the $\mathbf{v}\times\mathbf{B}$ field induced by the spacecraft orbital motion

across the Earth's magnetic field, and the electron density derived from the interpretation of the wave spectra from the WHISPER sounder (active or passive).

Auxiliary Sometimes a dataset includes other parameter(s), which are basically support or status information.

7.10.8 PARAMETER_TYPE

Parameters within a dataset are of two types :

Data These are the primary scientific parameters of the dataset.

Support_data These are parameters which provide information about properties of the dataset, such as energy bins of a particle detector, frequency channels of a spectrum analyser, or other such information which, while not being science data, describes the science data. These parameters may or may not be record-varying ; in the latter case, their values may be defined for the entire dataset using the "DATA" declaration described above (section 7.10.2). Support data cannot be specified once per file: it must be defined either once for the entire dataset, or on a per-record basis.

7.10.9 SI_CONVERSION

The conversion is expressed in terms of one of the SI units enumerated in Appendix B. It is a text string of the form number>SI unit , where number is the conversion factor to SI units. It is the factor by which the value of the variable must be multiplied to convert it to SI units. The string SI unit is the standard unit to which it converts. For example, a magnetic field in nT would have the predicate

```
SI_CONVERSION="1.0e-9>T"
```

For compound units the grammar will be of a standard form: distinct unit dimensions will be separated by space characters and powers (signed) will be preceded by the carat, . Non-dimensional qualifiers, which do not appear in the SI units list, are to be enclosed in braces "()", for example,

```
"m s^-1" or "(number electrons) m^-3" .
```

Non-integer powers are permitted, e.g., "Hz^–0.5". Similarly, braces may be used to provide user information (e.g., for labelling axes) on dimensionless quantities, such as

```
SI_CONVERSION="1.0E-2>(fraction)"
```

for a percentage.

7.10.10 SIGNIFICANT_DIGITS

This keywords is required at the parameters level to know the precision with which format conversion must preserve the data, for example, when converting from binary to CEF format.

7.10.11 TIME_RESOLUTION

The time resolution of numerical data is expressed in seconds, at the dataset level. If the time resolution is not constant, e.g., due to changes of telemetry rate, instrument or data processing mode, the two further keywords, MIN_TIME_RESOLUTION and MAX_TIME_RESOLUTION, must be used. Furthermore, in this case, DELTA_PLUS and DELTA_MINUS must be local (i.e., not global) attributes if the ratio of the sampling interval to the time resolution remains constant.

The attributes TIME_RESOLUTION, MIN_TIME_RESOLUTION and MAX_TIME_RESOLUTION apply only to time series data (products of type CP in Table 2), but not to event tables (CT) or quality/caveat (CQ) files.

Although the definition of MIN_TIME_RESOLUTION and MAX_TIME_RESOLUTION resolutions in Table 7 is unambiguous, there has been, perhaps not surprisingly, some confusion between the two. Users of CAA data are advised to do their own check to determine which is which.

7.10.12 VALUE_TYPE

This defines the type of information held within the field of a CEF file. It is essential for interpreting the ASCII of a CEF file. Allowed values are enumerated in Table 12. Note that

- time implies an ISO standard time code, and
- time span indicates the standard ISO specification of a time range, which is two successive ISO standard time codes separated by the sign "/".

7.11 Information Lists

Several concept keywords will point towards lists, including :

List	Description
Key mission personnel	A list names points of contact of key mission personnel, past and present
CAA personnel	idem, for Cluster Active Archive
Instrument team personnel	One list for each instrument, plus ESOC and JSOC: a list of pointers to items in the "Cluster Community" database.
Instrument bibliography	One list for each instrument, plus ESOC and JSOC Mission bibliography

These lists will contain pointers to individual persons, or bibliographic references. Again to reduce duplication of information, use will be made of databases which will be maintained by Cluster Active Archive for the Science Working Team.

Database	**Description**
Cluster Community database	The "Directory of Cluster Community Members" maintained by the Cluster Project Scientist, or something derived from it.
Bibliographic database	This will contain abstracts and pointers to sites which can give the full text, and possibly actual texts when desirable and not prevented by copyright considerations.

7.12 Miscellaneous Background Information

The keywords used for CAA have been recovered from several sources or existing metadata dictionaries, including :

- metadata included in the CEF file format specifications (ref. 3)
- the ISTP recommendations (ref. 7)
- CDPP

or from recommendations which will be ready soon, such as SPASE. SPASE is rather an interesting case, because it is a data dictionary recommendation which has been compiled jointly by several collaborating data centres to facilitate the implementation of interoperability. As far as possible SPASE recommendations (as of April 2004) have been followed by CAA, but they are neither binding nor necessarily complete.

8 Metadata Keywords for Other Products

Besides numerical data, Cluster Active Archive will preserve various other products. These must be described in sufficient detail for the would-be user to be able to find what he needs via his local CAA, or a world-wide SPASE-compatible, search interface. It is unlikely that these products will be used by application programs, although this cannot be totally excluded; for example, in the case of libraries of executable code (not presently foreseen for CAA).

The metadata for these products are discussed in the following sections. As for the numerical data products, the many of these descriptions are hierarchical, and frequent reference will be made to the tables of Appendix A.

8.1 Preplotted graphics

These are graphical products which have been produced by the PI and his team to facilitate the analysis of data from their experiment. Compared to use of the CAA interactive graphics, pre-plotted graphics offer the following advantages:

- The plots are immediately available. Those in png format are instantaneously available for visualisation on the end user's web browser, whilst those in PostScript format are available with a quality suitable for publication (subject to PI approval). Some plots may be available in both formats.
- The plots have been produced by the PI teams, and therefore :
    - they are plots which have been developed for their scientific utility ;
    - they show selected parameters plotted with optimum resolution ;
    - their utility has been validated by the PI team.

Their principal disadvantage is, of course, that they are not tailor-made to the CAA user's particular requirements of the day.

8.2 Event tables

Event tables differ from science data in that their records are characterised by two time tags, rather than one. Of course, each experiment datum is acquired during a certain interval of time described by DELTA- and DELTA+, and a greater sampling rate would enhance the information content of the data. This is not true for event tables, whose records are characterised by a pair of times: a start-validity t1 and an end-validity t2 time, with t1 <t2. Successive records are generally, but not necessarily, contiguous in time; gaps may occur, and even overlaps if the table contains more than one parameter (but overlaps must not contain contradictory information !).

Event tables may provide useful information in their own right, and/or supply conditions to be satisfied, for example, when searching for scientific data. Event tables may be used for at least two types of data:

Status data This data is quantised, so that it remains valid for some well-defined interval of time, then changes to another value valid during another interval, possibly of different duration. Data of this type include:

- spacecraft status (including telemetry mode);
- experiment status;
- data concerning the predicted (model based) region of the magnetosphere;
- geomagnetic activity indices, e.g., the Kp index has one value every 3 hours.

Scientific event data This includes observed crossings of the bow shock, the magnetopause, and other such features which have an "identifiable" start time and end time.

There are two reasons for placing status and scientific event information into a common type of archival entity:

- scientific, to be able to use both type of table interchangeably, e.g., for data searching;

- technical, to have only one type of data entity and associated generic software. It is shown in Appendix C that both types of data can indeed be included in the same type of data product, subject to certain conditions.

8.3 Caveat sets

Special event-type set containing information about the quality of the archived products. These sets will updated more frequently than the datasets, graphics sets, or other product to which they refer.

8.4 Calibration Data

Data returned from the instrument internal calibration cycle. Section still to be completed.

8.5 Documentation

All data, data products, and software products will be will be fully documented. The following standards are acceptable for documentation :

- html
- LaTeX
- Microsoft Word
- pdf
- plain ASCII.

Section still to be completed.

8.6 Software Products

CAA will provide access to software which may be recovered for exploitation at the user's home site. In addition to documentation as described in Sect 8.5, such software should be archived together with :

- the installation procedure,
- test data which will fully exercise all features of the software, and
- the corresponding output which should be obtained this running this test data. The CAA cannot accept responsibility for the maintenance of software provided by external groups.

Section still to be completed.

9 Representation of the Information

A representation of the concepts of entity, property and fluctuations which is readily understood is required, for example, for use on html pages or other documents examined by the human eye.

Let us consider a simple example, the electron density. This will be described by the predicates

```
ENTITY   =     electron
PROPERTY =     density
```

and may be represented by

```
electron>density
```

Similarly, for the magnetic field vector, the predicate representation is

```
ENTITY   =     Magnetic_field
PROPERTY =     Vector
```

and this will commonly be represented by

```
Magnetic_field>Vector
```

Very often it is the fluctuations of the vector field which are measured. Then is it necessary to add another predicate,

```
FLUCTUATIONS=spectrum
```

while the common representation becomes

```
Magnetic_field>Vector>spectrum
```

Another possibility is

```
FLUCTUATIONS=cross-spectrum
```

represented by

```
Magnetic_field>Vector>cross-spectrum
```

The absence of a predicate involving the concept `FLUCTUATIONS` is equivalent to the predicate `FLUCTUATIONS="waveform`", that is, the default is

```
Magnetic_field>Vector>waveform
```

It is also possible, in principle, to measure electron density fluctuations (plasma waves), although this term is usually used to describe the associated electric field fluctuations which must be present so as to satisfy Maxwell's equations)

```
ENTITY         =     electron
PROPERTY       =     density
FLUCTUATIONS   =     spectrum
```

represented by

```
electron>density>spectrum
```

We may note that it becomes possible to perform a search for experiments which measure waveform of either field OR of particle parameters. Such measurements do indeed exist: for Alfvén waves in the solar wind, and they are combined to derive the solar wind helicity, an important parameter in the study of turbulence.

10 Acquisition of the Information

Since the data description is designed to be both machine readable and also intelligible to the human eye, thought must be given to its representation.

Up to this point, all considerations have been scientific in nature. But to be able to order and handle the information automatically, it must be formatted and organised so as to be machine readable. The currently popular method of doing this is to use a mark-up language, and in particular, XML.

It may be desirable to develop tools :

- To validate the compatibility of the information received with the CAA metadata data dictionary. That is, verify that the concepts keywords used, and their associated value keywords, are compatible with the CAA dictionary. Such validation could be implemented using XML schema.
- To aid acquisition of the information. For example, in principal it is possible, using a XML Schema, to develop code which would display possible concept keywords and for each display the corresponding value keywords. Being forced to selected keywords from a list, the user would be both prompted for ideas, and be prevented from entering illegal (but not wrong!) information. Note that such software may become commercially available within the next year of so.

The above two initiatives are not within the scope of the current Metadata Dictionary Work Package.

11 Reference Documents

1. Users Guide to the Cluster Science Data System, edited by P.W. Daly, DS-MPATN-0015, Issue 1, Rev. 1, dated 19 April 2002, pdf file available from ftp://ftp.estec.esa.nl/pub/csds/task for/users guide/csds guide.html

2. Reference Document for CSDS CDF Implementation, by A.J. Allen, S.J. Schwartz and D. Burgess, DS-QMW-TN-0003 Issue 1, Rev. 6, dated 9 July 1999, ps file available from http://www.space-plasma.qmul.ac.uk/DOC/DS-QMW-TN-0003.ps

3. Cluster Exchange Format -Data File Syntax, by A. Allen, S.J. Schwartz, C. Harvey, C. Perry, C. Huc and P. Robert, DS-QMW-TN-0010 Issue 2, Rev. 0.1, (CEF-2) dated 25 May 2004, pdf file available from http://www.space-plasma.qmul.ac.uk/csds/welcome.html

4. Space Physics Archive Search and Extract, SPASE Conceptual Model, draft dated 26 June 2003, publicly available from http://www.igpp.ucla.edu/spase/data/model.pdf

5. Space Physics Archive Search and Extract, SPASE Conceptual Model, Draft 10, dated 06 February 2004 (the latest version), available (password protected) from http://www.igpp.ucla.edu/spase/data/candidate/index.cfm

6. Space Physics Archive Search and Extract, "Hierarchical SPASE Data Model", proposals from D.A. Roberts dated 1 June 2004 for modifications to the SPASE Conceptual Model, Draft 10.

7. A CDF dataset using ISTP/IACG guidelines, by definition forms a logically complete and self-sucient whole (data and descriptions) http://spdf.gsfc.nasa.gov/istp guide/istp guide.html

8. Unified Model for Solar Metadata, by Kevin Reardon et al., EGSO-WP4-D1-20030930, dated September 30, 2003, file EGSO.data.model.v1.4.pdf . Metadata description for the European Grid of Solar Observations.

9. Cluster Active Archive: Product Naming Convention, by C.H. Perry, CAA-ESTTN-0001, Issue 1.0, dated 2004-07-26.

10. Report of the Parameter Naming Task Group, DS-RAL-TN-0002,

11. Cluster Data Delivery Interface Document (DDID), CL-ESC-ID-2001, Issue 3.0, dated 19 May 2000.

12. Data Elements and Interchange Formats – Information Interchange – Representation of Dates and Times. International Standard ISO 8601:2000, 2nd ed. Geneva: ISO, 2000. Also available at http://www.ccsds.org/documents/301x0b3.pdf[1]

[1] ref does not work - New version released, ref. http://public.ccsds.org/publications/archive/301x0b4.pdf;

A Definitions of CAA Concept Keywords

The metadata are organised according to the hierarchy of levels described in Section 4. Tables 3 through 14 show respectively metadata corresponding to

Table	Concept Keywords for
3	Mission
4	Observatory
5	Experiment
6	Instrument
7	Dataset
8	Parameter : definition
9	Parameter : units
10	Parameter : coordinates
12	Parameter : field format
11	Parameter : record format
13	Parameter : plot information
14	File

In each table, successive columns show :

Header	Content
Metadatum	Name of the metadata concept keyword. Each keyword is unique, and is a logical name to identify the concept to which it refers
Occurrence	Whether use of the keyword is compulsory or optional, with values which are unique or multi-valued, as shown in Section 7.12
Source	Indicates whether the information is supplied by the Principal Investigator team (possibly with the help of CAA), or by CAA directly
Description	A short description of the concept, and the type selected from : • Numerical, a numerical value, • Logical, a logical name or pointer, • String(~n), a free text string of up to about n characters, • Formatted, a formatted text string, • Personnel reference, formatted Name>role>e-mail address • ISO time code, which has format YYYY-MM-DDThh:mm:ss • ISO time range, specified by a pair of ISO time codes, or • Enumerated, an enumerated text string, In the last case the list of CAA enumerated values is provided.
Ref.	Other parts of this document which provide further information.

Note: the lists of acceptable values for Enumerated keywords are given in Appendix B

Table 3: Mission Level Concept Keywords

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
MISSION	1..1	CAA	Enumerated	Mission or project under which the data was collected	
PARENT_MISSION	1..1	CAA	Enumerated	Mission or project from which the data was collected: (*not required for CEF*)	Sect. 4
MISSION_TIME_SPAN	1..1	CAA	ISO_TIME_RANGE	This specifies the time interval covered by the mission	
MISSION_AGENCY	1..1	CAA	Enumerated	The agency responsible for the mission	
MISSION_DESCRIPTION	1..1	CAA	String (~500)	A short text describing the mission, to be read by users when browsing the site of the archive	
MISSION_KEY_PERSONNEL	1..N	CAA		Personnel references to key mission personnel, past and present	Sect. 7.10
MISSION_REFERENCES	1..N	CAA	Logical	A pointer to a list of standard reference documents, including the URL of the Project home page	Sect. 7.11
MISSION_REGION	1..N	CAA	Enumerated	The regions of space which the mission studied.	
MISSION_CAVEATS	0..N	CAA	String (~320)	Miscellaneous information concerning the mission. May include reference(s) to external file(s)	

Table 4: Observatory Level Concept Keywords.

An observatory is a platform or facility which houses an instrument: for Cluster there are four platforms and two facilities.

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
OBSERVATORY	1..1	CAA	Enumerated	The platform on which the hardware flew (or the place from where the data originated). Multiple is used for compound parameters (Sect. 7.4.1), it must be followed by the enumerated values corresponding to each of the observatories used.	
PARENT_OBSERVATORY	0..1	CAA	Enumerated	the platform on which the hardware flies (*not required for CEF*)	Sect. 4
OBSERVATORY_DESCRIPTION	1..1	CAA	String (~500)	listing particularities with respect to the mission description, for example, orbital or data peculiarities, failures, etc.	
OBSERVATORY_TIME_SPAN	1..1	CAA	ISO_TIME_RANGE	if different from MISSION_TIME_SPAN	
OBSERVATORY_REGION	1..N	CAA	Enumerated		
OBSERVATORY_CAVEATS	0..N	CAA	String (~320)	of miscellaneous information concerning the observatory. May include reference(s) to external file(s)	

Table 5: Experiment Level Concept Keywords.

An experiment is identified by its Principal Investigator. Cluster has 11 hardware experiments, each of which provide an instrument for each of the four spacecraft

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
EXPERIMENT	1..1	PI	Enumerated	The name of the experiment	Table 1
PARENT_EXPERIMENT	0..1		Enumerated	the experiment of which this instrument is part (*not required for CEF*)	Sect. 4
EXPERIMENT_DESCRIPTION	1..1	PI	String (~500)	principles of the measurement and description of the sensor, suitable for displaying via a Web browser	
EXPERIMENT_REFERENCES	1..N	PI	Logical	list of reference documents. When pointers are used, the format is : *DATASET_ID for example : *CL_CQ_WHI	Sect. 7.11
INVESTIGATOR_COORDINATES	1..1	PI		Personnel reference of the Principal Investigator	Sect. 7.10
EXPERIMENT_KEY_PERSONNEL	1..N	PI		Personnel references of instrument key personnel (Experiment Engineer, Deputy-PI, Archive Scientist, etc.)	Sect. 7.10
EXPERIMENT_CAVEATS	0..N	PI	String(~320)	of miscellaneous information concerning the experiment. May include reference(s) to external file(s)	

Table 6: Instrument Level Concept Keywords.

The Cluster mission has 44 hardware instruments, identified by Observatory and Experiment.

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
INSTRUMENT_NAME	1..1		*Enumerated*	the instrument used to collect the data	Table 1
PARENT_INSTRUMENT	1..1	PI	Enumerated	the instrument used to collect the data (*not required for CEF*)	Sect. 4,
INSTRUMENT_DESCRIPTION	1..1	PI	String (~500)	describing particularities (if any) of this instrument with respect to EXPERIMENT_DESCRIPTION of Table 4	
INSTRUMENT_TYPE	1..1		Enumerated	the type of experiment, one or more	Sect. 7.10
MEASUREMENT_TYPE	1..N	CAA	Enumerated	the type of measurement, one or more (A general indication of what is observed, often all that is available for older \legacy" data systems)	Sect. 7.10
INSTRUMENT_CAVEATS	0..N	PI	String (~320)	of miscellaneous information concerning the instrument. May include reference(s) to external file(s)	Sect. 7.10

Table 7: Dataset Level Concept Keywords

These keywords refer to metadata pertaining to all the physical parameters included in one dataset. In general, a dataset will contain many different physical parameters.

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
DATASET_ID	1..1	CAA	Enumerated	unique identifier of the dataset.	Sect. 7.5
PARENT_DATASET_ID	1..1	CAA	Logical	identifier of the dataset of which this parameter is part Format : *DATASET_ID (*not required for CEF*)	Sect. 4
DATASET_TITLE	1..1	PI/CAA	String (~80)	a concise scientific name for the dataset, for example, "WHISPER high resolution data"	
DATA_TYPE	1..1	CAA	Formatted	description of the nature of the product of the form ID>Short-name, where allowed values of ID and the corresponding. Short names are Enumerated in Table 2	Sect. 5
DATASET_DESCRIPTION	1..1	PI	String (~1024)	a precise description of the dataset, suitable for displaying on a Web browser	
CONTACT_COORDINATES	1..1	PI		Personnel reference of the scientific point of contact for this dataset	
TIME_RESOLUTION	1..1*	PI	Numerical	the characteristic time interval (in seconds) between data samples	Sect. 7.10
MIN_TIME_RESOLUTION	1..1*	PI	Numerical	the maximum time interval (in seconds) between data samples	Sect. 7.10
MAX_TIME_RESOLUTION	1..1*	PI	Numerical	the minimum time interval (in seconds) between data samples	Sect. 7.10
PROCESSING_LEVEL	1..1	PI	Enumerated	the level to which the data has been processed	Sect. 7.10
ACKNOWLEDGEMENT	1..1	PI	String (~500)	for example, "Please acknowledge the instrument team and ESA Cluster Active Archive in any publication based upon use of this data"	Sect. 7.10
DATASET_CAVEATS	0..1	PI	String (~320)	of miscellaneous information concerning the dataset. May include reference(s) to external file(s)	

*Cardinality is optional for caveat data sets (Data Product Type "CQ" of Table 2, page 7)

Table 8: Parameter Level Keywords – description.
** Usage of the keyword for Support_Data is optional in the parameter metadata.*

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
PARAMETER_ID	1..1	CAA	Logical	identifier assigned to this parameter. This has the form "ParameterID_DatasetID"	Sect. 7.5
PARAMETER_TYPE	1..1	CAA	Enumerated	type of parameter	Sect. 7.10
CATDESC	1..1	PI	String (~80)	containing concise description of the parameter, e.g. "Ion differential flux at 12 energies in the range 67-1361 keV". The observatory ID is not required, it will be evident from the context.	
ENTITY	1..1*	PI	Enumerated	The type of entity whose property is measured	Sect 7.3
PROPERTY	1..1*	PI	Enumerated	The property of the entity which is measured	Sect 7.3
FLUCTUATIONS	0..1	PI	Enumerated	This concept keyword is present if the parameter is some measure of the temporal fluctuations of the observed property	Sect. 7.3
ERROR_PLUS and ERROR_MINUS	0..1	PI	Numerical	values of the probable error of the measured value of the parameter. Two values (PLUS and MINUS) allow partial description of a possible skew of the error distribution. May be a local attribute.	Sect. 7.10
COMPOUND	0..1	PI	Enumerated	Indicates a parameter which is not measured directly, but derived from two or more other observables.	Sect. 7.3
COMPOUND_DEF	0..1	PI		A precise definition, such as "The measured EFW despun electric field minus the $\mathbf{v}\times\mathbf{B}$ field induced by the spacecraft velocity $\mathbf{v}$"	Sect. 7.4

Table 9: Parameter Level Keywords - units

** Usage of the keyword for Support_Data is optional in the parameter metadata.*

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
UNITS	1.1*	PI	String (~20)	Units of the parameter, as to be indicated on the axes of a plot, for example : "nT", "ms", "keV", ... If unitless, this must be specified by UNITS="unitless". See section 7.6 for a vector expressed in polar coordinates	
SI_CONVERSION	1..1*	PI	Formatted Enumerated	The multiplicative factor required to take the archived value to the corresponding value in a standard SI unit “x”, expressed in the form "1.0E-5> x", where x is the appropriate SI unit. The following three are useful, but not SI radian, degree angle unitless [no units] The format must be scrupulously respected because the information will be parsed. Examples: Magnetic field in γ : SI_CONVERSION="1.0E-5>T" density in cm-3 : SI_CONVERSION="1.0E-6>(particles) m^-3" unitless : SI_CONVERSION="1>unitless" See section 7.6.2 for a vector expressed in polar coordinates	Sect. 7.10

Table 10: Parameter Level Keywords - coordinates

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
TENSOR_ORDER	0..1	PI	Enumerated	The order of the vector or tensor which represents a non-scalar physical observable.	Sect. 7.6
TENSOR_FRAME	0..1	PI	Enumerated		Sect 7.7
TARGET_SYSTEM	0..1	PI	Enumerated		
COORDINATE_SYSTEM	?..1	PI	Enumerated	For vectors, tensors or components thereof, an acronym indicating the coordinate system.	Sect. 7.7 Sect. 7.6
FRAME_ORIGIN	0..1	PI	Enumerated	The location of the origin of coordinates if different from the that implied by the name of the coordinate system. Presently only one such exception has been identified for Cluster.	Sect 7.7
FRAME VELOCITY	0..1	PI	Enumerated	describes the motion of the origin of the coordinate system	Sect 7.7
FRAME	?..1	CAA		Identifies the nature of the variable (scalar, vector, ...), and the frame of reference if applicable. Required for backward compatibility, see Ref. 2, page 57 for details	Sect. 7.10
REPRESENTATION	?..1	PI	Formatted	For vectors and tensors	Sect 7.6

Table 11: Parameter Level Keywords - record format

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
SIZES	?..N	PI	Formatted	Dimensions of the array required for any physical parameter represented by more than one component (vectors, tensors, spectral arrays, etc.). Examples : SIZES=1 for a scalar (default value) SIZES=3 for a vector of 3 values SIZES=5,5 for an array of 5×5 = 25 values	
DEPEND_0	0..1	PI		Explicitly ties the data variable to the independent variable parameter(s) on which it depends.	Sect. 7.9
DEPEND_i	0..1	PI		For arrays, defines dependencies other than those due to the vector or tensor nature of the physical parameter. Examples : energy channel, frequency channel, or caveats which vary within the dataset For vectors and tensors, see REPRESENTATION	Sect. 7.9
DATA	0..N	PI	Formatted	Used to define values of non-record-varying support data	
LABEL_i	0..1	PI	String (~40)	used to label an i-dimensional variable when one value of LABLAXIS is not sufficient to describe the variables or to label all the axes. Examples : "Bx","By","Bz"	

Table 12: Parameter Level Keywords - field values and quality

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
DELTA_PLUS and DELTA_MINUS	0..1	PI	Numerical	Number of units (of the parameter) to add to (or subtract from) the nominal value to obtain the upper (or lower) limit of the parameter interval within which the data was acquired. May be local attribute.	Sect. 7.10
VALUE_TYPE	1..1	PI	Enumerated	Identification of the value type (essential for ASCII conversion)	Sect. 7.10,
SIGNIFICANT_DIGITS	1..1*	PI	Integer	the number of decimal digits required to preserve the precision of the parameter (essential for ASCII conversion): 1 2 3 any integer	Sect. 7.10,
FILLVAL	1..1	PI	Numerical	The fill value used to replace bad or missing data. The type of fill value depends upon the VALUE_TYPE: for example, for an ISO TIME it could be 9999-12-31T23:59:59	
QUALITY	1..1*	PI	Integer	indicates the quality of the parameter, using the following definitions: 0 Not applicable 1 Major problems, check caveats 2. Minor problems, check caveats 3. Good data 4. Excellent data, has received special treatment Notes: • the quality flag can also be written record by record using a pointer quality table • most data are flagged with quality = 1-3. Values are primarily based on automatic procedures, so data flagged with quality=1-2 can sometimes be of good quality while in some other instances data of quality=3 can be of poor quality. • If you have any doubt of quality it is recommended to contact the data provider for further clarifications, see names in CONTACT_COORDINATES.	
PARAMETER_CAVEATS	0..1	PI	String (~320)	of miscellaneous information concerning the parameter. May include reference(s) to external file(s)	

Table 13: Parameter Level Keywords - graphical recommendations

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
FIELDNAM	0..1	PI	String (~40)	describing the parameter and which is used, for example, to label plots (use LABLAXIS for the axis). Examples : ion energy magnetic field etc. ... This could be the same as CATDESC. Note, however, that while FIELDNAM is optional, CATDESC (Table 8) is not!	
LABLAXIS	0..1	PI	String (~40)	which can be used to label the y-axis of a plot or to provide a column heading for a data listing. Examples : "Ne" "Te"	
SCALEMIN	0..1	PI	Numerical	Minimum data value to be plotted. Useful, for example, for plotting multiple files at the same scale. By default, the actual minimum data value found in the file is used	
SCALEMAX	0..1	PI	Numerical	Maximum data value to be plotted. Useful, for example, for plotting multiple files at the same scale. By default, the actual maximum data value found in the file is be used	
SCALETYP	0..1	PI	Enumerated	Indicates whether the variable should have a linear or a log scale. By default a linear scale is assumed.	
DISPLAYTYPE	0..1	PI	Enumerated	Recommended type of plot for graphical display.	

Table 14: File Level Keywords

Each dataset contains many files of identical format and content, but different numerical values. Each file has its particularities, as listed.

Keyword	Occurrence	Source	Type of attribute	Description of contents	Ref
LOGICAL_FILE_ID	1..1	PI	Logical	name of the file. This is the complete extended form "DatasetID_Instance", without the extension.	Sect. 7.5
PARENT_DATASET (not required for CEF)	1..1	PI	Enumerated	identifier of the parent dataset (Table 7)	Sect. 4
VERSION_NUMBER	1..1	PI	Integer	version of this file	
DATASET_VERSION	1..1	PI	Text	indicating the version of the dataset to which it corresponds (for instrument teams to track their own versions)	
FILE_TYPE	1..1	PI	Enumerated	Type of file format, specified by its extension	
FILE_TIME_SPAN	1..1	PI	ISO time range	indicating the date and time of the beginning and the end of the data in the file	
GENERATION_DATE	1..1	PI	ISO time code	when the file was produced.	
INGESTION_DATE	1..1	CAA	ISO time code	when the file was ingested into CAA	
FILE_SIZE	1..1	CAA	Numerical	size of the file (in bytes), supplied by CAA at ingestion	
METADATA_TYPE	0..1	PI	Enumerated	The Metadata Dictionary used to describe the data. This will usually be CAA (this document), but occasionally CSDS	
METADATA_VERSION	0..1	PI	Text	the version number of the above metadata description used	
FILE_CAVEATS	0..1	PI	String(~320)	of miscellaneous information concerning the data in the file. May include reference(s) to external file(s)	

B Glossary of all Enumerated Keywords

Keyword	Definition	
MISSION	the mission or project under which the data was collected	
	Cluster	
	DoubleStar	
PARENT_MISSION	not required for CEF	
	Cluster	Mission or project from which the data was collected:
MISSION_AGENCY	The agency responsible for the mission	
	ESA	European Space Agency
	CNSA_and_ESA	
MISSION_REGION	The regions of space which the mission studied	
	Solar_Wind	
	Bow_Shock	
	Magnetosheath	
	Magnetopause	
	Magnetosphere	
	Magnetotail	
	Polar_Cap	
	Auroral_Region	
	Cusp	
	Radiation_Belt	
	Plasmasphere	
	Ionosphere	
	Thermosphere	
	Atmosphere	
OBSERVATORY	The platform on which the hardware flew, or the place from where the data originated	
	Cluster-1	spacecraft also known as rumba
	Cluster-2	idem salsa
	Cluster-3	idem samba
	Cluster-4	idem tango
	ESOC	European Space Operations Centre, Darmstadt, Germany
	JSOC	Cluster Joint Operations Centre, RAL, Chilton, UK
	Multiple	
	DoubleStar-1	
	DoubleStar-2	
PARENT_OBSERVATORY		
	Same as for OBSERVATORY	
OBSERVATORY_REGION		
	as for MISSION REGION	

INSTRUMENT_TYPE	the type of experiment, one or more of :
Antenna	
Auxiliary	
Channeltron	
Data_Processing_Unit	A unit which processes data aboard the spacecraft so as to reduce the telemetry downlink data rate
Double_Sphere	
Electron_Drift	
Electrostatic_Analyser	
Faraday_Cup	
Flux_Feedback	
HF_Radar	
Langmuir_Probe	
Waveform_Receiver	
Long_Wire	
Magnetometer	
Mass_Spectrometer	
Micro-channel_Plate	
Monopole	
Particle_Correlator	An instrument to auto-or cross-correlate time variations of a particle count rate
Quadrispherical_Analyser	
Resonance_Sounder	
Search_Coil	
Spacecraft_Potential_Control	
Spectral_Power_Receiver	
Solid_State_Detector	
Waveform_Receiver	
EXPERIMENT	The acronym for the experiment, which is one of
For Cluster	
ASPOC	Active Spacecraft POtential Control
CIS	Cluster Ion Spectrometer
DWP	Digital Wave Processor
EDI	Electron Drift Instrument
EFW	Electric Fields and Waves
FGM	Fluxgate Magnetometer
PEACE	Plasma Electron And Current Experiment
RAPID	Research with Adaptive Particle Image Detectors
STAFF	Spatio-Temporal Analysis of Field Fluctuations
WBD	Wide Band Data
WHISPER	Waves of HIgh frequency and Sounder for Probing the Electron density by Relaxation
AUX	Auxiliary Parameters
Multiple	Multiple is used for compound parameters; see Sect. 7.4.1
For Double Star	
HEED	
HEPD	
HIA	
HID	
LEFW	
LEID	
NUADU	
STAFF-DWP	For the Double Star mission, the STAFF and DWP instruments are considered as a single instrument. The STAFF-DWP (experiment) and STAFF-DWP-D1 (instrument) tags are recommended for all datasets from originating from the DWP or STAFF on Double Star TC1.
PARENT_EXPERIMENT	
Same as for EXPERIMENT	

INSTRUMENT_NAME	the instrument used to collect the data : the instrument acronym plus spacecraft number
	ASPOC1 ASPOC2 ASPOC3 ASPOC4 CIS1 CIS2 CIS3 CIS4 CIS-CODIF1 CIS-CODIF2 CIS-CODIF3 CIS-CODIF4 CIS-HIA1 CIS-HIA2 CIS-HIA3 CIS-HIA4 DWP1 DWP2 DWP3 DWP4 EDI1 EDI2 E DI3 EDI4 EFW1 EFW2 EFW3 EFW4 FGM1 FGM2 FGM3 FGM4 PEACE1 PEACE2 PEACE3 PEACE4 RAPID1 RAPID2 RAPID3 RAPID4 STAFF1 STAFF2 STAFF3 STAFF4 STAFF-SC1 STAFF-SC2 STAFF-SC3 STAFF-SC4 STAFF-SA1 STAFF-SA2 STAFF-SA3 STAFF-SA4 WBD1 WBD2 WBD3 WBD4 WHISPER1 WHISPER2 WHISPER3 WHISPER4 AUX1 AUX2 AUX3 AUX4 AUXC ~~ISE (*Identified Science Events*)~~ ~~PCY PGP PMP PSE~~ Multiple (Multiple is used for compound parameters, see Sect. 7.4.1) *Double Star:* ASPOC-D1 DWP-D1 FGM-D1 FGM-D2 HEED-D1 HEED-D2 HEPD-D1 HEPD-D2 HIA-D1 HID-D1 HID-D2 PEACE-D1 PEACE-D2 STAFF-D1 STAFF-DWP-D1 LEFW-D2 LEID-D2 NUADU-D2 AUX-D1 AUX-D2
PARENT_INSTRUMENT	
	Same as for INSTRUMENT

MEASUREMENT_TYPE	the type of measurement, one or more of the following extracted from the SPASE Measurement Type Definitions in the document SPASE Data Model Draft 3.doc dated 21 July 2004 :
Activity_Index	An indication, derived from one or more measurements, of the level of activity of an object or region, such as sunspot number, F10.7 flux, Dst, or the Polar Cap Indices.
Electric_Field	Measurements of electric field vectors (sometimes not all components) as time series.
Electron_Drift	An active experiment to measure the electron drift velocity based on sensing the displacement of a weak beam of electrons after one gyration in the ambient magnetic field.
Emitted_Current	The current emitted by an electron gun or a device to control the spacecraft potential.
Energetic_Particles	Measurements of fluxes particles at above thermal energies, including relativistic particles of solar and galactic origin. May give simple fluxes, but more complete distributions are some times possible. Composition measurements may also be made.
Instrument_Status	
Ion_Composition	In situ measurements of the relative flux or density of (usually atomic) constituents of the space environment. May give simple fluxes, but full distribution functions are sometimes measured.
Magnetic_Field	Measurements of magnetic field vectors (sometimes not all components) as time series; can be space-or ground-based
Neutral_Atom_Images	Measurements of the quantity of ions of various particle species using the charge exchange of the ions with neutrals that then move freely from the source region to the detector.
Neutral_Gas	Measurements of neutral atomic and molecular components of the Earth and space environments.
Particle_Correlator	A particle instrument which correlates particle flux to help identify wave phenomena
Radio_and_Plasma_Waves	Measurements of low-frequency electromagnetic waves in spectral bands; can be given by peak values or spectral densities, and may give polarization information.
Radio_Soundings	Measurements of position, density and/or velocity of ionized constituents of the space environment using the active probing of the plasma by radio waves.
Spacecraft_Potential_Control	An instrument to control the electric potential of a spacecraft with respect to the ambient plasma by emitting a variable cur rent of positive ions.
Spacecraft_Status	
Status	Ephemeris, attitude, housekeeping, engineering, or other properties of an observatory or spacecraft relevant to the functioning of the scientific measuring devices present.
Thermal_Plasma	Measurements of the main thermal distributions of particles, typically electrons, protons, and/or alpha particles. Generally used to determine the main bulk properties of the plasma by moments or fits to distribution functions.

DATASET_ID	a unique identifier of the dataset, provided by the CAA Project as described in Sect. 7.5 under "extended form" (page 13)	
DATA_TYPE	a description of the nature of the product of the form ID>Short-name, where allowed values of ID and the corresponding Short name are enumerated in Table 2, Sect. 5	
PROCESSING_LEVEL	the level to which the data has been processed	
	Raw	Data which has decommuted but not otherwise modified.
	Uncalibrated	Data which has been pre-processed ready for calibration, but upon which no irreversible operation has been performed (no telemetry information lost).
	Calibrated	Data which is in known physical units described by the parameter SI CONVERSION. Particle flux rates (distribution functions) are in this category.
	Derived	Parameters derived from the calibrated data, such as the plasma moments derived by integrating over the particle distribution function. Parameters derived using some physical model to interpret the data, for example the plasma density derived from either wave spectra or a Langmuir probe, are also in this category.
	Auxiliary	Additional data which is not really part of the science dataset, but which is included because it is useful for the scientific analysis.
PARAMETER_TYPE	the type of parameter :	
	Data	A scientific parameter.
	Support_Data	In formation needed to interpret the scientific parameter.
ENTITY	The type of entity whose property is measured :	
	Aerosol	
	Alpha	
	Compound	data from more than one experiment, see Sect. 7.4)
	Dust	
	Electric_Field	
	Electron	
	Helium+	
	Instrument	
	Ion_CNO	
	Ion	
	Magnetic_Field	
	Molecule	
	Neutral	
	Observatory	
	Oxygen+	
	Other1, Other2, ..	*these two are used for the Auxiliary Data*
	Particles	
	Photon	
	Proton	

PROPERTY	The property of the entity which is measured :	
	"Particle-type"	
	Charge_Density	
	Coordinate_rotation	
	Corrected_Particle_Count_Rate	The Raw Particle Count Rate corrected for the detector efficiency.
	Current	
	Differential_Energy_Flux	
	Differential_Particle_Flux	The number of particles, per unit area, per unit time, per unit solid angle, per unit of energy
	Emitted_current	
	Energy	
	Heat_Flux	
	Integral_Particle_Flux	The number of particles, per unit area, per unit time, per unit solid angle, independent of their energy. This is the differential particle flux integrated over the complete energy range of the detector.
	Mass_Density	
	Mass_Flux	
	Number_Density	
	Particle_Energy_Flux	This is the differential particle flux multiplied by the particle energy, and integrated over the detector energy range. It is typically expressed in units keV cm^{-2} s^{-1} sr^{-1} keV^{-1} which may seem somewhat strange (keV ... keV^{-1})
	Phase_Space_Density	
	Pressure	
	Pressure_Tensor	
	Raw_Particle_Count_Rate	The count rate expressed in particles per second.
	Raw_Particle_Counts	The number of particles counted during an interval long enough to provide a statistically significant result, whose duration is given by one of the DEPEND_i variables
	Speed	
	Status	Used with ENTITY="Spacecraft" or ENTITY="Instrument"
	Temperature	
	Time_Offset	
	Time-of-flight	
	Vector_Mass_Flux	
	Velocity	
	"Field-type"	
	Component	
	Direction	
	Magnitude	
	Potential	
	Probe_Potential	
	Vector	
	"Optical"	
	Photon_Flux	

FLUCTUATIONS	This concept keyword is present if the parameter represents some measure of the temporal fluctuations of the observed property. These parameters are :	
	Waveform (default value)	it is the parameter itself which is provided.
	Bispectrum	
	Correlation	
	Covariance	
	Cross Correlation	
	Fluctuation_Level	For a signal x which has been averaged, this is the rms value $\sqrt{<(x-\bar{x})^2>}$ of the fluctuations1 with respect to the mean value $\bar{x} =< x >$ during the averaging interval. It is an indicator of how much "information" has been lost during averaging, i.e., the amplitude of the fluctuations resolved by the instrument but not transmitted to the telemetry (see Sect. 7.10, page 24).
	Fourier_Cross_power-spectrum	
	Fourier_Cross-spectrum	
	Fourier_Power-spectrum	
	Fourier-spectrum	
	Mean_Square_Level	The mean power spectral density over a wide range of frequencies obtained, for example, by averaging the power spectral density over a number of frequency channels, or by bandpass pre-detector filtering. Differs from the Fluctuation Level in that the spectral bandpass characteristic of the fluctuations is better known.
	Polarisation	
	Poynting_vector	
	Stokes_Parameters	
	Wavelet_Cross-power-spectrum	
	Wavelet_Cross-spectrum	
	Wavelet_Power-spectrum	
	Wavelet-spectrum	
COMPOUND	This concept indicates that a parameter is not measured directly, but is derived from two or more other observables parameters. So far the following have been identified :	
	Plasma beta	
	Alfven velocity	
TENSOR_ORDER	The order of the vector or tensor which represents a non- scalar physical observable	
	0	a scalar quantity (the default value)
	1	vector (i.e., a tensor of order 1)
	2	tensor of order 2
	3	tensor of order 3
COORDINATE_SYSTEM	For vectors, tensors or components thereof, an acronym indicating the coordinate system. Allowed acronyms are in the left column, with those recommended for Cluster in bold :	
	FAC	Field-Aligned Coordinates
	GEI	Geocentric Equatorial Inertial
	GEOC	Geographic (geocentric)
	GSE	Geocentric Solar Ecliptic
	GSEQ	Geocentric Solar Equatorial
	GSM	Geocentric Solar Magnetic
	HAE	Heliospheric Aries Ecliptic
	HEE	Heliospheric Earth Ecliptic
	HEEQ	Heliospheric Earth Equatorial
	ISR2	Inverted (about x-axis) SR2
	MAGD	Geomagnetic (dipole)
	MFA	Magnetic Field-Aligned
	SC	SpaceCraft (body-build) coordinates
	SM	Solar Magnetic
	SR2	Despun spacecraft (spin-reference) frame
	Instrument	Instrument specific
	Data	Data-derived

SI_CONVERSION	This is the multiplicative factor required to take the archived value to the corresponding value in a standard SI unit “x”, expressed in the form "1.0E-5>x", where x is the appropriate SI unit.
m	metre
N	newton
kg	kilogram
Pa	pascal
s	second
Hz	hertz
A	ampere
V	volt
K	kelvin
W	watt
rad	radian
J	joule
sr	steradian
C	coulomb
T	tesla
ohm	ohm
mho	mho (siemens)
H	henry
F	farad
FRAME_ORIGIN	The location of the origin of coordinates if different from . the that implied by the name of the coordinate system
Observatory	
FRAME_VELOCITY	A string which describes the motion of the origin of the coordinate system.
Observatory	
Inertial	
Earth_Corotating	
REPRESENTATION	For vectors and tensors, the representation as described in Sect. 7.6
VALUE_TYPE	Identification of the value type (essential for ASCII conversion) .
CHAR	
DOUBLE	
FLOAT	
INT	
ISO_TIME	
ISO_TIME_RANGE	
QUALITY	An integer indicating the quality of the parameter.
0	*Not applicable or scientifically unusable data (e.g. due to incorrect instrument settings)*
1	*Data contain errors very likely (if used, discussion with the data provider is recommended)*
2	*Data can have systematic errors, e.g. offsets (if used, discussion with the data provider is recommended)*
3	*Good for publication*
4	*Good for publication; data have been manually calibrated/confirmed or have received special treatment*
SCALETYP	Indicates whether the variable should have a linear or a logarithmic scale.
Linear	default
Log	
DISPLAYTYPE	The recommended type of plot for graphical display.
Time_Series	A normal 2-D line plot, with the DEPEND_0 variable (which is normally time) plotted along the abscissa and the physical parameter plotted along the ordinate.
Spectrogram	A 3-D plot (spectrogram), in which the DEPEND_0 variable is plotted along the abscissa, the DEPEND_1 variable is plotted along the ordinate, and the physical parameter is represented by variations of colour, or shades of grey.
Stack_Plot	An ensemble of time series plots, for which each 2-D line plot has been offset in the ordinate direction by an amount which depends upon the DEPEND_1 variable. Successive line plots appeared to be “stacked” one above the other.

FILE_TYPE	The type of file format, specified by its extension.	
	cdf	Common Data Format
	cef	Cluster Exchange Format
	txt	An ASCII text file
	ps	PostScript
	png	
	gif	Graphical Interface Format
	jpg	Abbreviation for jpeg
	jpeg	Joint Photographic Experts Group format
	pdf	Portable Document Format
	tex	A TEX or LATEX document source file
	doc	Microsoft Word document file
	tiff	
METADATA_TYPE	The Metadata Dictionary used to describe the data.	
	CAA	Cluster Active Archive
	CSDS	Cluster Science Data System metadata dictionary may be used for the CSDS products
TENSOR_FRAME		
	Instrument	
	MFA	
TARGET_SYSTEM		
	Other	
	TENSOR_FRAME	*can also take any value supported by TENSOR_FRAME*

C Scientific Events

It is a characteristic of collisionless plasma that energy dissipation is confined to limited regions of space, such as shocks, discontinuities, and regions of magnetic field re-connection, where strong non-linear wave turbulence gives rise to energy dissipation. The macroscopic plasma parameters, such as density, temperature, and magnetic field, change abruptly in these regions; elsewhere the plasma is statistically relatively homogeneous.

The plasma physicist is particularly interested in studying these regions of strong turbulence which determine what happens in large regions of space elsewhere. As it advances along its orbit a spacecraft encounters these frontiers from time to time; but as the magnetosphere is highly variable, these crossings are somewhat unpredictable. Therefore it is highly desirable to use the scientific data to identify these regions of strong scientific interest. But the identifications are rather difficult, and somewhat controversial: what parameters should be used to identify a shock or discontinuity, where does a shock "start", and where does it "stop" ? Nevertheless, any information, even incomplete, is of great scientific value although, clearly, it must be documented with great care.

An observed scientific event is something which is characterised by a start time t1 and a stop time t2. For asymmetrical events may be helpful to use the order of t1 and t2 to indicate direction: for example, magnetopause or shock crossings could have t1 <t2 for outward crossings, and t1 >t2 for inward crossings.

The following table lists the differences between status data and scientific event data requirements:

	Status Type	Scientific Event Type
1	t1 < t2	No ordering of t1 and t2
2	Records are time ordered	Records are not time-ordered; data will be produced irregularly according to the scientific priorities of the moment
3	The data is quasi-irrefutable (models excepted)	Depending upon its nature, the data may be subject to controversy
4	One or more conditions are satisfied continuously during each time interval	There is no concept of "validity" within the interval; the two times merely "delimit" some "event" seen on one spacecraft. Two-spacecraft use of event tables has also been proposed.
5	Level 1 or Level 2 data	Certainly Level 3 data
6	The data is quasi-continuous during the whole mission	The datasets will generally be orders of magnitude smaller than for status data
7	Each files is identified by the event table to which its belongs, and its start time (as for numerical data)	File nomenclature and identification still need to be resolved

There are both scientific and technical advantages of archiving status and scientific event data in the same type of archival storage entity (see Sect. 8.2). This table suggests that this is possible in tables with t1 <t2 provided that additional fields supply the following information:

1. The nature of the event ([perpendicular/parallel] shock, magnetopause, etc.) if different types of event are included in the same event table.
2. The scientist responsible for identifying the event, if several scientists contribute to the same event table.
3. The dataset(s) and/or method used to identify the event (if not specified in the global metadata).
4. The time, t1 or t2, which corresponds to the "inner" frontier of the event. This field could even be expressed in terms of the (normal component of the) shock velocity relative to the spacecraft (provided it can be determined); then the user (and/or his application) has all the information required to perform a Lorentz transformation on each time profile and produce, for example, a stack plot of the spatial profiles of successive shock crossings (the time offsets being provided by averaging t1 and t2).

Clearly thought needs to be given to the organisation of event table sets; for example, they could be organised by:

- by type of event,
- by institute or scientist who performs the identifications,
- by Cluster workshop,
- by the intersection (many simple tables) or union (fewer, more complicated, tables) of these and/or other criteria.