THE ESA CLUSTER ACTIVE ARCHIVE

C. Perry[(1)], T. Eriksson[(2)], P. Escoubet[(2)], S. Esson[(2)], H. Laakso[(2)], S. McCaffrey[(2)], T. Sanderson[(2)], H. Bowen[(2)], A. Allen[(3)] and C. Harvey[(4)]

[(1)] Rutherford Appleton Laboratory, UNITED KINGDOM
[(2)] ESTEC, NETHERLANDS
[(3)]Imperial College University of London, UNITED KINGDOM
[(4)]CESR, FRANCE

ABSTRACT

The Cluster Active Archive (CAA) is a database of high-resolution Cluster data and other allied products and services that is being established and maintained by ESA. The aims of the CAA are i) to ensure that the CAA contains all of the Cluster high resolution data ii) that the data archived should be of the best quality achievable, iii) that the data should be suitable for science use and publication by the scientific community, and finally that iv) the CAA will provide user friendly services for searching and accessing these data and ancillary products such as orbit information.

We present the current status of the CAA system, data products and services as it prepares to make the data from the first year of mission operations publicly available to the Cluster community.

1. INTRODUCTION

The Cluster mission, together with the Solar Heliospheric Observatory (SOHO), constitutes 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. Although 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. A further extension of four years to December 2009, with a mid-term review planned for 2007, was approved by the ESA SPC in February 2005.

Cluster is fulfilling its promise as a revolutionary magnetospheric space mission. 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 already acquired have validated the Cluster mission concept and its scientific objectives. The Cluster observations have shown that, on the scales so far explored, 3-dimensional plasma dynamics clearly plays an essential role in shaping the larger scale structures of our space environment. The second extension will further extend these observations by using multiscale spacecraft separations to probe the link between phenomena acting on different spatial scales and by probing new regions not explored earlier in the mission.

The mission is successfully delivering raw data to the Principal Investigator teams, and summary and prime parameter data through the Cluster Science Data System (CSDS). The Cluster Active Archive (CAA) aims to augment the mission by providing a facility 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. The scientific rationale underpinning the CAA activities are:

- To maximise the scientific return from the mission by making all Cluster data available to the worldwide scientific community.
- To ensure that the unique data set returned by the Cluster mission is preserved in a stable, long-term archive for scientific analysis beyond the end of the mission.
- To provide a major contribution by ESA and the Cluster science community to the International Living With a Star (ILWS) programme.

The CAA will help to remove some of the barriers that currently limit the efficient analysis of Cluster data by providing a system for uniform access to the high-quality and high-resolution instrument data together with the necessary data descriptions and auxiliary information to aid its interpretation. The CAA will facilitate analyses that up to now have been difficult due to data or information inaccessibility. Key to achieving this is the handling of relatively small (the total archive is expected to be only a few tens of Terabytes) but complex and diverse amounts of data.

In the short-term the provision of Cluster data in this uniform way will result in an improved and faster science return. It is also hoped that the CAA, in conjunction with other national and international initiatives, such as the international Space Physics

Proceedings Cluster and Double Star Symposium – 5th Anniversary of Cluster in Space, Noordwijk, The Netherlands, 19 - 23 September 2005 (ESA SP-598, January 2006)

Archive Search and Extract (SPASE) consortium, will encourage broader standardisation, thus helping to promote the re-use of analysis systems and long-term security of data.

1.1 Background

The nationally funded Cluster Science Data System (CSDS) was set-up to distribute quicklook and processed data to all Cluster Principal and Co-Investigators, as well as to the scientific community. To distribute the data efficiently to all users, a system of nine data centres, located in Austria, China, France, Germany, Hungary, Sweden, the United Kingdom, the United States and at ESTEC, was established. Each data centre, interconnected with the others, stores the full database of processed and validated data (low and medium resolution) from all instruments. Data from the start of the science phase of the mission, February 2001, onwards are available at these centres and can be accessed through the web.

CSDS was set-up to enable exchange of data from different instruments soon after acquisition and to allow the user to browse through data to identify interesting events. For that purpose, the data contained in the CSDS are of low and medium resolution. Once events have been identified, their detailed analysis usually requires the best quality, highest detail data. These high-resolution products are processed and validated at the PI institute. Due to limited manpower in the PI teams, high-resolution data have up to now been processed and validated on an event-by-event basis. To maximise the scientific return of the mission and to ensure the long-term security of the data it is necessary to make all Cluster high-resolution data available to the wider scientific community.

The CAA is intended as a database of high-resolution data and other allied products and services established and maintained under the overall control of ESA. In view of the shortage of manpower in most institutes processing Cluster data, ESA-supported manpower is deployed in institutes where the relevant expertise exists, to assist in the preparation, validation, and documentation of the high-resolution data to be deposited in the archive. In return, teams are expected to find a matching level of support for the archiving activities from their national funding bodies to support the archiving effort. It is also recognising that much of the in-house expertise might be lost at the end of the Cluster mission and it was therefore imperative that the archiving activities started during the operational phase of the mission.

This CAA development is part of ESA's contribution to ILWS. The ILWS programme was launched (kick-off meeting in Washington September 4-6 2002, and first public announcement at COSPAR in Houston, October 2002) as a long-term inter-agency collaboration enterprise in Solar and Solar-Terrestrial Physics, following the highly successful ISTP programme. The ILWS programme was initiated by the Inter Agency Consultative Group (IACG) and involves initially NASA, ISAS, RSA, CSA and ESA. The principal charter of ILWS is to stimulate, strengthen and coordinate space research to understand the governing processes of the Sun-Earth System as an integrated entity (for a preliminary ESA strategy of participation on ILWS see ESA-document SPC2002-39).

The CAA conceptual design phase started in 2003, followed by a development and implementation phase, comprising software integration and data preparation in 2004/5. Processing and preparation of data to be archived has proceeded in parallel within each of the experiment teams, with planning for data from all instruments to enter the database at an average rate of about two years of data per calendar year. Data from any individual experiment may be archived more or less rapidly, subject to the requirement that the archiving of all data should be completed at the conclusion of CAA phase. The archive will be accessible to all scientists. Once data are included in the archive they will be public. The active phase of the archive is expected to last for the duration of the mission and is a recognition that the data, calibrations and processing software are still evolving. However, it is important to make the data available early to maximise the science return from the mission. The CAA architecture has been developed so that it can respond quickly to changes and updates so that the data provided is of the best quality available at the time. The CAA is also intended to provide an extensible set of services to facilitate the active use of the data including advanced search, data visualisation and manipulation capabilities, support for science workshops, and interoperability with other related space physics archives (Allen 2003 and Perry 2003).

To achieve the stated aims, the requirements on the CAA can be summarised as:

- The CAA should contain all the Cluster high resolution data
- The data should be of the best achievable quality
- The data should be suitable for detailed science investigations
- Data should be publicly accessible
- Development should take place while all the necessary expertise still available

In addition the CAA will hold ancillary products and support information, including orbit and attitude data, survey plots and documentation. For the Cluster mission extension the CAA shall also take on the responsibility for the network distribution of newly acquired raw data to the Cluster instrument teams replacing the current delivery of these data on CD-ROM.

2. STANDARDS

Detailed science investigations using Cluster data inevitably require the examination, and in many cases combination, of data from the different instruments and spacecraft. Individual instrument teams have developed their high-resolution data systems independently resulting in a diversity of data product descriptions and formats that represents a significant technical impediment to the use of these data. A key issue for the CAA in terms of improving accessibility to the Cluster data is standardisation of products. One of the first activities of the CAA was to set up working groups to address the fundamental issues of data formats and metadata descriptions.

2.1 Data Format

The importance of a standard data format for the exchange of Cluster data had already been identified prior to the conception of the CAA. A single ascii format data file syntax was proposed by the CSDS Archive Task Group for the exchange of science data between instrument teams but without the necessary backing it had not been widely adopted. This format was intended as an exchange format to allow translation between the large number of native data formats used by science tools and data bases within the Cluster community. Adopting a single common format, that could be written and read by all teams, allows exchange of data between any of the other formats and storage systems. This format was adopted by the Cluster Active Archive design team as the basis for the format to be used for delivery of science data to the CAA, and as a format for delivery to science users.

The Cluster Exchange Format (version 2) file syntax and minimum header content (Allen et al., 2004) is specified to describe the products sufficiently for science use. The open format also allows for inclusion of the extra metadata required by the CAA to fully specify the contents and provenance of the data. This format serves to facilitate the storage of commonly used science products in a robust and easily accessible form. It is intended to assist delivery of data products to workers outside of the instrument team with different science and database software and to future scientists without access to specific software in use at the time of data archival.

2.2 Data Descriptions

The CAA Metadata Dictionary (Harvey et al., 2005) defines the terms that are to be used to describe the Cluster data down to the specification of individual parameters within the file. It is a superset of the terms required by the CEF-2 file format and builds on previous space physics metadata descriptions developed by the International Solar Terrestrial Physics Programme (ISTP), CSDS, SPASE and the Centre de Données de la Physique des Plasmas (CDPP). The metadata includes the following categories of information:

1. Essential scientific information to enable the scientist to identify the dataset(s) of interest and to understand the data once it has been recovered. This is the semantic description. It may also includes less essential, but exceedingly useful, practical information to help the scientist, or his application programmes, 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.

3. Information to enable the CAA 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 redundant copies of information. This is curation information.

The approach adopted by CAA is that all information is held in a homogeneous way, and the various CAA applications, such as search, data extract and preparation for delivery, shall select the metadata parameters as required. 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 (Figure 1). This framework allows the development of generic metadata handling tools, independently 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 collection of terms used to populate the data model.

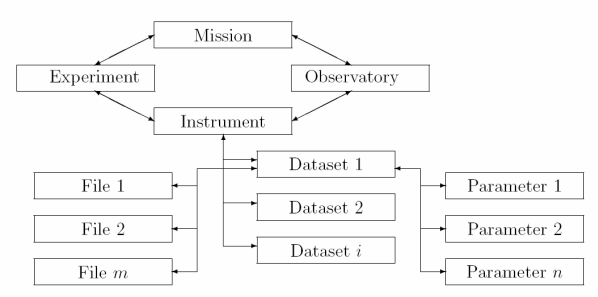

Figure 1: The CAA Metadata Hierarchy.

3. CAA INFRASTRUCTURE

The main CAA infrastructure is built using standard workgroup level computing hardware and open source software. This is combined with specialised software for the production, handling, manipulation and visualisation of the Cluster instrument data and associated products.

3.1 Hardware

The CAA hardware infrastructure is located at ESA/ESTEC in Holland. The three core elements of the CAA hardware infrastructure (Figure 2) are the storage, processing and network:

1. ***Data storage and backup*** .
   The total product volume to be handled by the CAA is expected to be in the range of a few tens of Terabytes, built up over the course of the mission. The CAA will keep all current data available online to ensure the user has rapid access to the data that they require. This level of online storage is well within the capabilities of current commercial-of-the-shelf components and the CAA system is based on a scaleable storage area network (SAN) architecture using SATA based RAID devices with an initial capacity of some 12 TB. Backup and disaster recovery are provided via a LTO-2 tape library with tapes stored in a separate location.

2. ***Data processing system.***
   The CAA processing system handles i) data processing for those instruments that provide software rather than pre-processed products, ii) data ingestion, iii) data management iv) data manipulation and visualisation and v) data delivery. The system is based on a high-availability dual node system running the Debian distribution of the Linux operating system. The two main system nodes are identical consisting of dual Xeon 3 Ghz processors coupled with 4GB of memory. A further system is used for software development and testing activities.

3. ***Network connectivity***.
   The significant volumes of high-resolution data to be collected from the instrument teams and delivered to the science community require that the CAA has good network connectivity to the public internet. This requirement has taken on additional significance with the decision to use the CAA for network distribution of the raw data to the instrument teams during the extended mission. To ensure the necessary connectivity, the CAA is directly connected to the existing ESAgrid research network infrastructure that provides a Gigabit link to the Dutch academic network, Surfnet.

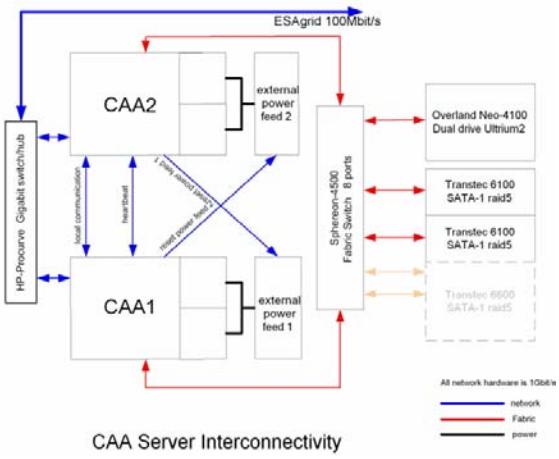

Figure 2: CAA Hardware Configuration.

3.2 Software

A simplified view of the CAA system architecture is shown in Figure3.

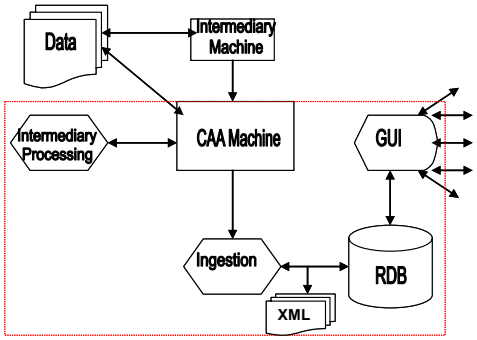

Figure 3: CAA System Overview.

For nine out of the eleven Cluster instrument teams, pre-processed data, plots and documentation are either delivered to, or collected by the CAA. For the two remaining particle experiments (PEACE and RAPID), software, calibrations, validation information and reformatted raw data are used to generate the data products on the CAA system. These processed data sets are then moved to the ingestion area and treated in the same way as the delivered products. Options for on-demand processing that would allow users to tailor the processing parameters or produce value added products are under consideration for several of the other instruments but have not yet been implemented. User access to the system is via a web based forms interface and associated underlying web services.

The core components of the CAA data management system are shown in Figure 4. In the current implementation these consist of the data ingestion system, a relational database system and the web based graphical user interface.

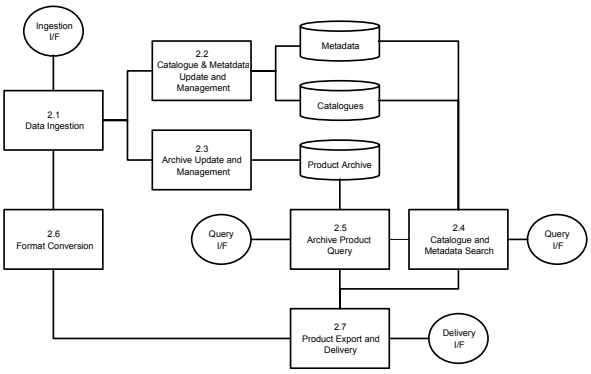

Figure 4: Components of the CAA Data Management System.

The data ingestion system is aimed at providing a stable and automated system for ingestion of the multiple types of data delivered by the eleven experiment teams. To ensure consistency of products, a stringent validation process is performed that verifies that the products are formatted correctly, are consistent and contain all the necessary metadata information. More detailed product and quality assurance activities that involve comparison of similar products generated from different instruments are subsequently carried out as part of testing by a CAA working group consisting of the instrument teams and science users.

The relational database system (RDB) is used to hold the main file catalogue, system configuration information, file processing status and all associated tables. The full product descriptions are stored using an XML schema that reflects the CAA model hierarchy described above. Subsets of this information are exported to the RDB to facilitate standard SQL based queries. The data is stored as a collection of individual files each covering a certain interval of time that may vary from minutes to many hours depending on the complexity and temporal resolution of the data. The collection of files with a common metadata description is conceptually treated as a single timeline. Bespoke software is used to seamlessly concatenate or extract segments from files within a given dataset based on a user requested time interval.

The initial web access system (Figure 5) provides a basic GUI that allows the simple search of ingested CEF files. A standardised look and feel (Figure 6 and 7) is provided via the use of XML encoded content that is transformed using a style sheet prior to delivery via the web server. Access to the file catalogue database is via a collection of standard Perl modules and scripts that have been designed specifically for the purpose and that are used in conjunction with a set of stand-alone tools that are used for extraction, manipulation, packaging and visualisation of the actual data.

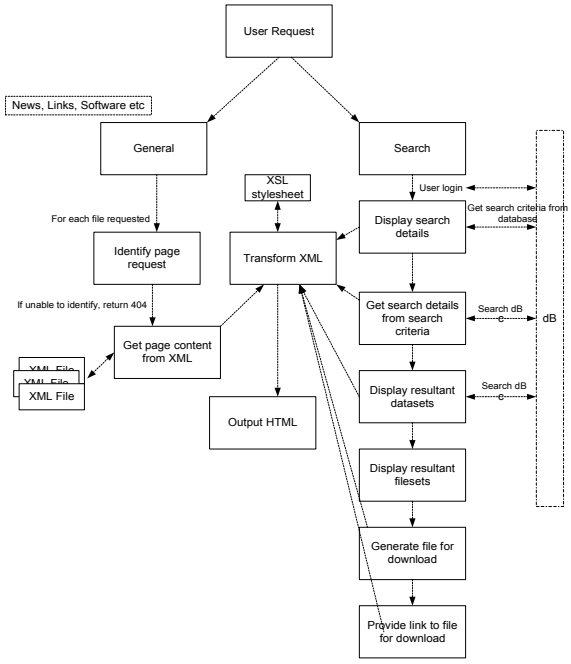

Figure 5: CAA Web Based Data Access System.

The search capabilities are currently limited to selections based on time range, spacecraft, instrument, instrument type and measurement type. Options for additional search criteria including the ability to select datasets based on any of the terms held within the XML metadata descriptions such as data type or processing level are planned for the next major release of the CAA system.

Any data products that have been successfully ingested are publicly accessible via the web interface. To streamline the handling and delivery of the potentially very large data volumes associated with the highest resolution data, the CAA provides the capability to select, and have returned, those specific intervals of interest which may be only a few minutes in duration and correspond to just a small fraction of the stored data file.

Figure 6: CAA Web Based Graphical User Interface.

Figure 7: CAA Initial data selection web page.

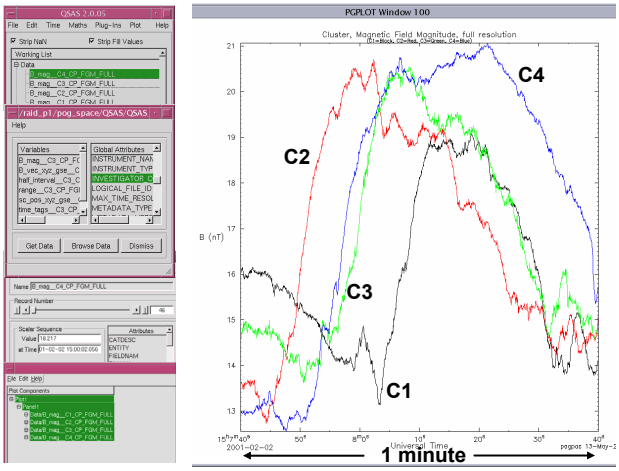

Figure 8: QSAS Displaying 4 spacecraft high-resolution magnetic field data from the FGM experiment.

3.3 Tools

To assist with the development of the standard product formats and descriptions used within the CAA a set of low-level tools have been produced. Although these are used internally within the CAA data management and web front-end systems, they are stand-alone and are also widely used within the instrument teams to verify the technical compliance of their products prior to delivery to the CAA. The software are downloadable from the CAA web site and currently include tools for verification of the CEF syntax (CEFpass), conversion of the CEF header information into standard XML metadata descriptions (CEF2XML), resolving CEF include files (CEFresolve) and for combining, splitting and extracting intervals from one or more files (CEFcombine). Further tools such as a library of routines for reading data into the widely used Interactive Data Language (IDL) package are currently under development.

Other groups within the Cluster community are already adapting existing tools and analysis packages to be compatible with the CAA. The CIS instrument team are developing a CEF-2 interface for their CL analysis package and Imperial College have already updated their Qtran format conversion software and QSAS analysis package. QSAS provides a complete data visualisation and analysis package and with the built-in support for the CEF-2 data format is able to directly handle the digital data products downloaded from the CAA. An example is shown in Figure 8 where one minute of high-resolution magnetic field data from each of the four Cluster spacecraft has been downloaded from the CAA, read in and over-plotted using the QSAS software. QSAS also provides many specialised tools for handling time series and vector data, such as minimum variance analysis and many of multi-spacecraft techniques described by the ISSI working group on "Advanced Analysis Methods for Data from Clusters of Spacecraft" (Paschmann and Daly, 1998)

4. SCIENCE DATA DELIVERIES

Development of the best achievable quality, standardised, high-resolution data is the primary aim of the CAA. To ensure this, ESA has for the first time provided direct support to the Cluster instrument teams to assist with the development and production of the archival products. The detailed descriptions of the supplied instrument team products are given in the Interface Control Documents (ICD) for each team, a summary of which can be found in the individual instrument team archiving papers in these proceedings. The steps leading to the production of the instrument ICD from the original outlines given in the instrument Archive Plans (AP) and development of test products are outlined in Figure 9.

The instrument teams have been split into two groups that consist of i) the fields and waves instruments, and, ii) the particle related instruments. An ESA/CAA archive developer has been assigned to each of these groups to assist with the development process and to help ensure consistency and compatibility across the full range of CAA data products.

To assist the science user in locating intervals of interest that can then be studied using the best-quality, high-resolution CAA data, the full CSDS prime, summary and quicklook survey products are being converted to CAA compatible formats and metadata descriptions. The CSDS versions of the PP/SP will continue to be available first from the national data centres and will subsequently appear on the CAA once the corresponding high-resolution products have been delivered. In addition to the core science products, information on the Cluster spacecraft status, planning and operations are being provided by the ESA Space Operations Centre (ESOC).

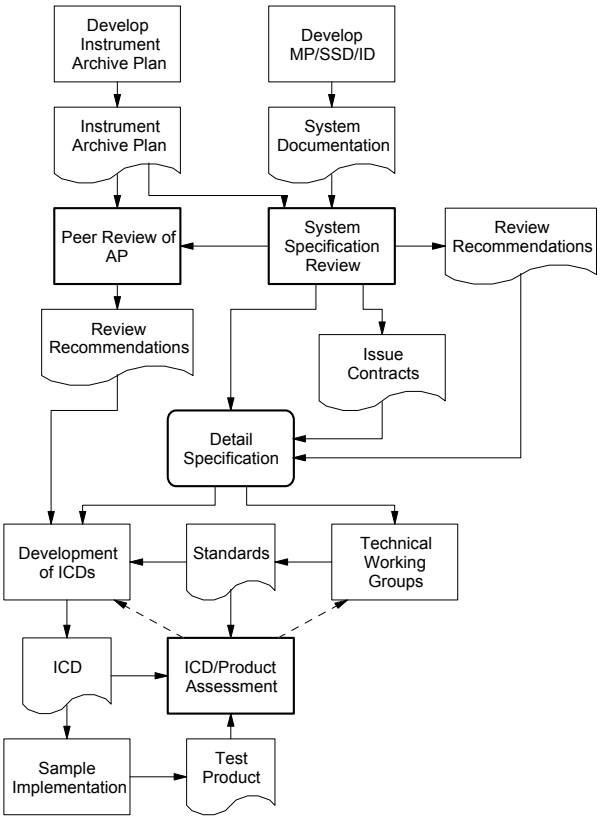

Figure 9: CAA Product Design Process

A preliminary review of the instrument team products, looking primarily at the technical aspects of the ICDs and example products, was held in December 2004 and was followed by a full archive implementation review in May 2005. The CAA became publicly available for Beta testing in September 2005 an activity that will continue until the next major release planned for early 2006. A summary of the ICD version, data products and intervals over which they have so far been provided is given in Table 1 for the wave instruments and Table 2 for the particle instruments.

TEAM	ICD Updated	Data Products	Interval Provided
DWP	Issue 1 04-May	TCOR, LOG, COR, UT_PIOR	20010201
EFW	Issue 1 23-Apr	L1_P12, L1_P1, L1_P2, L1_P34, L1_P3, L1_P4, L2_E, L2_P, L3_E, L3_P	20010202 -> 20011231
FGM	Issue 1 22-Apr	FGM_FULL, FGM_5VPS, FGM_SPIN, FGM_CALF, FGM_CALA, FGM_CAVF, FGM_GAPF, FGM_VALF	20010203->20010501
STAFF	Issue 2 09-May	DWF_HBR, DWF_NBR, AGC, PSD, SM, spectrograms,	20010202->20011231
WBD		Spectrogram plots	20010201->20011231
WHISPER	Issue 1 04-May	NATURAL, WAVE_FORM_ENERGY, ACTIVE, PASSIVE, ACTIVE_TO_PASSIVE_RATIO, HK (Sounding), ACTIVE_EVENT, NATURAL_EVENT, ELECTRON_DENSITY	20010202 -> 20010831

Table 1: Product Delivery Status – Fields and Waves

TEAM	ICD Updated	Data Products	Interval Provided
ASPOC	Issue 1.1 11-May	STAT, IONC, IONS, CMDH	20010201-> 20010630
CIS	Issue 1.0 4-May	HIA: IONS_CS,IONS_PEF, IONS_PF,IONS_RC CODIF: TBC	20010201->20020630 Only C1/C3
EDI	Draft 2 4-May	DOC, EGD, CAVEATS, MPD, MSF, PP, Ppplus, QSTAT, Survey plots	20010201 -> 20021231
PEACE	Issue 1.0 5-May	Many, see ICD	20010201 -> 20011231
RAPID	Issue 1.0 21-Apr	Many, see ICD	20010201 -> 20010630

Table 2: Product Delivery Status – Particles

The longer-term schedule is to deliver at the rate of at least two years of data per year until the backlog has been removed, after which data is expected to be delivered approximately one year after acquisition. The recently approved 2nd extended of the mission means continued operation of CAA active phase until the end of 2010 although at reduced level from 2008.

Ongoing quality assurance of the delivered products is being undertaken by a CAA cross-calibration working group drawing on expertise within the instrument teams and from the broader Cluster science community. Teams within this working group are using multi-instrument, multi-spacecraft comparison techniques to compare related products, such as plasma density, that can be derived in different ways from several instruments e.g. direct particle measurements or from wave observations.

5. SUMMARY

The CAA and associated activities within the instrument teams will ensure a long-term legacy from the Cluster mission that will allow scientific exploitation of the Cluster data to continue for many decades after the mission itself has completed.

The CAA has completed the initial development phase. The file format and data description standards, that are vital to ensuring the long term accessibility and usability of the science products, have been completed. These standards have been applied to the detailed specification and documentation of the science products provided by each of the instrument teams. The CAA system has been installed and the initial releases of the data management system and web based front-end access have been implemented.

Routine deliveries of the raw, level-1 and processed science data products are taking place and most of the instrument teams have submitted a significant proportion of the data products covering the first year of operations, and in some cases are making good progress on deliveries of data from subsequent years.

The CAA is currently available for public Beta testing and can be accessed at http://caa.estec.esa.int/. During the active operations phase the capabilities of the web interface will be extended to provide access to non-digital data products and for value added services such as on demand plot production and the ability to pose queries, not just on the catalogues, but also on the data itself. Routine submissions of data from the instrument teams shall continue as well as improvement of and update to the existing data as a result of the cross calibration activities and feedback from the science community.

ACKNOWLEDGEMENTS

This paper is dedicated to the memory of Tobias Eriksson, one of the core ESA development team at ESTEC, who died unexpectedly, shortly after the workshop.

REFERENCES

Allen, A – CAA User Requirements Document, ESA, CAA-QMW-UR-0001, Issue 1.2, 4 Nov 2003

Allen, A., S.J.Schwartz, C. Harvey, C. Perry, C. Huc, P. Robert, Cluster Exchange Format - Data File Syntax, ESA, DS-QMW-TN-0010, Issue 2.03, 21 Sept 2004.

Harvey, C.C, F. Deriot, A.J. Allen, C. Huc, M. Nonon-Latapie, C.H. Perry, S.J. Schwartz, T. Eriksson & S. McCaffrey, Cluster Metadata Dictionary, ESA, CAA-CDPP-TN-0002, 2.0, 17 Mar 2005.

Paschmann, G and P.W. Daly, Editors, Analysis Methods for Multi-Spaceraft Data, ISSI, SR-001, July 1998.

Perry, CH, CAA System Specification Document, ESA, CAA-EST-SS-0001, Issue 1.0, 7 Nov 2003.