Chapter 15
Flexible Tools for Accessing the Cluster Archives

E. Gamby, J. De Keyser, and M. Roth

Abstract Nowadays, multi-spacecraft analysis is becoming common practice. The lack of homogeneity in data archive structure, access types and data formats results in a lot of time spent to search, retrieve and reformat data. We have developed a software model dealing with these issues. In this model, we integrate an abstraction layer for accessing and caching archive data from different providers. A second abstraction layer allows converting the specific data formats into a common working format. This model has been implemented in the MIM software, developed by the Belgian Institute for Space Aeronomy. In particular, it provides a framework for accessing the CAA (Cluster Active Archive) and for exploiting Cluster data with multipoint analysis tools such as, e.g., gradient computation and magnetopause reconstruction.

15.1 Introduction

Space science data are usually stored in large, centrally managed archives. Scientists that access these data must deal with a plethora of access types, archive structures and data formats. This lack of homogeneity makes it more difficult to implement algorithms based on data from multiple sources. We present a software model that we have developed to handle this situation.

15.2 Software Infrastructure

The whole sequence of activities that scientists perform on data can be seen as a single workflow. In this workflow, we identify three main processes:

1. Accessing the remote archive and retrieving data
2. Converting these data into a common format
3. Data analysis and visualization

E. Gamby (✉), J. De Keyser, and M. Roth
Belgian Institute for Space Aeronomy, Brussels, Belgium
e-mail: emmanuel.gamby@aeronomie.be

H. Laakso et al. (eds.), *The Cluster Active Archive*, Astrophysics and Space Science Proceedings, DOI 10.1007/978-90-481-3499-1_15,

The output of the second process provides a uniform data model to the scientific algorithms of the third one, so that these algorithms don't have to deal with data heterogeneity. We base our software model on this three-fold decomposition.

15.2.1 The Retrieval Process

This process (see Fig. 15.1) is responsible for downloading data from remote archives. Each connection to such an archive is called a channel. A channel specifies where and how to find spacecraft data for a given time interval. The process keeps a queue of pending user requests, or jobs.

A channel has two software components:

1. A data source, which encapsulates the communication protocol used by the remote archive and which hides the access details from the other processes.
2. A local cache, in which the previously downloaded data are stored. This caching strategy aims at saving time and bandwidth when the same data are requested several times.

15.2.2 The Formatting Process

The formatting process (see Fig. 15.2) merges the different formats into the same data model and organizes these data into datasets. Each conversion is steered by a format description that specifies how to extract the data and metadata from the source format. A dataset is a collection of instrument data, such as measurements of the magnetic field, of ion densities ... and of the associated ancillary data, such as the spacecraft position and velocity. These quantities are called data items in our software model.

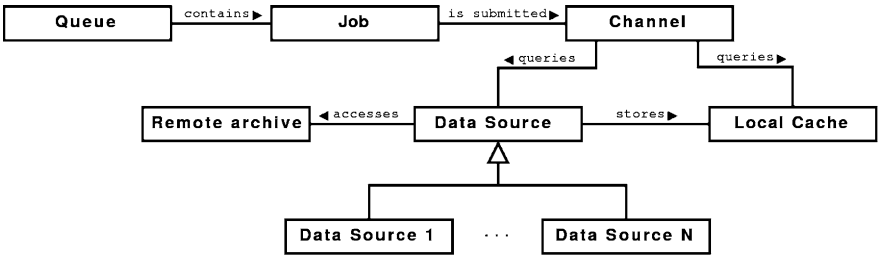

Fig. 15.1 Retrieval and caching of data from various data sources. Black arrows represent associations between concepts; white arrows are specializations of concepts

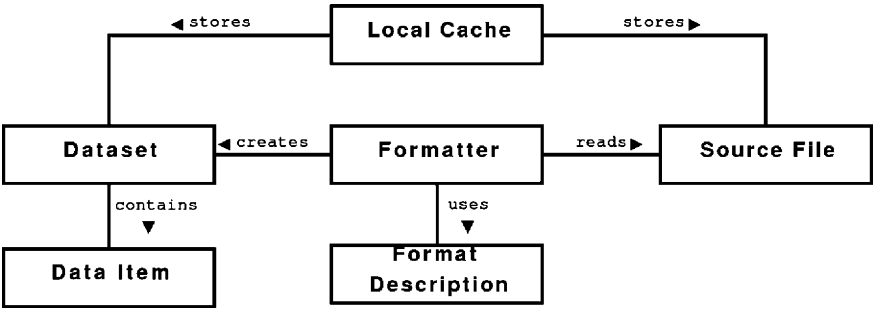

Fig. 15.2 Importing science data into the MIM internal format

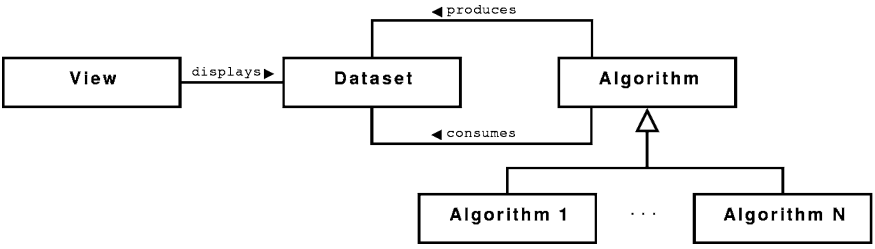

Fig. 15.3 Analyzing and visualizing data. Algorithms can both consume and produce datasets, making the system very flexible

15.2.3 The Analysis and Visualization Process

This process includes the scientific algorithms and visualization techniques (see Fig. 15.3). As all algorithms share the same data model, the result of one of them can be used as input of another one. Moreover, adopting such a flexible model eases the combination of data issued from different sources (spacecraft or numerical simulations) in various algorithms. For instance, multi-spacecraft data can be fed into algorithms specifically tailored for multi-point analysis and be compared to numerical predictions.

15.3 The MIM Implementation

The above model is implemented into the MIM (Manager of Interactive Modules) software, a data management and modeling tool developed by the Belgian Institute for Space Aeronomy. It provides an interactive environment for accessing and processing data from a multitude of experiments. MIM is a MATLAB application, written in an object-oriented way. It consists of a set of modules, amongst which the **channel manager** and the **dataset formatter**, respectively

implementing the retrieval and formatting processes. Other modules are dedicated to **scientific algorithms**, such as multi-spacecraft gradient computation and magnetopause reconstruction.

In the **channel manager**, we define channels for accessing the CAA (Cluster Active Archive) and CSDS (Cluster Science Data System) http-based archives, as well as the Themis ftp site. Since data download can lead to significant wait times, cache hits are important. A high cache hit rate can be achieved as scientists often work for a prolonged time with a limited set of events.

The **dataset formatter** is the module responsible for creating format descriptions and applying them to retrieved data. We have defined templates to import CDF (Common Data Format), HDF (Hierarchical Data Format) and CSV (Comma Separated Values) files. In particular, we can import CDF data from the Cluster and Themis repositories.

Amongst the **scientific modules**, algorithms are available to process multi-point data, e.g. gradient calculation [2], magnetopause reconstruction [1], time delay analysis … An example of a calculation of the least-squares gradient during a Cluster pass through the inner magnetosphere is shown in Fig. 15.4.

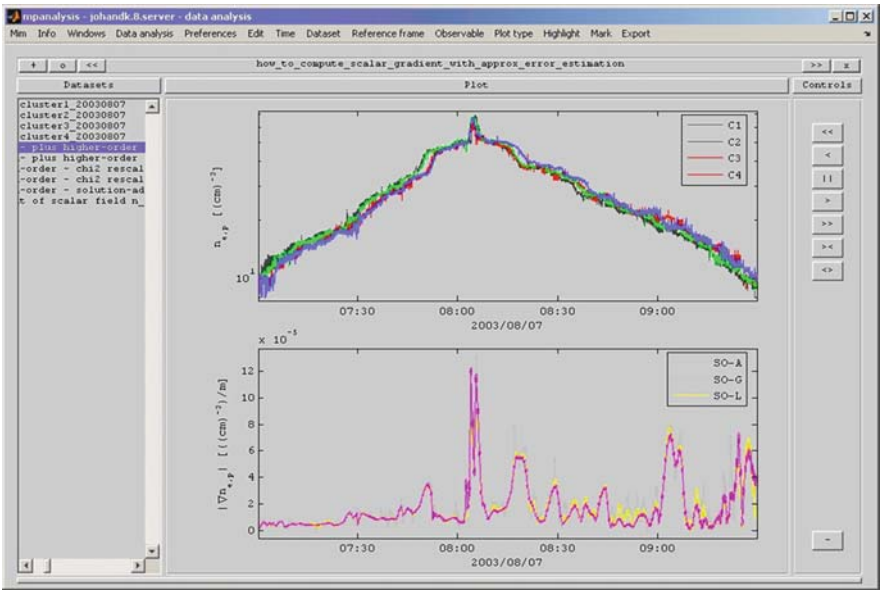

Fig. 15.4 Calculation in MIM of the least-squares gradient of plasma density obtained from the WHISPER instruments on the four Cluster spacecraft during a pass through the inner magnetosphere on 7 August 2001; perigee is encountered at about 4 R_E around 08:05 UT. The increasing and decreasing density before and after the perigee crossing reflect the entry into the outer regions of the plasmasphere. The model assumes the gradient to be locally constant. The distance over which the gradient can safely be considered constant is estimated automatically. The output is the computed gradient vector with its error margins

15.4 About the Importance of Metadata

In the software model introduced in the previous sections, extensive use is made of metadata. Indeed, meta-information is very useful to automate data processing. For example, normalizing data from heterogeneous sources requires information such as the units or the coordinate system in which data are expressed. Moreover, data visualization can benefit from metadata such as scale type (linear or logarithmic), minimum and maximum values, axis labels ... Unfortunately, so far, no standard dictionary of metadata has been defined. However, with projects such as ISTP/IACG, CSDS, SPASE and the Cluster Metadata Dictionary, there is an ongoing effort towards standardization:

- The ISTP/IACG (International Solar-Terrestrial Physics/Inter-Agency Consultative Group for Space Science) Guidelines describe a standard way of structuring data in CDF files [3].
- The CSDS (Cluster Science Data System) is an extension of the ISTP guidelines including science tool driven metadata that is machine readable.
- The SPASE (Space Physics Archive Search and Extract) data model provides a rigorous hierarchy of metadata but is so far essentially oriented to data access and retrieval [5].
- The Cluster Metadata Dictionary, which has been specifically developed for the Cluster mission, borrows concepts from the two preceding dictionaries [4] to provide a complete set of metadata accounting for both the retrieval and the science processing in advance of their adoption by SPASE.

One of the main advantages of the latter dictionary is that it provides metadata down to the level required by science tools, while very often, standards tend to stop at metadata needed for search and delivery.

15.5 Conclusion

Manipulation and interpretation of data from multiple sources require specific functionalities to deal with heterogeneities in data formats and archive structures and with the lack of standardization amongst metadata. These functionalities are often not available in standard data processing environments. We have proposed a software model that integrates at least some of the required functionalities. This model has been successfully implemented in the MIM application (downloadable from http://www.spaceweather.eu/software/mim).

Acknowledgements The authors wish to thank the Belgian Science Policy (BELSPO) and the Solar Terrestrial Center of Excellence (STCE) for their financial support.

References

1. De Keyser, J., Roth, M., Dunlop, M.W., Rème, H., Owen, C. J., Paschmann, G.: Empirical reconstruction and long-duration tracking of the magnetospheric boundary in single- and multi-spacecraft contexts. Ann. Geophys. **23**, 1355–1369 (2005)
2. De Keyser, J., Darrouzet, F., Dunlop, M.W., Décréau, P.M.E.: Least-squares gradient calculation from multi-point observations of scalar and vector fields: methodology and applications with Cluster in the plasmasphere. Ann. Geophys. **25**, 971–987 (2007)
3. Goddard Space Flight Center: ISTP/IACG guidelines. Space Physics Data Facility Website. http://spdf.gsfc.nasa.gov/sp_use_of_cdf.html (2006). Accessed 2 June 2008
4. Harvey, C.C., Allen, A.J., Dériot, F., Huc, C., Nonon-Latapie, M., Perry, C.H., Schwartz, S.J., Eriksson, T., McCaffrey, S.: Cluster Metadata Dictionary, version 2.3. The Cluster Active Archive Website. http://caa.estec.esa.int/documents/DataDic.pdf (2008). Accessed 2 June 2008
5. SPASE Consortium: A Space and Solar Physics Data Model, version 1.2.1. SPASE Website. http://www.spase-group.org/data/doc/spase-1_2_1.pdf (2008). Accessed 2 June 2008