Mem. S.A.It. Vol. 74, 819

Memorie della

SOHO Long-term ARchive: Information Retrieval Approach

A. Cora[1], E. Antonucci[1], G. Dimitoglou[2], C.A. Volpicelli[1] and S. Giordano[1]

[1] Istituto Nazionale di Astrofisica – Osservatorio Astronomico di Torino, via Osservatorio 20, 10025 Pino Torinese e-mail: cora@to.astro.it

[2] SOHO ESA/NASA Project Scientist Team, Laboratory of Astronomy & Solar Physics, NASA Goddard Space Flight Center, Greenbelt, MD, USA

Abstract. The SOho Long-term ARchive (SOLAR) is one of the three European data archives of the SOlar and Heliospheric Observatory (SOHO) instruments. The SOHO archives adopt the Internet and the World Wide Web (WWW) as the platform to search, retrieve and disseminate science data and mission information. This paper presents the architecture and design overview of the archive built at the Turin Astronomical Observatory and a brief description of the available web-based Graphical User Interfaces developed to have access to the data remotely. SOLAR is foreseen to be operational not only during the SOHO mission (recently extended to 2007) but also for a 10-year period following the mission end (2007-2017).

Key words. SOHO – data archive – solar databases

1. Introduction

SOLAR is one of the three European archives of SOHO, mirroring the NASA Goddard Space Flight Center (GSFC) archive. The other two mirrored archives are located at the Rutherford Appleton Laboratory (UK) and the Multi-Experiment Data Operation Centre (MEDOC) of the Institut d'Astrophysique Spatiale (France). Each site has implemented the archive using the same software but different hardware platforms and different Graphical User Interfaces. SOLAR stores the following types of datasets and data products:

- calibrated scientific data: the observational data of the 12 instruments working on board to SOHO.
- summary data: a simple collection of daily observations of the instruments.
- software: packages developed by the instrumental teams to reduce and analyze the observations.
- basic documentation useful to guide the catalog navigation (information related to Joint Observing Programmes, Campaigns, etc).

Send offprint requests to: Alberto Cora
Correspondence to: via Osservatorio 20, 10125 Pino Torinese Italy

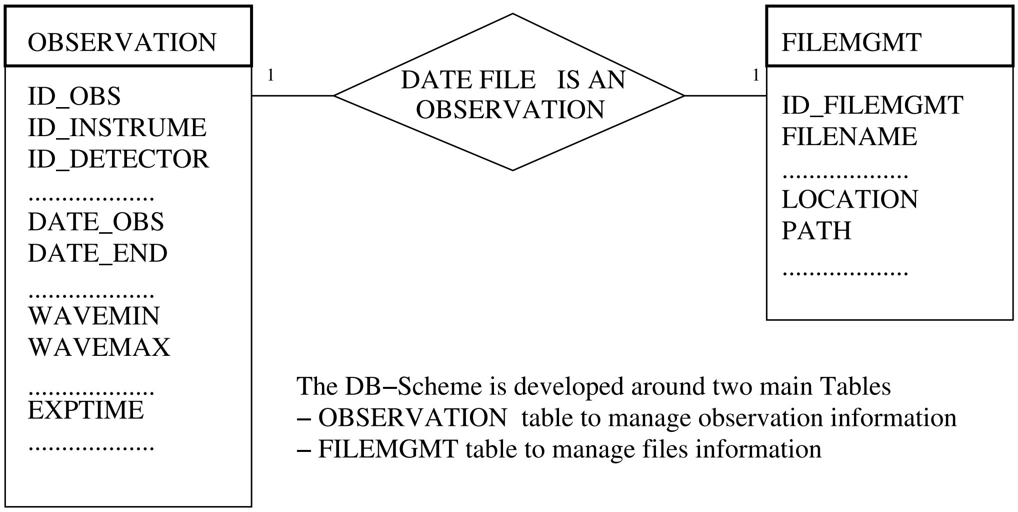

Fig. 1. Entity-relationship between observation and file information

2. Information Retrieval Approach

In SOLAR the information retrieval is a process in support of the user's task finalized to retrieve data sets. The scientists engage in a variety of information seeking strategies, all of which should be supported by the Information Retrieval System (IRS). On the base of the experience developed by our partners on the other three SOHO archives, SOLAR combines the functionalities of a conventional IRS and hypertext systems in order to take into account a broad range of information needs that different users may have at different times. The retrieval approach adopted by SOLAR is twofold. On the one hand the system is designed to retrieve useful information from a catalog to access SOHO data. On the other hand the system offers documentation and software to reduce and analize the data.

3. DataBase Structure

SOLAR is designed to build and update the catalog of the SOHO observations; the information is extracted directly from the FITS files through a software called *extract-metadata*, developed by the GFSC/ESA team. This software creates and populates the SOLAR archive database that will be consulted by the users. From this point of view the FITS files undergo a process of representation, leading to a formalized query; lets us extract the essential properties of data while ignoring unnecessary details. The conceptual model for organizing data files within the SOHO archive is simple: a data file is equivalent to an 'observation'. The core of database is costituted by two tables that relate the data file information to the information defining the observation(see Fig.1). Several observations can be grouped in a 'study'. Typically, all the observations belonging to a study share some parameters, like a scientific objective, or the object being observed. The only requirement we impose is that all observations of a particular study should have been taken with the same instrument. Some instruments organize observations in studies, and some do not. Another level to organize series of observation is the so called 'campaign'. Observations related to a campaign can be from different instruments. Also, an observation can be related to more than one campaign. The system should provide quick access to the desidered documentation and guide the user througt the cat-

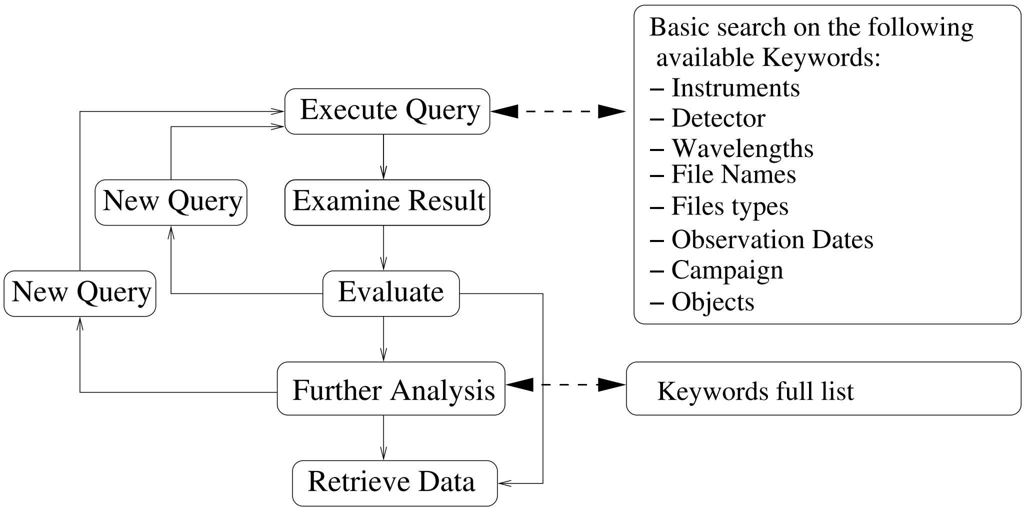

Fig. 2. Model showing the feedback loop for query refinement which leads to the multi-step query process

alog (meta-data) to choose the data. The database operates as a catalog. It does not store the actual scientific observations but information about the data (meta-data) along with pointers that link each catalog entry with the actual observation. An Oracle Relational Database Management System (RDBMS) is used to store the catalog information, C routines are used to extract metadata from FITS files and Oracle PL/SQL routines are used to populate the database tables. As a typical relational database, it consists of tables of highly structured records with fixed-length fields. Keywords and fields are defined on the basis of the keywords determined by the instrument science teams.

4. Remote Catalog Access

It is widely recognized that there is a need for good methods for disseminating large amounts of digital data generated by space missions. The SOHO archives adopt the World-Wide-Web (WWW) as the platform to disseminate mission information and allow access for data retrieval. The Graphic User Interfaces (GUI) developed by the MEDOC team allows to build very complex operations within a single query. An alternative non-Java, web-enabled interface has also been developed by the ESA team at GSFC providing catalog search and retrieval facilities. Both interfaces has been installed and are operational at the SOLAR site. Through the WWW interface is possible to do a multi-step query process in which queries are progressively refined. The users perform a query on a dataset, examine the results of that query and then make a scientific judgements about these results. It is also possible to have further information on a single observation selecting it on the list. At this stage of the catalog exploration, it is possible to proceed to a new query or to refine the original query. In general the final result will be achieved only after stepping through many queries (see Fig.2). At the end of the process, the user decides to either save the information related to the requested data files or retrieve them; in this case the requested observations are compressed into a cache area and then made available for download at a temporary web address (specific to the request).

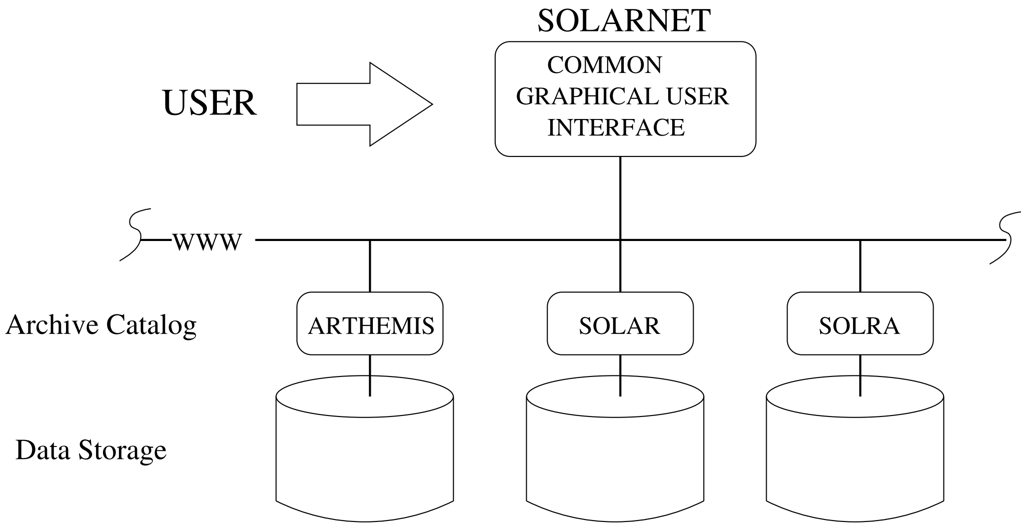

Fig. 3. SOLARNET as a Distributed Database

5. Future Developments

SOLAR together with SOLRA, the archive of the radio data at the Astronomical Observatory of Trieste, and ARTHEMIS, the archive of ground-based italian data at the Astronomical Observatory of Naples, forms SOLARNET, the first complete network of thematic (solar physics) archives of astrophisycal data in Italy. SOLAR and SOLARNET are becoming an essential part of the European Grid for Solar Observations (EGSO), a 3–year project funded by the European Union under the Information Society Technologies Programme. EGSO will lay the foundations of a Worldwide Virtual Solar Observatory.

6. Conclusions

To take full advantage of the scientific observations from space missions, the data need to be easly and quickly accessible and available to the scientists. The development, implementation and operation of SOLAR at the Astronomical Observatory of Turin provides exactly that functionality as a robust, highly available, around the clock access of SOHO data via the WWW. The SOLAR system contributes in making data sets available to the local participating science teams and the wider solar physics community. This will be accomplished by maintaining the data, obtained by the scientific teams working with SOHO, for 10 years after the end of the mission (until 2017); providing the necessary basic software to analyze scientific data; providing access to a wider community that will allow the use of SOHO data for public outreach. SOLAR Archive is operating since the beginning of March 2002 at URL: http://solar.to.astro.it

References

Dimitoglou G., Sanchez L. 2001, "The SOHO Data and Information System", ASP Conference Proceedings, Vol. 225.Conference, p.173

Cora A., Antonucci E.,Volpicelli C.A., Dimitoglou G. Mem. S.A.It., in press "SOho Long-term ARchive (SOLAR)"