Astronomical Data Analysis Software and Systems XIX **O01**
ASP Conference Series, Vol. XXX, 2009
Y. Mizumoto, K.-I. Morita, and M. Ohishi, eds.

ESA New Generation Science Archives: SOHO and EXOSAT

P. Osuna, C. Arviset, D. Baines, I. Barbarisi, J. Castellanos, N. Cheek, H. Costa, N. Fajersztejn, M. Fernandez, J. Gonzalez, A. Laruelo, I. Leon, I. Ortiz, J. Salgado, A. Stebe, D. Tapiador

ESA-ESAC, Science Operations Department, PO Box 78, 28691 Villanueva de la Canada, Madrid, Spain

Abstract. The ESAC Science Archives and VO Team (SAT) has developed a new infrastructure for the development and maintenance of the ESA space based missions' Science Archives. This infrastructure makes use of state-of-the-art technology to overcome some of the already known limitations of older technologies, used for the building of the current archives, the older of which has been live since 1998. This paper describes how the SAT approached the issue of re-engineering their infrastructure to result in a more flexible, reusable, robust and cost-effective way of building their archives. It also describes how the new technology has been applied to the building of two Science Archives from scratch: the SOHO Science Archive (a Solar physics mission) and the EXOSAT Science Archive (an astronomy mission).

1. Introduction

The European Space Astronomy Centre (ESAC) located near Madrid, Spain, is the default location for ESA's Science Operations Centres for space based missions. It is also the default location for most of the Science Archives for those missions. A dedicated Science Archives and Virtual Observatory Team (SAT henceforth) is in charge of the design, development, operations and maintenance of the aforementioned archives. ESA's space based missions cover different areas in science: Astronomy, Solar physics, planetary physics, fundamental physics.

2. ESAC Archives History

Prior to 1996, ESA was not holding the scientific data of its missions. Such was the case of IUE or EXOSAT projects, where the data were archived at LAEFF (Spain) and HEASARC (USA) respectively. After the successful build of the ISO Post Mission archive by the SAT at ESAC, the idea to keep all data for the missions and have a team building all Science Archives for the missions started to take shape. The SAT started henceforth building archives for the different missions (see [Arviset 2006]) among which we can cite the ISO Data

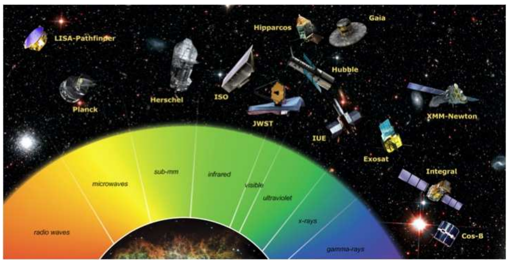


Figure 1. ESA's Space based Astronomy and Fundamental Physics missions.

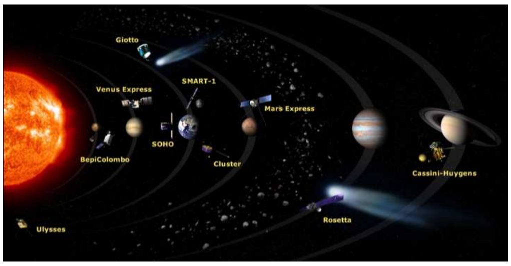


Figure 2. ESA's Space based Planetary and Solar Physics missions.

Archive (IDA[1]), the XMM-Newton Science Archive (XSA[2]), the Integral SOC Science Data Archive (ISDA[3]), all ESA's Planetary mission archives (Rosetta, Mars Express, Venus Express, Smart-1, Huygens and Giotto) (PSA[4]), and the Herschel Science Archive (HSA[5]), currently being used during the operations phase.

All these archives were built in a flexible and modular way, permitting not only a seamless access to our mission's data, but also integration within the Virtual Observatory distributed paradigm by implementation of the International Virtual Observatory Alliance (IVOA) protocols that allow our archived data to take part in the VO world.

Despite all this accomplishments though, the implementation of the aforementioned archives had been done making use of technology which is today out of date. Taking into account that the team started building the ISO archive back in 1997, it is easy to understand how the IT technology has evolved from those times. To give a simple example, some of the nowadays everywhere available scroll-bars had to be implemented by hand in the case of some of our subsystems in the ISO archive. The network was starting to be born, and there were no

[1] `http://iso.esac.esa.int/ida`

[2] `http://xmm.esac.esa.int/xsa`

[3] `http://integral.esac.esa.int/isda`

[4] `http://www.rssd.esa.int/psa`

[5] `http://archives.esac.esa.int/hsa/hsa.html`

firewalls, nor issues with ports, etc., nearly no "hackers" that could threat the systems' safety.

The sector of Software and Systems engineering has changed drastically since the times we started to build our science archives, and in 2006 the SAT took the decision to review the technology used and to try to adapt it to the new trends. The concept of an "Archives Building System Infrastructure" (ABSI henceforth) was studied, and two missions' archives were selected to measure the goodness of the new technology: SOHO and EXOSAT.

3. ESAC Archives Evolution

The Science Archives and VO team undertook an exercise of self-auditing, where the technology used for building our science archives was scrutinised to find eventual possible problems. As a result of this exercise, a document called "The Archives Building System Infrastructure Study Report" was produced. This document lists the different difficulties found in using a somehow obsolete technology and possible areas for improvement. An example of the type of issues found can be seen here:

- Communication Client-Server done through home made RPC (Java serialised objects)
- Transport done through TCP-IP in compressed mode using ports like 5433, 5443, ... and standard port 80 for users behind firewall, but...
- Business layer makes uses of port 80 and blocks its usage to any other application running on the same machine
- TCP-IP is a lower level protocol than HTTP, which supports tunneling, encryption, connection timeouts, request compression, ...
- Business layer mixing service, transport and logic in a single implementation
- load balancing, security and proxy redirection not available in our home-made server
- ad-hoc persistence layer difficult to maintain and expand

These issues were investigated and eventually, internal requirements were set on the type of infrastructure we would ideally like to have in order to be able to overcome all the difficulties, resulting in an internal "ABSI Requirements Document". Examples of this type of requirements are given here:

Main points

- the infrastructure should be as open as possible, with a community big enough to ensure estability and permanence
- should be as light as possible (both client and server)
- should be modular and flexible

Client layer

- should be light

Server layer

- should be robust and flexible
- DB and persistence layer
- focus on Open source DB
- find a proper persistence layer

After all the requirements were set, decisions had to be taken among the different options in the market, sometimes quite numerous making the process a difficult one. An example of the decisions taken follows:

Client Layer

- Eclipse RPC versus InfoNode/JGoodies on Swing. Decided for lighter InfoNode/JGoodies/Swing

Server Layer

- Application framework: Spring (used in all layers)
- Server container: opted for the light way with Tomcat rather than heavier JBoss, Glassfish, Jonas....

Persistence layer

- Hibernate vs Ibatis. Both very good pros and cons. Decided finally for Hibernate, with better integration with overall project dependant Data Models

4. The "Archives Building System Infrastructure" concept

After studying all the possibilities at hand, the design of the whole infrastructure took place, resulting in the Archives Building System Infrastructure (ABSI henceforth) concept. The ABSI is based on a standard three-tier architecture:

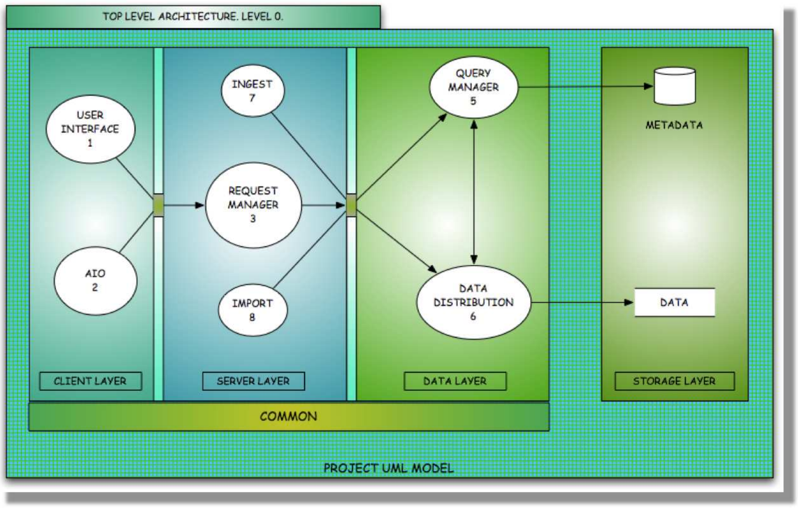


Figure 3. ABSI top level three-tier architecture.

The three-tier arhitecture consists of:

Client layer This layer encapsulates all the interfaces for accessing the system from the external world. Two different types of interfaces are normally offered:

- a standard Graphical User Interface. This interface is intended for human access, and contains search panels that allow the user to select items to filter the searches to the archive
- a machine interface. With the name of Archive Inter-Operability system (AIO) this interface allows scripts to interact with the system. The AIO is also the layer that allows for VO access to the archives built within the ABSI paradigm.

More details about this subsystem can be found in these proceedings ([Fernandez-Barreiros 2009])

Server Layer This layer encapsulates all the back-end machinery internal to the system. The heart of this layer is the RequestManager subsystem that deals with the complexities of the filtering requests coming from the client layer. More details about this subsystem can be found in these proceedings ([Leon 2009])

Data layer This layer encapsulates the data storage and data access subsystems. In the diagram, both appear separated to clearly indicate that the storage part of the data layer can be changed without affecting the rest of subsystems, i.e., if the decision to change database vendor is taken, simply changing the database would be sufficient and the system would not be affected by the change. More details about this subsystem can be found in these proceedings ([Laruelo 2009])

The whole of the ABSI framework is surrounded by the "Project UML Model". This means that although the infrastructure to build different archives is unique, the specificities of the different projects have to be coded according to a data model that has to be built in collaboration with the project, and reflecting the specific needs of the particular mission we are dealing with. We shall see later how important it is to have a well built UML model for the project (and will see differences between a UML for a Solar physics mission and a UML for an Astronomy mission).

4.1. The ABSI elements

The above top level architecture is implemented through the use of several "building blocks" that allow for modularity, flexibility and simplicity in the building of a new archive. These building blocks are:

Interfaces: Object Oriented type of Interface

Modules: It's a software package that packs self-contained functionality. Must be accompanied by an API or similar that gives information on how to consume it.

Component sample: Wraps-up sample code implementing certain functionality. It may contain Modules and/or GlueCode.

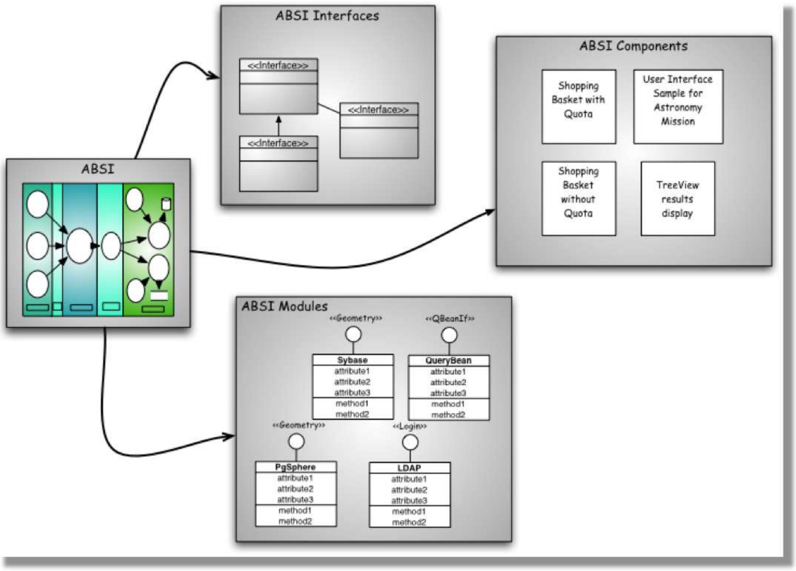


Figure 4. ABSI building blocks.

5. The SOHO Science Archive case

The Solar and Heliospheric Observatory (SOHO) was chosen as the first scientific archive to be implemented with the new technology, becoming the first -together with the EXOSAT one- of the so-called ESA's 2nd generation Scientific Archives. SOHO was one of the ESA Horizons 2000 cornerstone of ESA's Science Programme, implemented as an ESA/NASA collaboration. It is currently in operation, studying helio-seismology, the solar atmosphere, and the solar wind. Launched in December 1995 and placed at Lagrangian point L1, it consists of 12 instrumentes including imagers, spectrometers, radiometers, photometers, particle analysers... It works mainly EUV, UV and visible spectral ranges.
The initial archive for SOHO was developed by the Project Scientist's team during 1995-1998, in a collaboratin with NASA, where ESA was providing the servers and software and NASA was providing the storage. However, since the archive was located at Goddard Space Flight Centre, issues related to long term availability, ESA staff permanence in expatriation and difficult maintainability made ESA reconsider where and how to reallocate the SOHO Science Archive. The decision was then taken to assign the building of the SOHO science archive

to the SAT team at ESAC, using the new technology under development, to benefit from better cost, longer availability, maintainability and longevity, and state of the art technology.

5.1. Building the SOHO archive

As already mentioned above, the most important -and maybe, difficult- part when starting an archive from scratch is the creation of a proper data model. Guided by the SOHO Archive Scientist, the SAT built a UML model for the project. The first peculiarity of the SOHO project is that the hierarchy of data is organised by "Campaigns", "Studies" and "Observations", where an observation basically corresponds to a single file (a snapshot of the Sun, for instance), a Study groups one or more observations and a Campaign groups one or more studies and Observations. Understanding the hierarchical relations among the data, in this specific case quite different to other astronomical archives built by the team, is of crucial importance for the success of the archive. A snapshot of the central part of the SOHO UML model is attached here for reference.

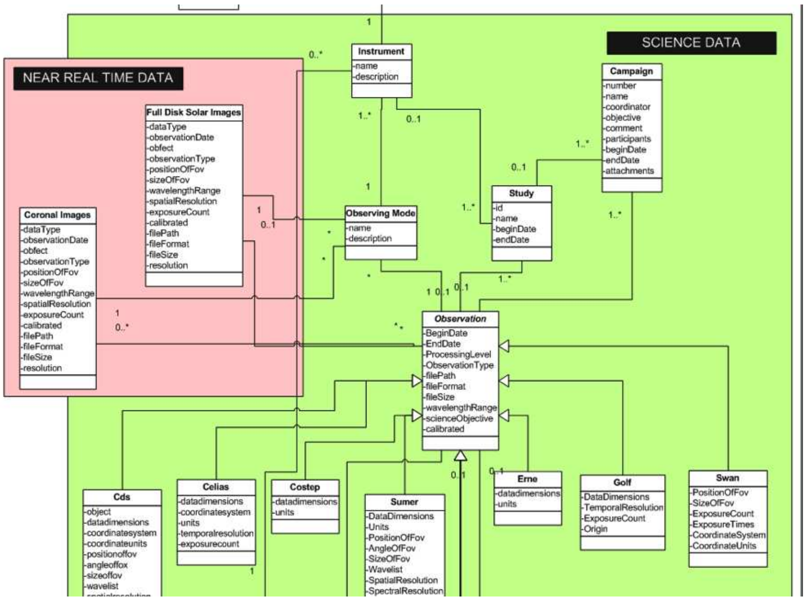


Figure 5. SOHO UML.

5.2. Dealing with *millions* of observations

One of the biggest problems that the SAT had to face when implementing the ABSI infrastructure over SOHO data was the fact that the number of observations in SOHO is in the range of the ***millions*** (around 3 million observations right now, and increasing daily). Joining relations within the relational database became a burden when dealing with these numbers, and therefore the SAT had to look for ways to improve the handling of searches in the DB.

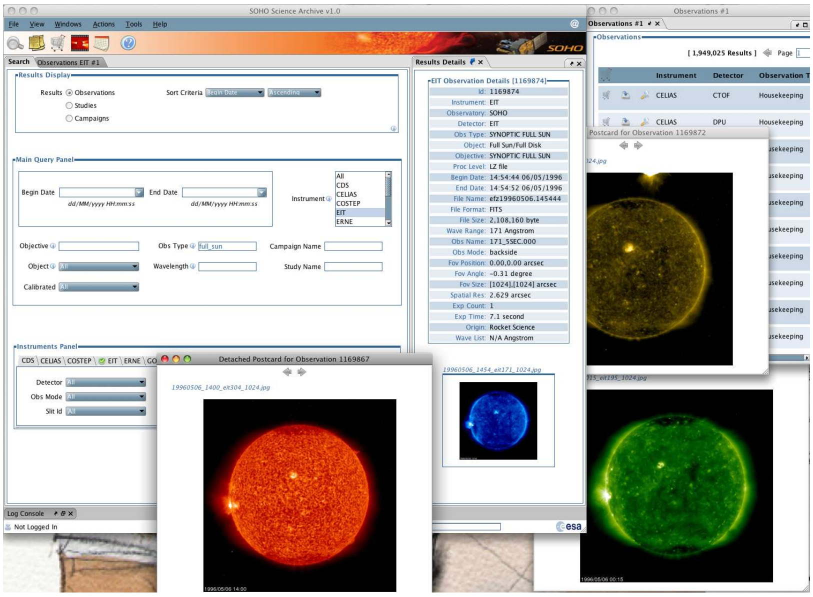
The solution came through the implementation of the Dijkstra algorithm (see

[Dijkstra 1959]). The algorithm, is a graph search algorithm that solves the single-source shortest path problem for a graph with nonnegative edge path costs, producing a shortest path tree. This algorithm is often used in routing. For a given source vertex (node) in the graph, the algorithm finds the path with lowest cost (i.e. the shortest path) between that vertex and every other vertex.. The algorithm was applied when executing relational database joins, resulting in an important improvement in seacrhes response times. Current searches for the whole archive's contents take now less than one second to give a paginated result of million(s) of observations.

5.3. The SOHO Science Archive Graphical User Interface

The ABSI framework provides a standard Model-View-Controller paradigm User Interface, that is adapted to the specific needs of each of the archives. In the case of SOHO, the time domain is one of the most important parameters. Also, the existence of many instruments (12 in total) makes the user interface implementation quite demanding.

Apart from the specific items required by the project, the user interface (as stated [Fernandez-Barreiros 2009]) implements state-of-the-art technology that allow seamless freedom in the interaction of the user with the interface. Tabs can be detached or reattached at will, details panels can be expanded, overlayed to allow for comparison of, e.g., different filter images, etc.



Results Display

Results ● Observations ○ Studies ○ Campaigns

Sort Criteria: Begin Date, Ascending

Main Query Panel

Begin Date: dd/MM/yyyy HH:mm:ss

End Date: dd/MM/yyyy HH:mm:ss

Instrument: All, CDS, CELIAS, COSTEP, EIT, ERNE

Objective:

Obs Type: full_sun

Campaign Name:

Object: All

Wavelength:

Study Name:

Calibrated: All

Instruments Panel

CDS | CELIAS | COSTEP | ✓ EIT | ERNE | GO

Detector: All

Obs Mode: All

Slit Id: All

Detached Postcard for Observation 1169867

19960506_1400_eit304_1024.jpg

1996/05/06 14:00

EIT Observation Details [1169874]

Id:	1169874
Instrument:	EIT
Observatory:	SOHO
Detector:	EIT
Obs Type:	SYNOPTIC FULL SUN
Object:	Full Sun/Full Disk
Objective:	SYNOPTIC FULL SUN
Proc Level:	LZ file
Begin Date:	14:54:44 06/05/1996
End Date:	14:54:52 06/05/1996
File Name:	efz19960506.145444
File Format:	FITS
File Size:	2,108,160 byte
Wave Range:	171 Angstrom
Obs Name:	171_5SEC.000
Obs Mode:	backside
Fov Position:	0.00,0.00 arcsec
Fov Angle:	-0.31 degree
Fov Size:	[1024],[1024] arcsec
Spatial Res:	2.629 arcsec
Exp Count:	1
Exp Time:	7.1 second
Origin:	Rocket Science
Wave List:	N/A Angstrom

19960506_1454_eit171_1024.jpg

Observations

[1,949,025 Results] Page 1

Instrument	Detector	Observation T
CELIAS	CTOF	Housekeeping
CELIAS	DPU	Housekeeping

Postcard for Observation 1169872

024.jpg

915_eit195_1024.jpg

1996/05/06 00:15

Figure 6. SOHO Science Archive at work.

5.4. The "Time Animator"

Since time is such an important parameter in SOHO, an on-the-fly video creation tool to animate in seconds events that span for hours days or months was implemented in the ABSI Infrastructure from scratch. The system, called the ”Time Animator” is purely implemented in Java and does not require the use of plugins of any type. It can be adapted to any images to create on-the-fly animations. In the case of SOHO, choosing an instrument and start and end times, the systems makes a database search, locates images of the Sun that match the period and offers videos showing the Sun’s evolution in the selected range. A snapdhot of images from the Time Animator in action is shown.

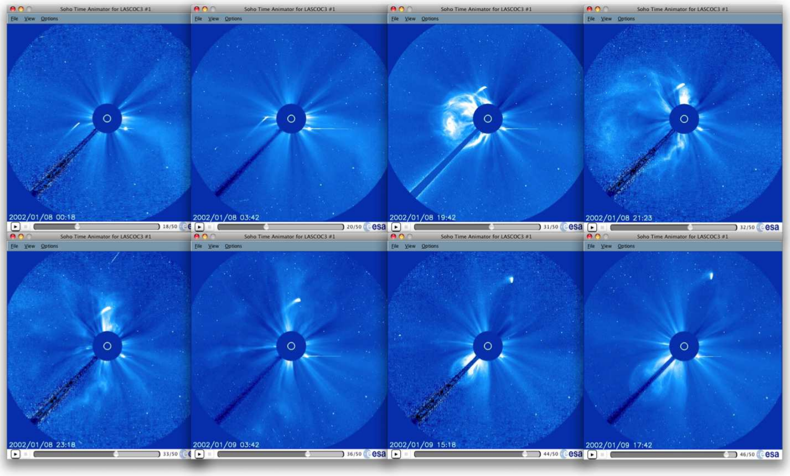


Figure 7. The SOHO Time Animator: a Comet just avoids the Sun under heavy Sun activity.

The SOHO Science Archive was released in September 2009 and can be accessed from the following URL:

`http://soho.esac.esa.int/data/archive/index_ssa.html`

6. The EXOSAT Science Archive case

The case of the EXOSAT Science Archive (EXSA henceforth) was of a completely different nature than the one of SOHO. While SOHO is still a live project, EXOSAT stopped working back in 1986. However, the recovery of the data from this mission was felt important for ESA, and the decission was taken to build a brand new archive recuperating the data from the only place where they were currently available at Heasarc.

EXOSAT was ESA’s first X-Ray mission, operational from May 1983 to April 1986. It operated in the 0.05 to 50 keV with two soft X-Ray imagers, one grating spectrometer, a medium-energy X-Ray collimated spectrometer, producing images, spectra and lightcurves.

In the post-operational phase, a data access software was written by ESA astronomers to keep the data. However, the HEASARC centre was born by that time and decision was made to trade data with HEASARC for software to translate orginial EXOSAT FOT files to emerging FITs standard. From that moment on, EXOSAT data were accessible from GSFC.

The possibility to implement the EXOSAT Science Archive from scratch and keep it under ESA premises using the new technology took place in 2007 and the EXOSAT Science Archive was finally made available in 2009.

6.1. The problem of handling "geometrical searches" within astronomical archives

One of the issues that the SAT was facing in the new overall ABSI framework was how to perform geometrical searches over a spherical surface. Traditionally, this problem had been undertaken by direct SQL queries to the database, making the seacrhes very slow, since indexing the coordinates and searching for squares within tables of thousands of entries is not very performant.
The solution to this problem came through the use of PgSphere (`http://pgsphere.projects.postgresql.org/index.html`), a module that implements spherical types in PostgreSQL, an open source database that, for these reasons, was chosen as the heart of the EXOSAT Science Archive metadata storage. PgSphere provides:

- input and output of data
- containing, overlapping, and other operators
- various input and converting functions and operators
- circumference and area of an object
- spherical transformation
- indexing of spherical data types
- several input and output formats

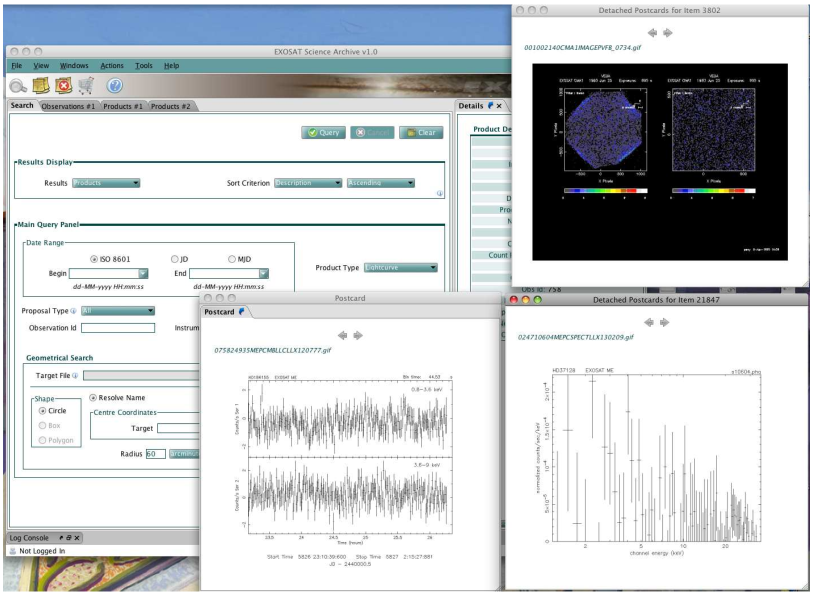
6.2. The EXOSAT Science Archive Graphical User Interface

As can be seen when accessing the EXOSAT Science Archive, the Graphical User Interface is very similar to that of SOHO. This gives and added value since users get very easily adapted to the use of different archives from cross-disciplines, and do not have to learn new technologies for different archives. The following is a snapshot of the EXOSAT Science Archive at work.

The EXOSAT Science Archive can be accessed at: `http://www.rssd.esa.int/index.php?project=EXOSAT&page=archive`

7. Conclusion

Flexibility, modularity, scalability and long term availability are paramount in designing durable strategies for Science Archive building. Software and Systems technology advances very fast making it specially difficult to foresee which technologies will be stable in ten years time. However, with proper analysis of the tools at hand, decisions can be taken and proper test beds implemented to check that the technologies chosen can at least last for a period of around other ten years.
The Science Archives and VO Team undertook an investigation of current state-of-the-art technologies and made decisions to implement a framework that would allow for better, cheaper, more reliable and expandable scientific archives, the Archives Building System Infrastructure (ABSI).
The ABSI has allowed the SAT to create two new archives from scratch. This


EXOSAT Science Archive v1.0

Results: Products
Sort Criterion: Description, Ascending

Date Range: ● ISO 8601 ○ JD ○ MJD
Begin dd-MM-yyyy HH:mm:ss
End dd-MM-yyyy HH:mm:ss
Product Type: Lightcurve
Proposal Type: All
Observation Id
Shape: ● Circle ○ Box ○ Polygon
● Resolve Name
Radius: 60 arcminut

Not Logged In

Postcard
075824935MEPCMBLLCLLX120777.gif

Detached Postcards for Item 3802
001002140CMA1IMAGEPVFB_0734.gif

Detached Postcards for Item 21847
024710604MEPCSPECTLLX130209.gif

Obs Id: 758

Figure 8. EXOSAT Science Archive at work.

Archives are already available to the community and are performing well, being efficient and scalable, even under demanding requirements. As an example, SOHO Science Archive currently has almost two million observations in the database and nearly four million files in the data repository (and growing) while providing fast and reliable service.

More new ESA space based missions' archives are being implemented within the new Infrastructure, c.f., Planck, Lisa Pathfinder, and some of the already existing ones will soon be re-engineered (Herschel, Planetary Science Archive, and others). SOHO and EXOSAT science archives have given birth to ESA's New Generation Science Archives.

References

Leon, I. 2009, in ASP Conf. Ser. YYY, ADASS XIX, ed. Y. Mizumoto, K.-I. Morita& M. Ohishi (San Francisco: ASP), [Leon 2009]

Fernandez-Barreiro, M. 2009, in ASP Conf. Ser. YYY, ADASS XIX, ed. Y. Mizumoto, K.-I. Morita& M. Ohishi (San Francisco: ASP), [Fernandez-Barreiro 2009]

Laruelo, A. 2009, in ASP Conf. Ser. YYY, ADASS XIX, ed. Y. Mizumoto, K.-I. Morita& M. Ohishi (San Francisco: ASP), [Laruelo 2009]

Arviset, C. 2006, in ASP Conf. Ser. 376, ADASS XVI, ed. R. A. Shaw, F. Hill, & D. J. Bell (San Francisco: ASP), [Arviset 2006]

Dijkstra, E. W. Numerische Mathematlk l, 269 - 27 I (l 959)A. 2009, in ASP Conf. Ser. YYY, ADASS XIX, ed. Y. Mizumoto, K.-I. Morita& M. Ohishi (San Francisco: ASP), [Dijkstra 1959]