Archives Building System Infrastructure: Re-engineering ESA Space Based Missions' Archives

P. Osuna [(1)], J. Salgado [(1)], I. Ortiz [(1)], C. Arviset [(1)]

[(1)] ESA-ESAC

PoBox 78, 28691 Villanueva de la Cañada, Madrid, Spain

EMail: Pedro.Osuna@sciops.esa.int

ABSTRACT

The Science Archives and VO Team (SAT) located at the European Space Astronomy Centre (ESAC) near Madrid, Spain, is in charge of building the Scientific Archives of all ESA space based missions. The team started in 1997 with the building of the Infrared Space Observatory data Archive (IDA) and, since then, has undertaken the building of several other scientific archives that span from high-energy astronomy mission like XMM-Newton to Planetary missions like Mars Express or Venus Express.

While the SAT was pioneer in the usage of certain technologies back in the mid nineties (the ISO archive is probably the first ever astronomical archive written completely in Java) some of the ad-hoc solutions that had to be implemented from scratch are nowadays offered as standard solutions within certain environments.

After an intensive exercise of self-auditory, the Science Archives and VO team undertook the task of re-engineering of the infrastructure for building scientific archives. This resulted in a new concept called the Archives Building System Infrastructure (ABSI). Building the SOHO Science Archive has recently proved this concept.

This article will describe the lessons learned during the years of science archives building and the solutions found to the identified problems that conform the ABSI. The article will also describe the basics of the SOHO Science Archive.

Keywords: science archive, SOHO, ABSI, archive, building, system, infrastructure, astronomy, VO

INTRODUCTION

When the Science Archives and VO Team (SAT) started building archives back in 1997, the overall world IT picture was quite different to the one we can see now. For instance, firewalls did not exist; data were held (and distributed) in CD-ROMs of 640 Kbytes storage capacity; big jukeboxes were holding the data and giving access to them through the Archive system; Java technology was starting; and a long etcetera of remarkable points that today sound as if they belonged to the Pleistocene.

The evolution of the technologies poses a challenge in the design and maintenance of software engineering systems, especially when the technologies used were in an infant state at the beginning of the development cycle. Therefore, some of the new introduced paradigms, like firewalls blocking ports, etc., posed a serious problem for the science archives developed for ESA missions.

Those issues were being solved whenever they were appearing, as the overall architecture for the archives had proven to be a correct and robust one, therefore allowing for changes to be introduced in the system without breaking the overall architecture.

Thus, problems were being overcome as years were passing, and the SAT managed to release science archives for the following ESA missions: XMM-Newton, Integral, Planetary missions (Mars Express, Venus Express, Rosetta, Giotto, soon SMART-1), and Herschel (very recently).

However, problems were becoming more and more costly in terms of low-level architecture modifications and implementation and maintenance. Thus, the team decided to make a study on the

different issue affecting the systems used for building archives, and started an investigation towards the creation of an archives building system infrastructure that would use state of the art technology and would allow the building of scientific archives for ESA missions in an easy and powerful way.

ARCHIVES BUILDING SYSTEM INFRASTRUCTURE

The process towards finding the right infrastructure was done in several stages:

1. Investigation of the issues affecting the current archives
2. Research on current trends and technologies in the fields of database access, user interface design, etc.
3. Creation of an internal System requirements Document that the new Architecture should obey
4. Design of the Archives Building System Infrastructure
5. Implementation of the different Modules, Components, etc. of the general ABSI architecture
6. Application to building a Science Archive using the new technology: the SOHO case.

Investigation of current issues

A compilation of the current issues affecting our archives was done. The following are examples extracted from the document that was written with the current facts, status and possible solutions for the different issues found:

[...]

User Interface Architecture

Current facts and status

Communication with server is done through a home-made Java RPC (Java serialized objects).

Transport is done over TCP-IP in a compressed streaming mode.

Default Ports are 5433, 5453 and 80 (for users under a firewall).

Remarks and possible problems

Compression is disabled on port 80, so transfer is heavier for users under a firewall.

TCP-IP is a low level protocol, while HTTP supports tunneling, encryption, connection timeouts, etc.

Our Java-RPC implementation is far for being standard (RMI more suitable)

Possible solutions

Use standard HTTP as transport protocol and benefit from its features.

Use standard communication protocols (RMI, SOAP, XML-RPC, ...etc).

RMI and XML-RPC are best suitable (see presentation XML-RPC vs. RMI-SOAP).

XML as communication, HTTP as transport implies:

Flexibility: different clients in different languages, same back-end.

Scalability: it can be plugged into Java application servers easily.

Interoperability: HTTP and XML are consolidated standards opened to interoperability.

[...]

Research on current trends and technologies for the new infrastructure

During the analysis of the problems affecting our archives, we also researched on possible new technologies that could alleviate the issues we were identifying. The resulting document gave some recommendations. Examples of them are shown in this extract from the produced document:

[...]

FRAMEWORK CONSIDERATIONS AND RECOMMENDATIONS FOR METADATA ACCESS

Discussions have been held within the team on the best approach for the Persistence Layer and its access mechanisms.

After some research, and taking into account the experiences of the members of the team, it was decided to research with a bit more detail the following two technologies:

- *Hibernate (http://www.hibernate.org/)*
- *Ibatis (http://ibatis.apache.org/)*

In order to clarify ideas and be able to implement using the proposed technologies, a Use Case was designed and two different and independent implementations were done making use of the discussed technologies.

Recommendation

*After looking at the details of the aforementioned implementations, the **final recommendation** was to use the **Hibernate** Technology. Some of the reasons given by the members of the SAT are given without a specific order below.*

- *Hibernate has got more functionality and flexibility allowing all types of queries to be defined easily. On the downside it looks bigger than iBatis. However, iBatis might be short for what we need. iBatis looks like the perfect option for other applications like administration tools, etc.*
- *As we start from scratch, the persistence data model on the DB can be driven from the project data model. This is one of the main reasons where Hibernate is recommended. It allows a lot of flexibility on the data model design.*
- *It is very powerful, basically it can perform any type of query, as it is done navigating through objects rather than having to write the SQL.*
- *Once it is setup, inserts and updates are as simple as calling save() on the Java object, without any explicit JDBC call. This can be very useful for ingestion.*
- *Smoothly integrated with Spring.*
- *Very big community support. All the comments from users after using Hibernate are positive, hardly any complain.*
- *we should have one single DB access layer (for archive, stats, admin, ingestion, ...) and hence, need one as flexible/powerful as possible. We should also use SUN's APIs everywhere possible, and in this case the new Java Persistence API. For this, mainly, I think Hibernate is the best solution (best JPA implementation at the moment), used through SUN's API, so that we can switch to any other if needed (TopLink, Kodo, GF JPA, OpenJPA, ...).*
- *JPA (Hibernate) Pros : standard SUN API, very flexible, full DB encapsulation, various implementations, Hibernate best impl at the moment. Cons : heavyweight, steep learning curve*
- *iBatis Pros : lightweight, easy to learn (if SQL already known). Cons : non-standard, some flexibility limits, no encapsulation (only SQL queries mapping), unique implementation.*
- *Hibernate is quite similar in concept to the current BL, so adapting to it might seem natural to our systems.*

[...]

ABSI System Requirements, Design and Implementation guidelines

After the investigation of the problems found in our archives, and the possible solutions to them, the SAT created an internal document describing the requirements needed for an infrastructure that would be able to adapt easily to the User Requirements to build powerful science archives.

Following the system requirements document, the top level design was created for the different subsystems, and the implementation of the different modules, components, etc was initiated. Excerpts of the documents describing the terms and definitions of the new architecture are given here for reference:

[...]

Terms and Definitions

Important terms and definitions for the ABSI are given in what follows. They form the basis of the subsystems that the developer will count on when making use of the ABSI Infrastructure to build a specific new Science Archive "from scratch".

ABSI

Stands for "Archives Building System Infrastructure". It describes the Infrastructure that should be used when creating a new Archive from Scratch. This doc describes the ABSI Top Level architecture plus first level of decomposition to clarify the concepts. It also gives code storage and naming conventions that must be followed when writing the software.

ABSI Archive Sample

Implements a set of Use Cases providing a certain number of functionalities. It is composed of:

1. *BuildingBlocks*
2. *GlueCode*

Building Block

It is a logical definition. It can take form as one of the following:

1. *Module*
2. *ComponentSample*

Module

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. Its code is normally not modified. If modification to it is needed, it shall be agreed within the SAT, and new version of Module shall be uploaded to repository.

ComponentSample

Wraps-up sample code implementing certain functionality. It may contain Modules and/or GlueCode. It is subject to code modifications to be adapted to specific needs, hence the name "sample". However, these modifications should not be substantial.

[...]

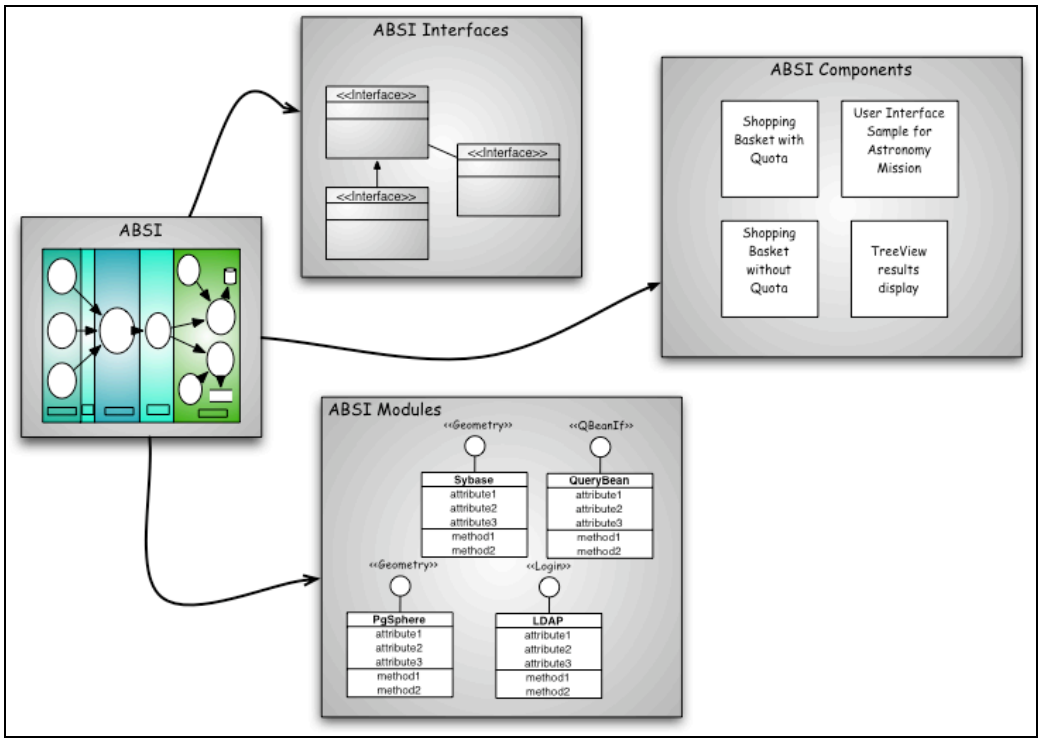


Figure 1 graphical interpretations of terms and definitions

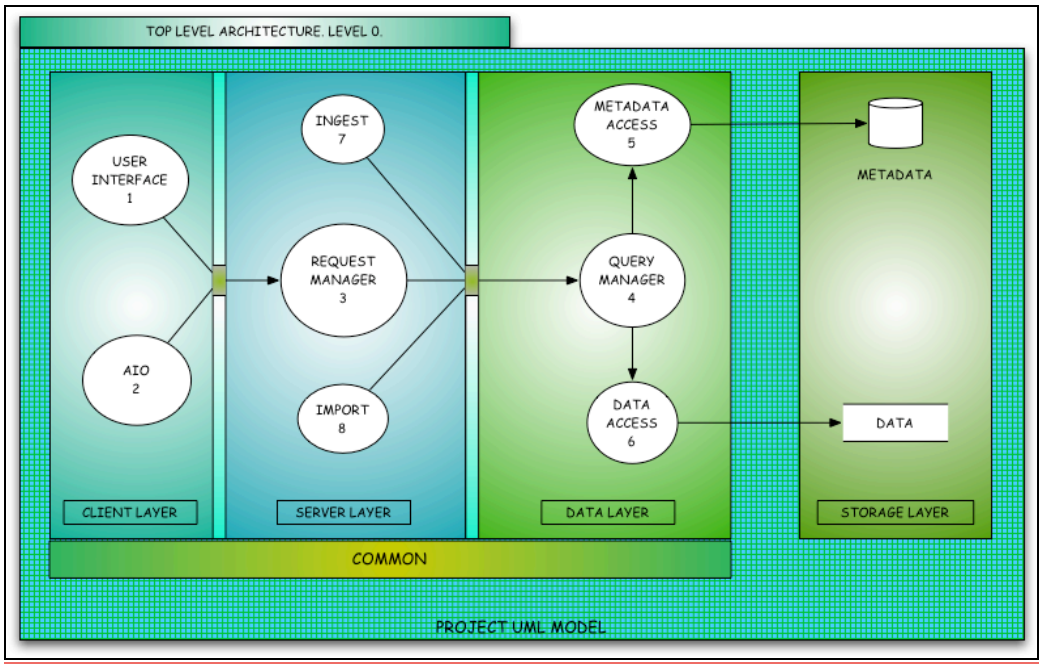


Figure 2 ABSI Top level architecture diagram

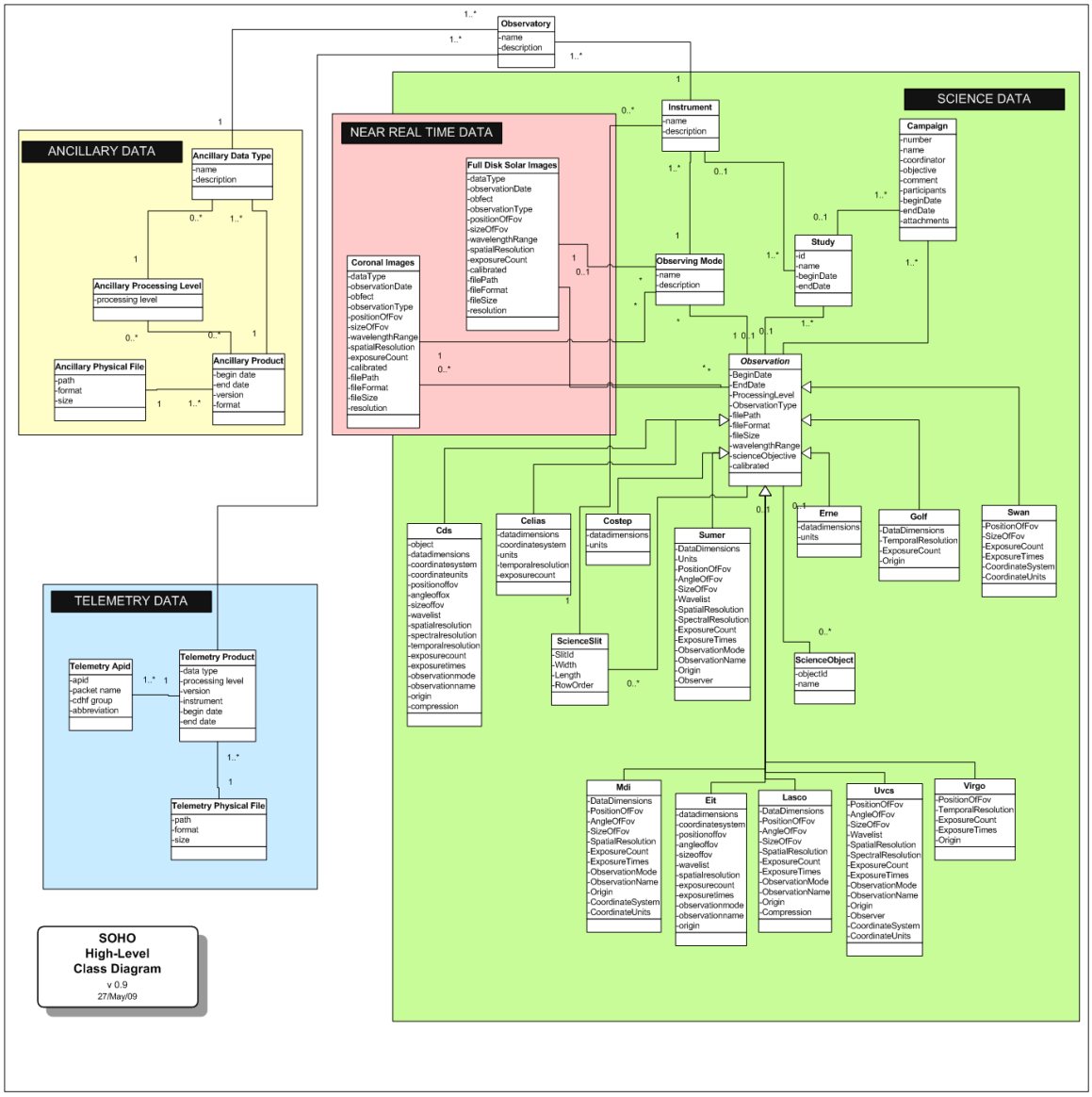
USING THE INFRASTRUCTURE TO BUILD THE SOHO SCIENCE ARCHIVE

The "Solar and Heliospheric Observatory" was chosen as the first science archive to be built under the new Archives Building system Infrastructure.

With requirements gathered from the community and converted into a User Requirements Document by Luis Sánchez (SOHO Archive Scientist), the archive was implemented with the new technology and released to the public in October 2009 (see ESA news at http://sci.esa.int/science-e/www/object/index.cfm?fobjectid=45900). The SOHO science archive can be accessed from http://soho.esac.esa.int/data/archive/index_ssa.html.

Basic Building Blocks for the SOHO Science Archive

The first step in the building of the new SOHO science archive using the ABSI paradigm was the creation of the logical model for the project. The UML created for that purpose is attached:



From the UML model, the metadata model for the Database and the repository model for the storage of the physical data are built. At this stage, the specific items for the processes of ingestion of metadata into the database and of import of the data into the repository are also built.

In the case of SOHO, the extraction of metadata had to be done over different file formats: ASCII, CDF, FITs. Also, the big amount of observations made with SOHO (above two million up to date) posed a problem in the handling of the insertion of the metadata and data to the database and file system respectively.

The following two images show amounts of metadata and data ingested into the SOHO archive DB and repository:

Data Type	Instrument	Data Format	Number of files
REALTIME	EIT	JPG	883356
REALTIME	LASCO	JPG	734085
REALTIME	MDI	JPG	20434
SCIENCE	CDS	FITS	285641
SCIENCE	EIT	FITS	443241
SCIENCE	GOLF	FITS	3817
SCIENCE	LASCO	FITS	532719
SCIENCE	MDI	FITS	56004
SCIENCE	SUMER	FITS	116418
SCIENCE	SWAN	FITS	7214
SCIENCE	UVCS	FITS	85336
SCIENCE	VIRGO	FITS	19685

Figure 3 File metadata ingestion into the DB

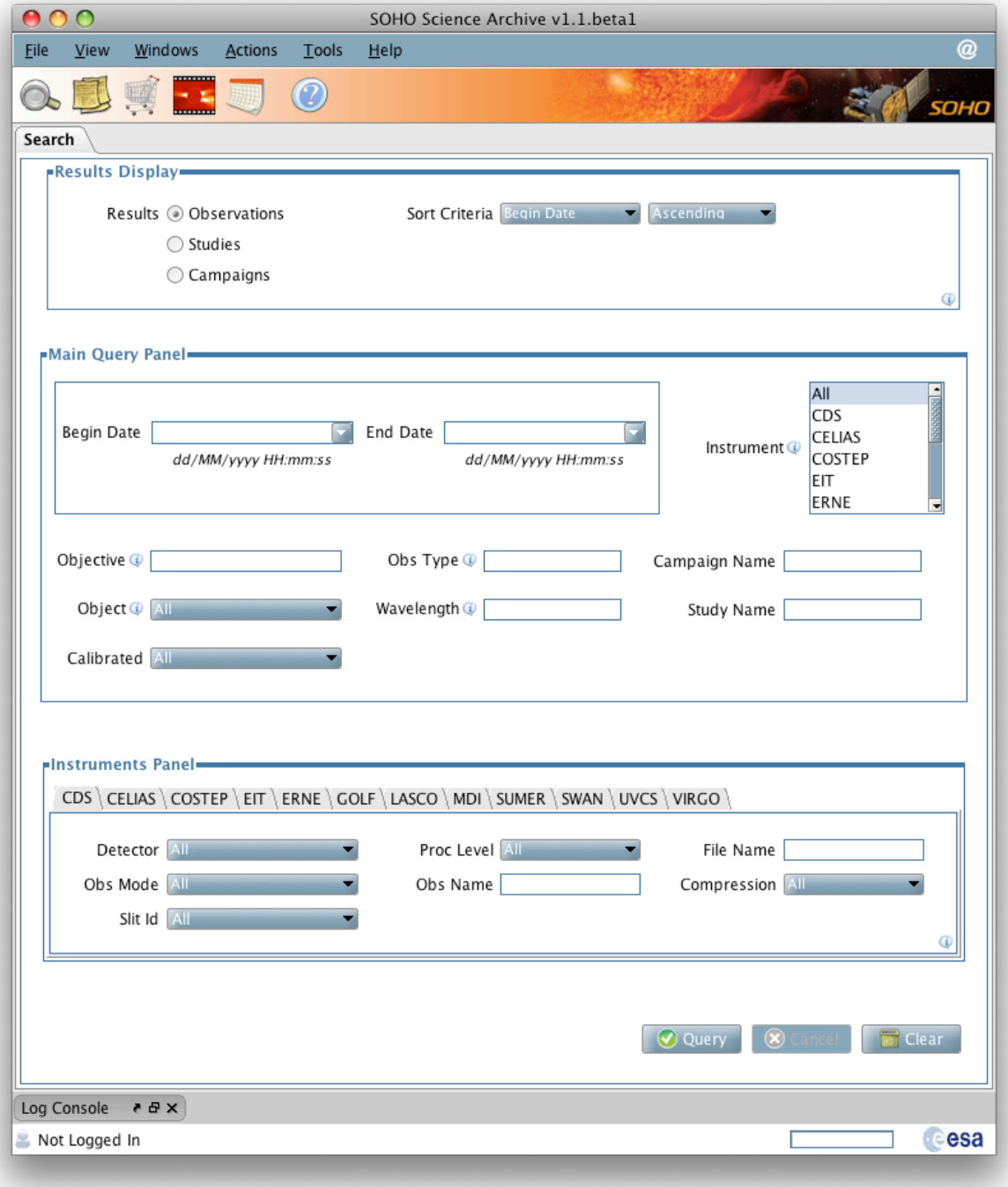
Instrument	File type	Number of files
CDS	FITS	285641
EIT	FITS	443241
GOLF	FITS	3817
LASCO	FITS	532719
MDI	FITS	56004
SUMER	FITS	27897
VIRGO	FITS	19679
LASCO	JPG	734085
EIT	JPG	883336
MDI	JPG	20434

Figure 4 File data import to repository

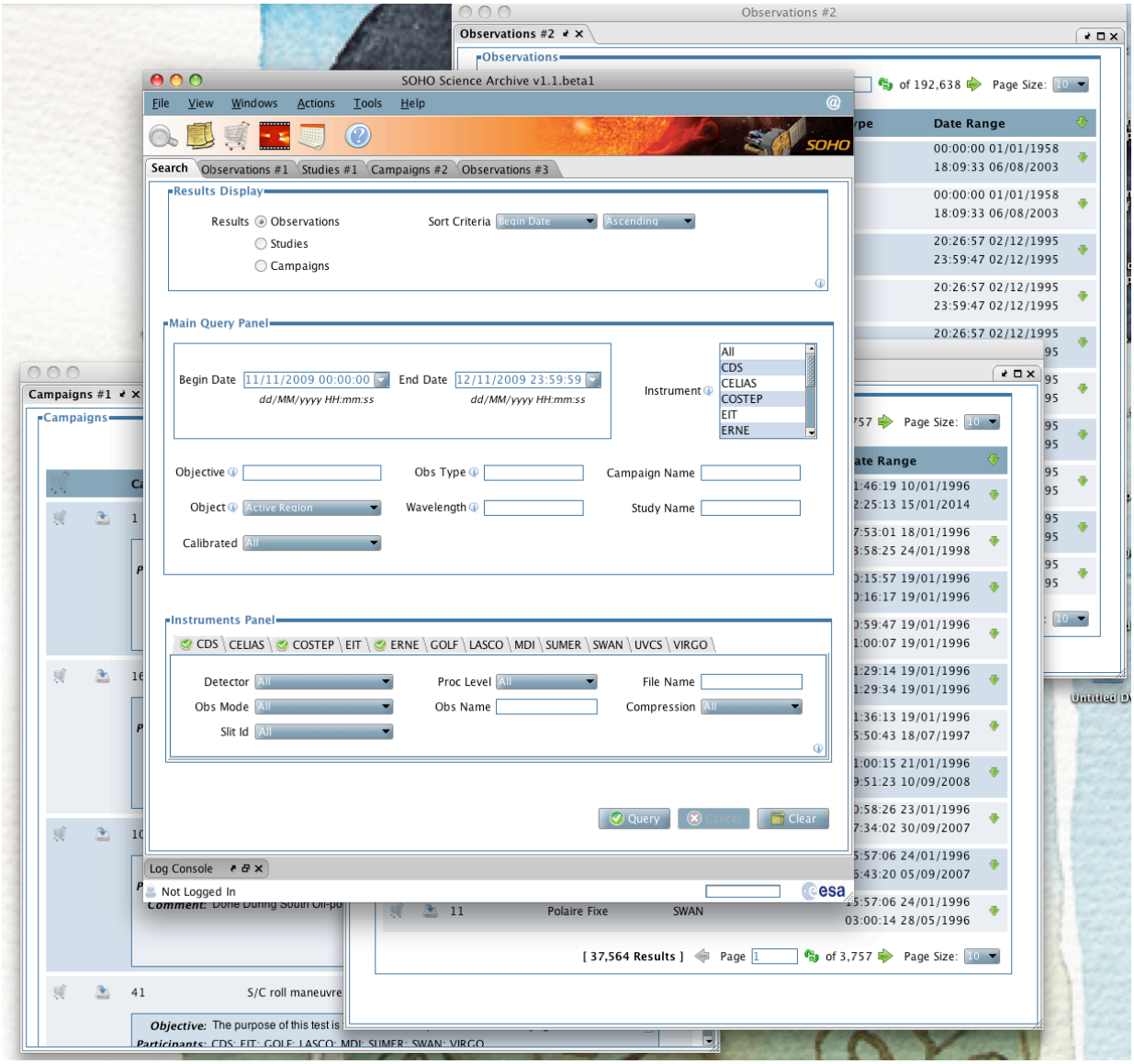
Building of the User Interface for SOHO Science Archive

The User Interface was designed to be easy to handle by laymen, but powerful enough for expert users. The flexible and modular design of the ABSI infrastructure together with the functionalities offered by open source APIs, make the SOHO User Interface a powerful tool for science research.

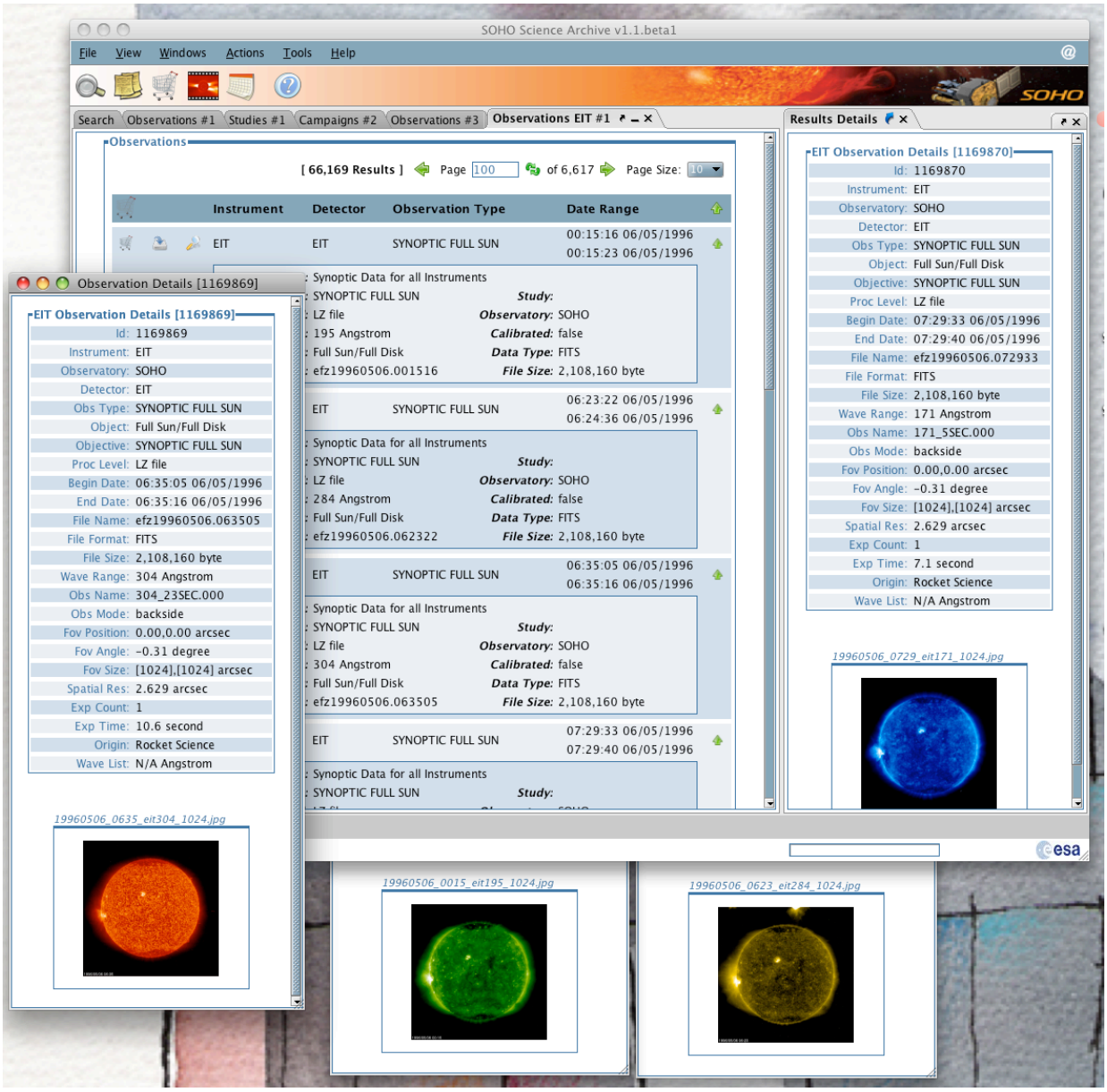
The following is an image of the SOHO User Interface Search panel, where the different selection parameters can be seen for the overall mission and for specific instruments:



The results from the searches from the User Interface can be detached for easy handling:

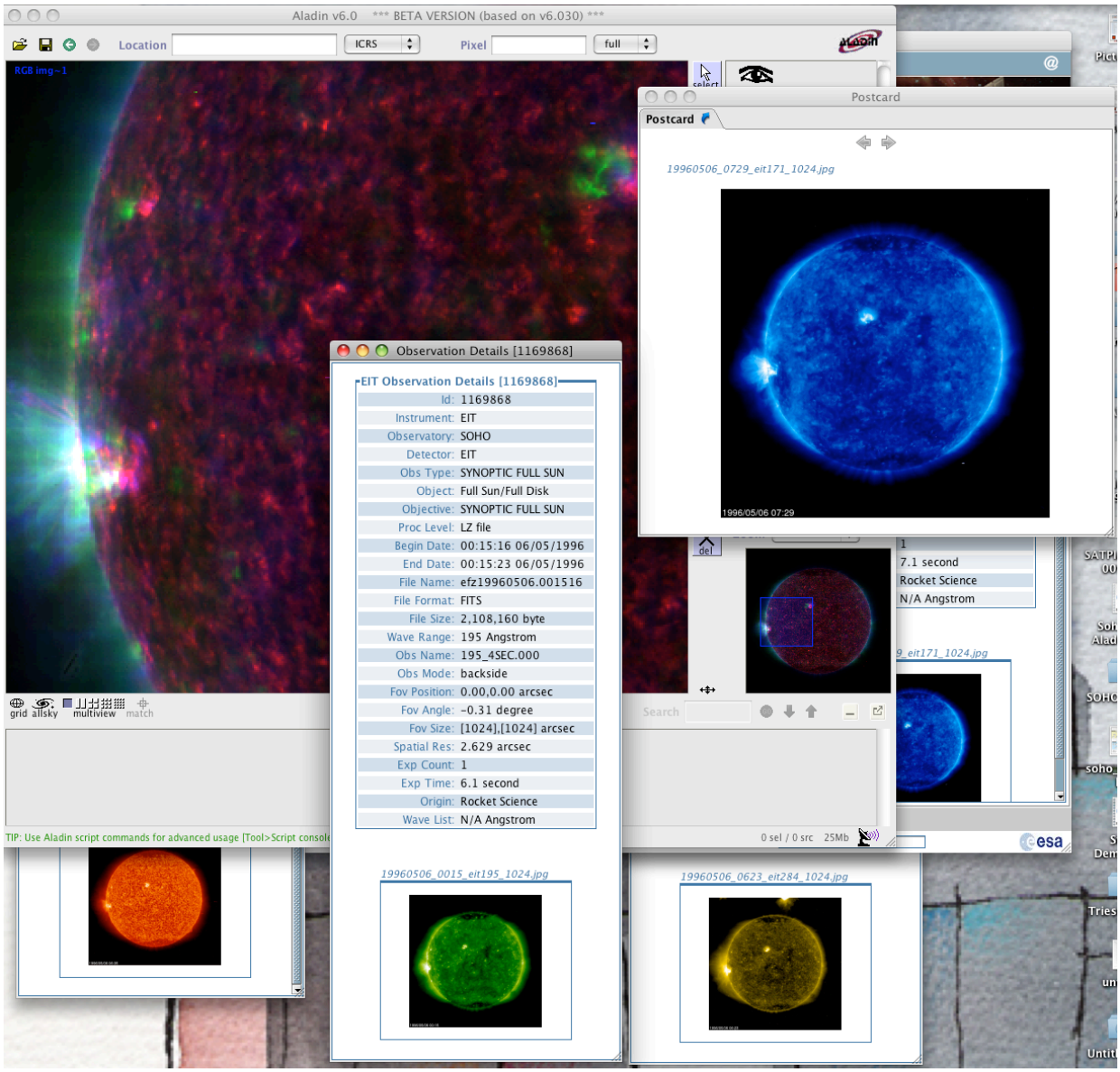


Detailed panels allow the user to get more specific information for one observation. Details panels can also be detached:


SOHO Science Archive v1.1.beta1
File View Windows Actions Tools Help
Search Observations #1 Studies #1 Campaigns #2 Observations #3 Observations EIT #1
Observations
[66,169 Results] Page 100 of 6,617 Page Size: 10
Instrument Detector Observation Type Date Range
EIT EIT SYNOPTIC FULL SUN 00:15:16 06/05/1996 00:15:23 06/05/1996
Synoptic Data for all Instruments
SYNOPTIC FULL SUN Study:
LZ file Observatory: SOHO
195 Angstrom Calibrated: false
Full Sun/Full Disk Data Type: FITS
efz19960506.001516 File Size: 2,108,160 byte
EIT SYNOPTIC FULL SUN 06:23:22 06/05/1996 06:24:36 06/05/1996
Synoptic Data for all Instruments
SYNOPTIC FULL SUN Study:
LZ file Observatory: SOHO
284 Angstrom Calibrated: false
Full Sun/Full Disk Data Type: FITS
efz19960506.062322 File Size: 2,108,160 byte
EIT SYNOPTIC FULL SUN 06:35:05 06/05/1996 06:35:16 06/05/1996
Synoptic Data for all Instruments
SYNOPTIC FULL SUN Study:
LZ file Observatory: SOHO
304 Angstrom Calibrated: false
Full Sun/Full Disk Data Type: FITS
efz19960506.063505 File Size: 2,108,160 byte
EIT SYNOPTIC FULL SUN 07:29:33 06/05/1996 07:29:40 06/05/1996
Synoptic Data for all Instruments
SYNOPTIC FULL SUN Study:
Observation Details [1169869]
EIT Observation Details [1169869]
Id: 1169869
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: 06:35:05 06/05/1996
End Date: 06:35:16 06/05/1996
File Name: efz19960506.063505
File Format: FITS
File Size: 2,108,160 byte
Wave Range: 304 Angstrom
Obs Name: 304_23SEC.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: 10.6 second
Origin: Rocket Science
Wave List: N/A Angstrom
19960506_0635_eit304_1024.jpg
Results Details
EIT Observation Details [1169870]
Id: 1169870
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: 07:29:33 06/05/1996
End Date: 07:29:40 06/05/1996
File Name: efz19960506.072933
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_0729_eit171_1024.jpg
19960506_0015_eit195_1024.jpg
19960506_0623_eit284_1024.jpg
esa

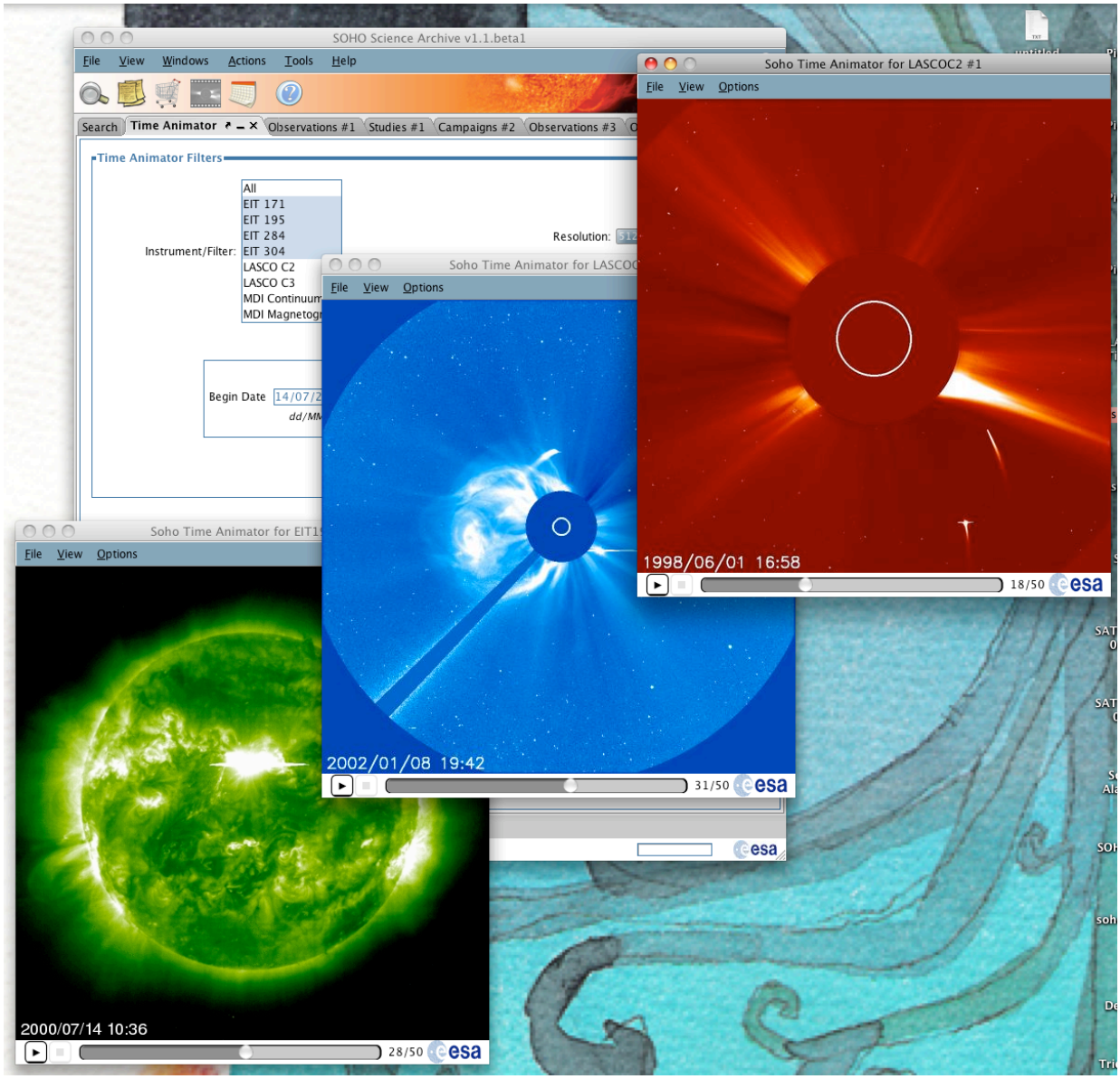
Interoperability with other applications through international protocols

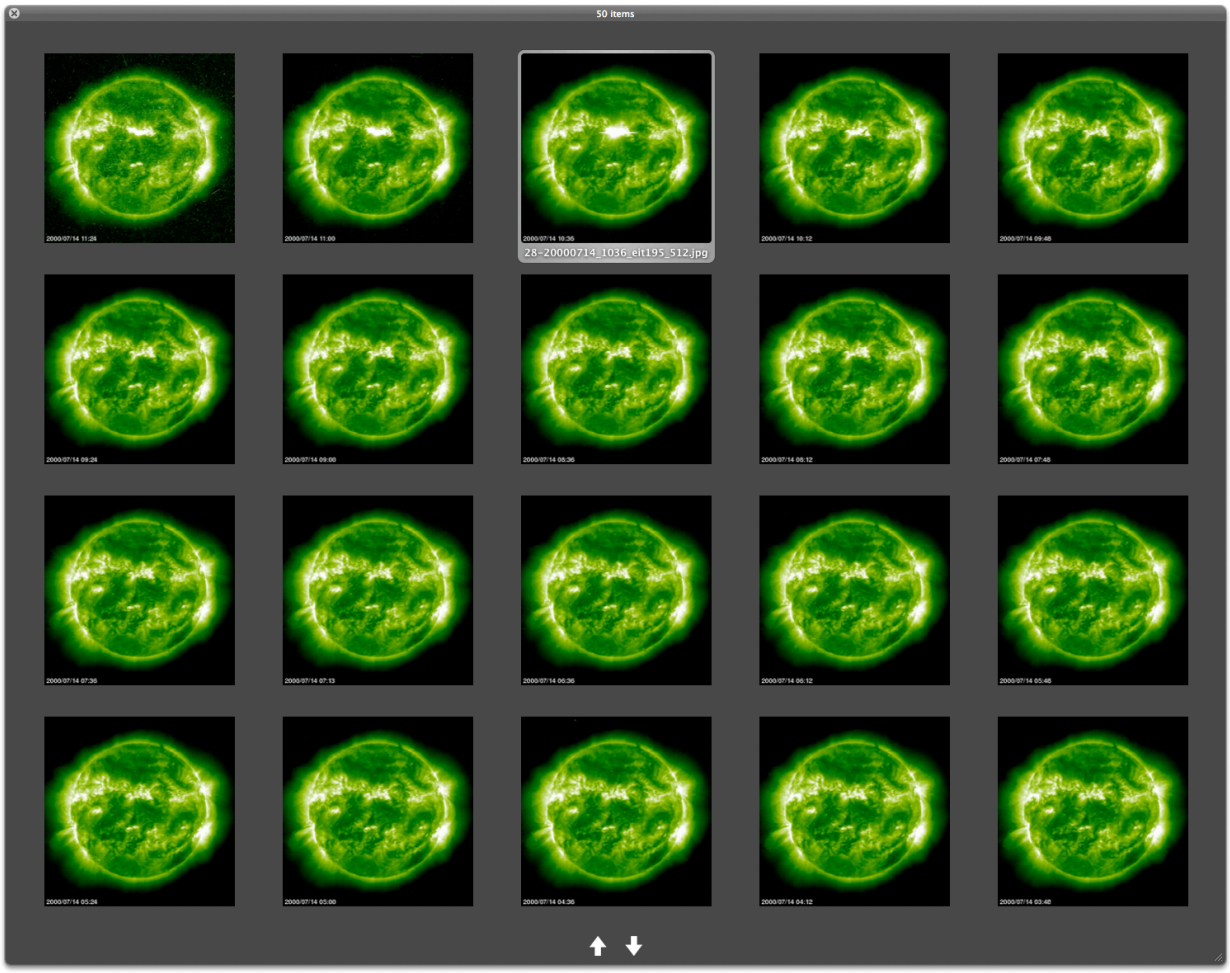
The ABSI contains a module, which implements the Simple Application Messaging Protocol (SAMP), a protocol defined by the International Virtual Observatory Alliance (IVOA) for interaction between applications. In this way, a simple right click in an image of the Sun on the SOHO Science Archive allows to send the image to an image manipulation tool such as the CDS Aladin, where RGB images can be constructed in minutes, enhancing the functionality of the archive and its scientific usage possibilities:



On the fly video creation capabilities in SOHO Science Archive

Within the ABSI, a module for image animation was created from scratch to be able to animate images from the Sun. Completely platform independent and not requiring any external plug-in, the image animator provides the user with the ability to produce videos of relevant events happening on the sun within the click of a button:




50 Items
28-20000714_1036_eit195_512.jpg

CONCLUSION

Software system technologies evolve with time, and the recent years could be considered "inflationary" with respect to how much the technology has advanced. Design of software systems is always trying to be done flexible enough to adapt to possible future changes, but when technologies are in their initial state, it is very difficult to foresee the evolution.

The Science Archives and VO Team has undertaken a self-assessment exercise to identify where the current ESA science archives are finding issues related precisely to the advance in the different technology areas. As a result of this exercise, a new infrastructure has been created that allows the creation of more flexible, modular, powerful and performant science archives.

The first science archive created using this new technology is the SOHO Science Archive, released in October 2009. Feedback from the community has been very positive. This technology has also been applied to the creation of the archive for a historical mission, EXOSAT, and will be used for the creation of the PLA archive, the Herschel archive and others.

REFERENCES

[1] Science Archives and VO Team web pages http://www.sciops.esa.int/index.php?project=SAT

[2] International Virtual Observatory Alliance (IVOA) http://ivoa.net

[3] SOHO http://soho.esac.esa.int/

[4] SOHO Science Archive http://soho.esac.esa.int/data/archive/index_ssa.html

Acknowledgements: the authors represent the whole Science Archives and VO Team, whose work has made the ABSI project a reality. The SAT members are the following (in alphabetical order):

Deborah Baines, Isa Barbarisi, Javier Castellanos, Neil Cheek, Hugo Costa, Nico Fajerzstejn, Monica Fernández Barreiros, Juan González, Andrea Laruelo, Ignacio León, Aurelien Stébè, Daniel Tapiador.

The authors also wish to acknowledge the SOHO Archive Scientist, Luis Sánchez.