Archives System Building Infrastructure: Re-engineering ESA's space based missions' archives

Pedro Osuna

Science Archives and Computer Support Engineering Unit (SRE-OE)
Science Operations Department (SRE-O)
Science and Robotic Exploration Directorate (SRE)

Introduction

3. Lessons learnt

- Examples of working systems and the lessons learnt that can be derived from them
- Lessons learnt from archive migration to new technologies
- Implications of new technologies for engineering processes, data storage, operations costs and system performance
- Advantages and difficulties of building interoperables services
- Common or re-usable systems for archives building

Presentation Overview

ESAC Archives History

ESAC Archives evolution

The “ABSI” concept

The SOHO Science Archive case

The EXOSAT Science Archive

Conclusion

Presentation Overview

ESAC Archives History

ESAC Archives evolution

The “ABSI” concept

The SOHO Science Archive case

The EXOSAT Science Archive

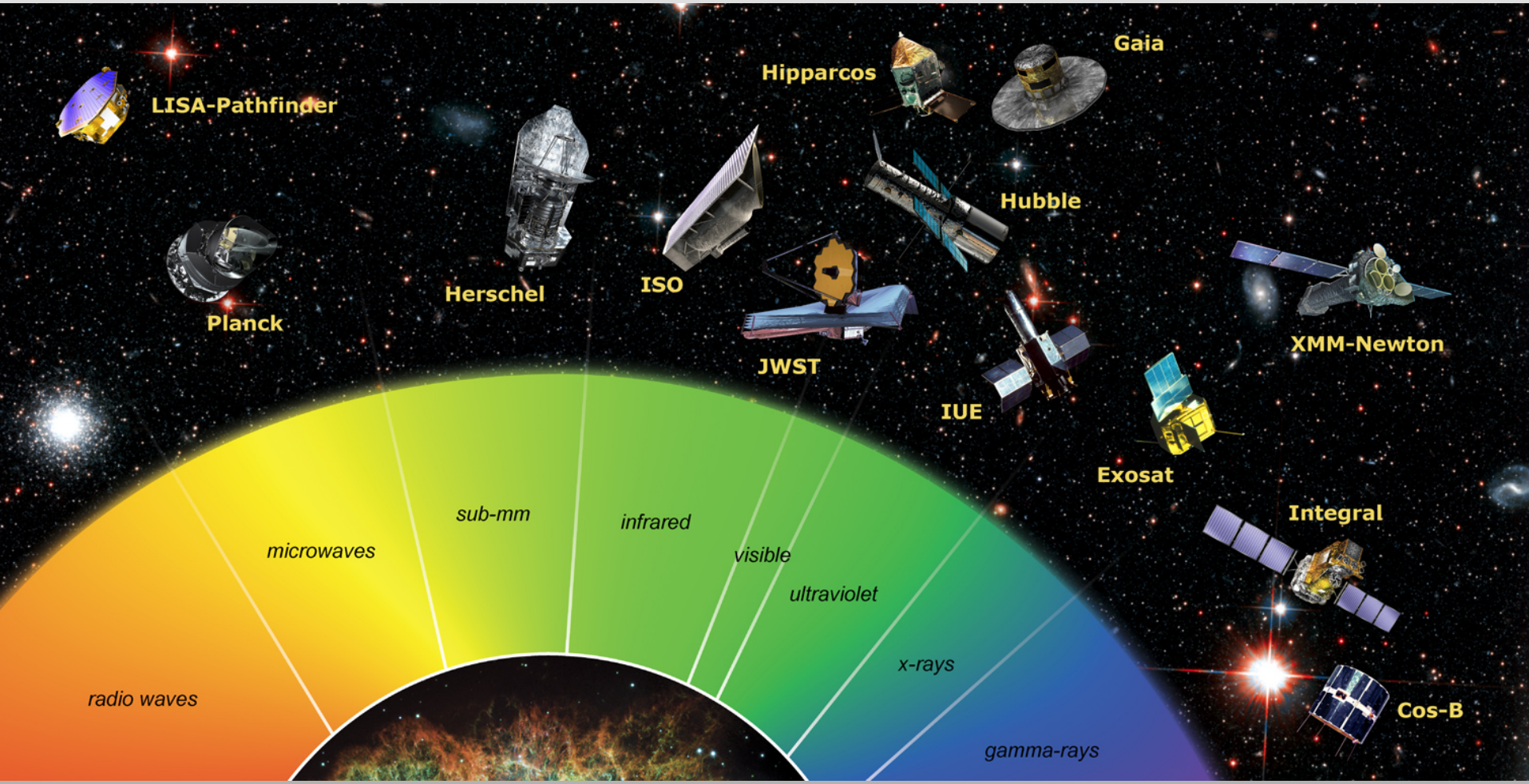
Conclusion

European Space Astronomy Centre (ESAC)

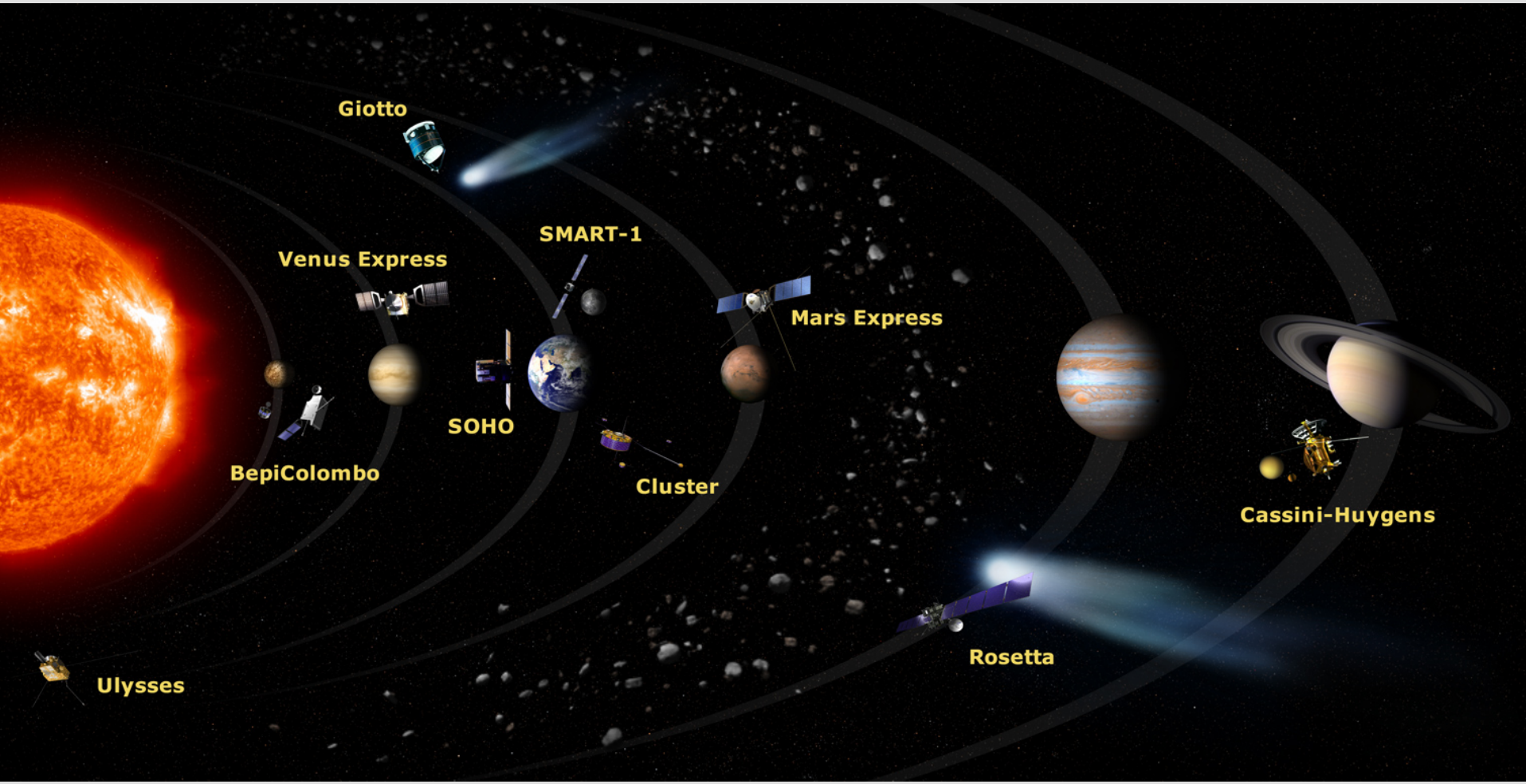
ESAC default location for ESA's:

- Science operations:
  - long history with astronomical missions
  - expanding with solar system missions
- Science archives:
  - Astronomy
  - Planetary
- ESA VO activities:
  - ESAC is the European VO node for space-based astronomy

Space Based missions on Astronomy and Fundamental Physics

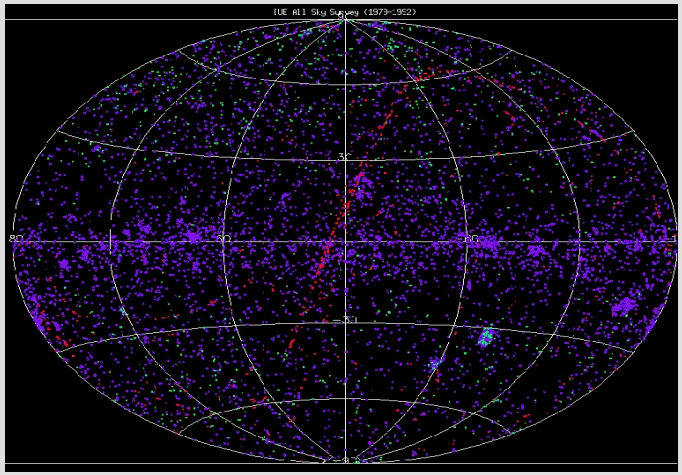


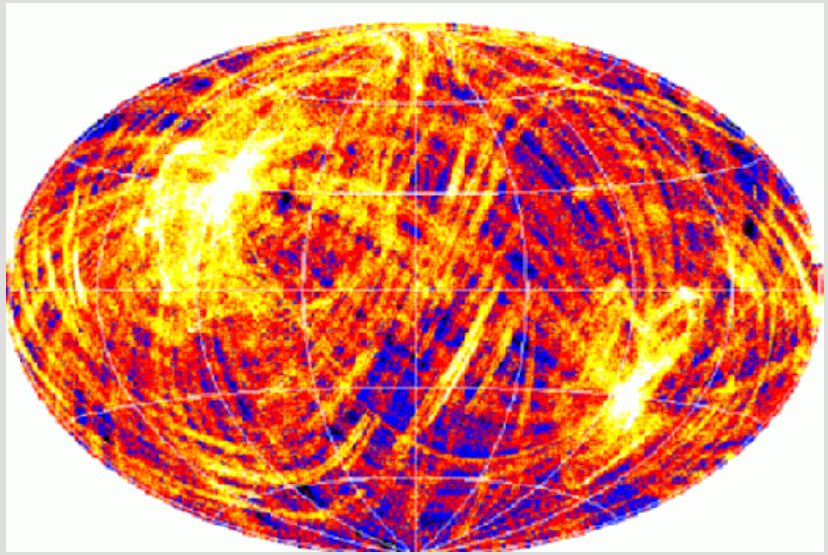
Solar System Space based missions



Before 1996: archives not at ESA

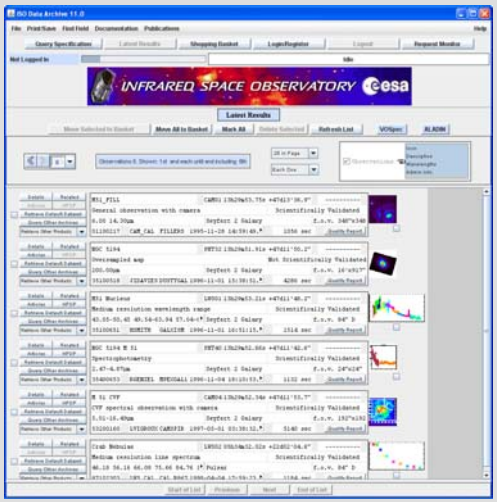
The International Ultraviolet Explorer (IUE) (1978-1996)
The European Space Agency X-RAY Observatory (EXOSAT)

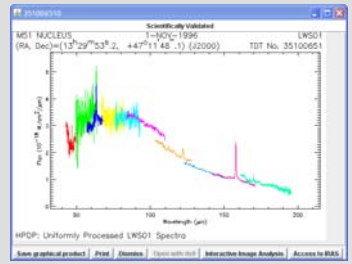


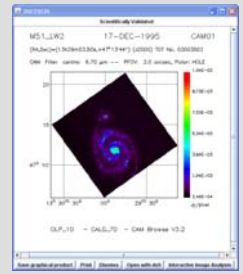


Starting a new era in archiving: The Infrared Space Observatory (ISO) Archive

- Available since December 1998:
  - http://iso.esac.esa.int/ida/
- Developed and released for the post mission phase
- Active development up to 2006, low maintenance now
- Content stable since 2002
- Around 400 GB of data (FITS) on hard disk of all levels of data (standard processing + high level data products)

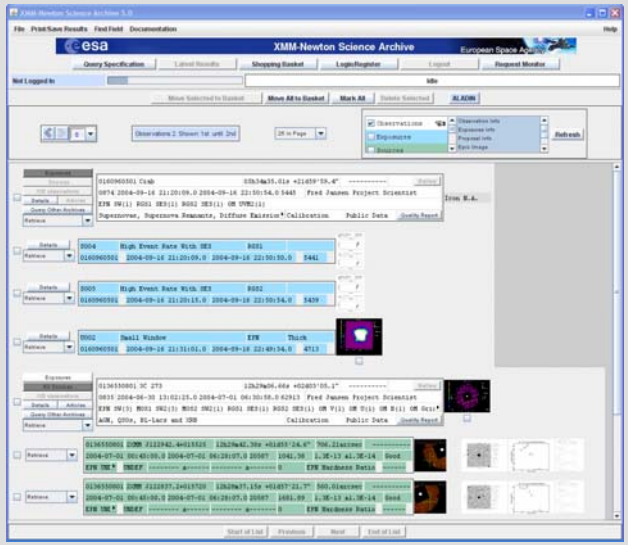


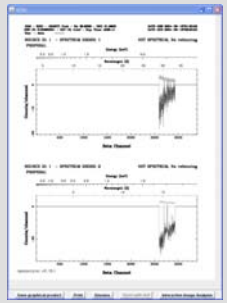


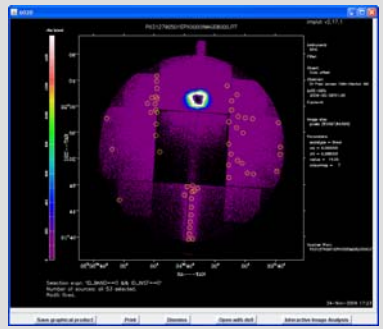


XMM-Newton Science Archive

- Available since April 2002:
  - http://xmm.esac.esa.int/xsa/
- Developed and released for the operational phase
- Still active development
- New data coming on daily basis
- 2 TB of data (FITS) on hard disk of raw and processed data, some catalogues
- Data processing done at Leicester
- On the fly reprocessing system available from the Archive (run at ESAC)







Archives at ESAC since ISO

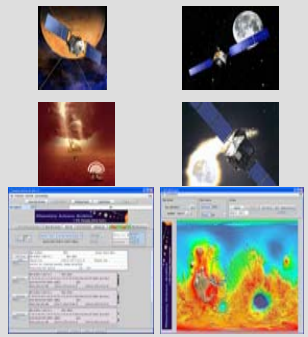
- ESAC is the Centre where most of ESA Scientific Archives are developed, maintained and operated.
- ESAC Science Archives Team is giving support to various projects
  http://www.rssd.esa.int/index.php?project=SAT

ISO Data Archive
Since December 1998

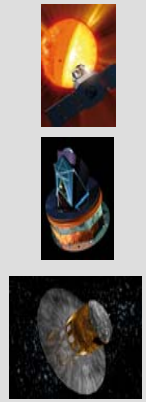
XMM-Newton Science Archive
Since April 2002

Integral SOC Science Data Archive
Since July 2005

Herschel Science Archive
Since 2009



Planetary Science Archive
Giotto, Mars Express
Rosetta, Venus Express
Smart-1, Huygens
Since March 2004



Soho, Exosat, Planck, GAIA,
… in the future

Presentation Overview

ESAC Archives History

ESAC Archives evolution

The “ABSI” concept

The SOHO Science Archive case

The EXOSAT Science Archive

Conclusion

Investigation of possible improvements

- ISO Data Archive (IDA) was pioneer in using three tier architecture and Java technology, state-of-the-art technology at the time (~1995)
- Initially thought to serve only ISO data, the IDA archive technology was later applied to the XMM-Newton Science Archive (XSA)
- Other archives followed suite (see before...)
- This architecture has allowed not only the building of ESAC archives but also their integration within the VO, but...
- A lot of the technology used at the time is obsolete
- An example: some of the nowadays available-everywhere scroll-bars were implemented “by hand” in certain window interfaces

Setting the ground: listing possible issues

- Made an exercise of “self-auditory”....:
  - Communication Client-Server done through home made RPC (Java serialised objects)
  - Transport done through TCP-IP in compressed mode
  - Business layer mixing service, transport and logic in a single implementation
  - Business layer makes uses of port 80 and blocks its usage to any other application running on the same machine
  - load balancing, security and proxy redirection are not available in our “home-made” server
  - .....
  - .....

Investigation on frameworks and available technology

- Main points:
  - 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 (!!)

The options and decisions

- 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, Galshfish, 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

Presentation Overview

ESAC Archives History

ESAC Archives evolution

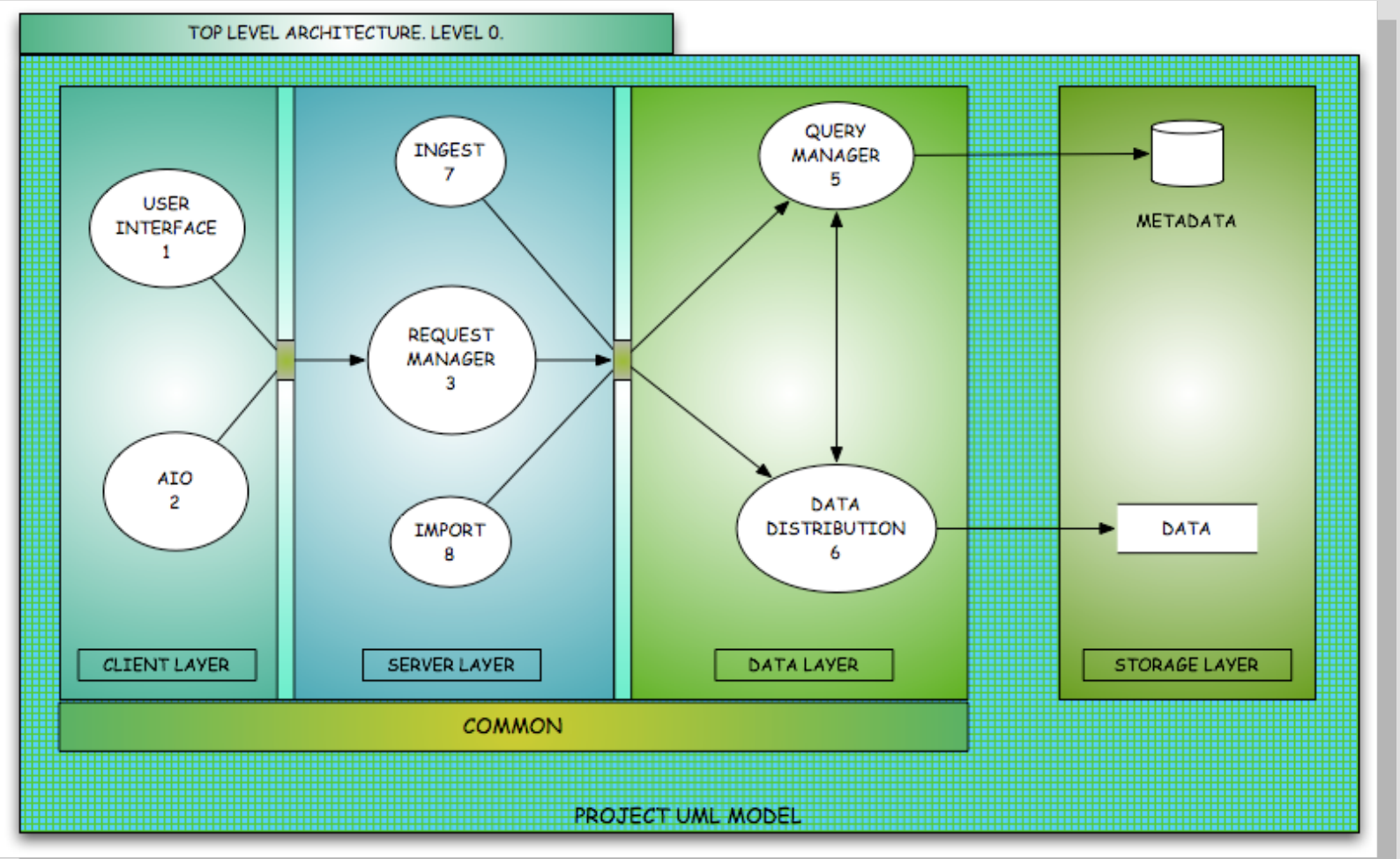
The “ABSI” concept

The SOHO Science Archive case

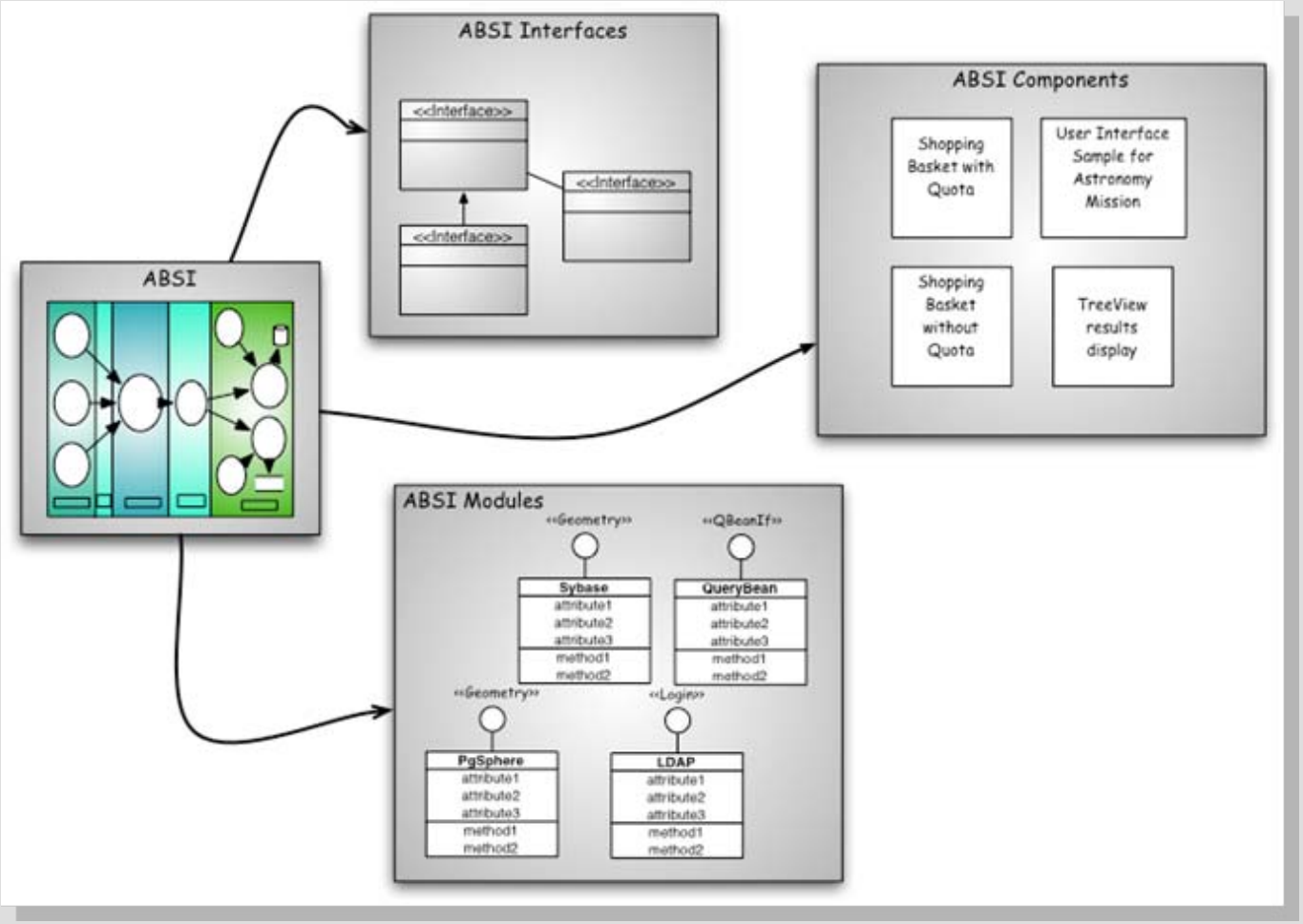
The EXOSAT Science Archive

Conclusion

Standard three-layer architecture



ABSI Elements (II)



The problem of handling big amounts of metadata

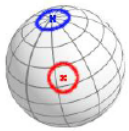
- We are dealing with more than a million observations (granularity in SOHO different from Astronomical cases). Database Table indexing gets overly complicated and joins poorly performant
- To know how to apply the joins to the different attributes requested ("where" part of the query), we implement the **Dijkstra** algorithm (shortest path algorithm, graph theory).
- **Dijkstra's algorithm**, conceived by Dutch computer scientist Edsger Dijkstra in 1959, is a graph search algorithm that solves the single-source shortest path problem for a graph with non negative edge path costs, outputting a shortest path tree.
- This algorithm is often used in routing.
- We have applied it to our database tables and relationships
- On-line examples:

  http://www.carto.net/papers/svg/dijkstra_shortest_path_demo/

Indexing spherical data in DB

- Index
- Coor
- Gets
- PgSp
- Provi
  - inpu
  - con
  - vari
  - circ
  - sph
  - indexing of spherical data types
  - several input and output formats
- Implemented in EXSA for the

3.4. Circle



A spherical circle is an area around a
the radius of the circle. Usage cases ar

- sites on earth having a maximum
- round cluster or nebula on sky spl
- a position with an undirected posi

A circle is specified using a spherical

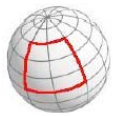
```
< point , radius >
```

Valid radius units are RAD, DEG, and I
than zero.

Example 7. A circle around the Nor

```
sql> SELECT scircle '< (0d
```

3.9. Coordinates range



A spherical box is a coordinates range. Hence, you can select objects within a longitude range and latitude range. The box is represented using two spherical points: the southwest (*pos_sw*) and the northeast (*pos_ne*) edge of the box. The input syntax is:

```
( pos_sw, pos_ne )
```

or

```
pos_sw, pos_ne
```

Note:

- If the latitude of the southwest edge is larger than the latitude of the northeast edge, pgSphere swaps the edges.
- If the longitude of the southwest edge is equal to the longitude of the northeast edge, pgSphere assumes a full latitude range, except that the latitudes are equal, too.

Example 12. Input of a full latitude range

A full latitude range between +20° and +23°.

```
sql> SELECT sbox '( (0d,20d), (0d,23d) )';
```

Example 13. A simple coordinates range

A coordinate range between -10° and +10° in latitude and 350° and 10° in longitude.

```
sql> SELECT sbox '( (350d,-10d), (10d,+10d) )';
```

New Archive creation Process

- Bottom-up approach: First build the general UML for the overall project. Then start building from there.
- UML → DB Design → Repository design → DAO (Data Access Objects) design → User Interface design
- Good UML design for project extremely important.
- Proper knowledge of the data by the SAT is crucial in order to build good Data repository and Data Distribution systems hundreds of mails interchanged between SOHO Archive Scientist and SAT Team on Data issues)

Presentation Overview

ESAC Archives History

ESAC Archives evolution

The “ABSI” concept

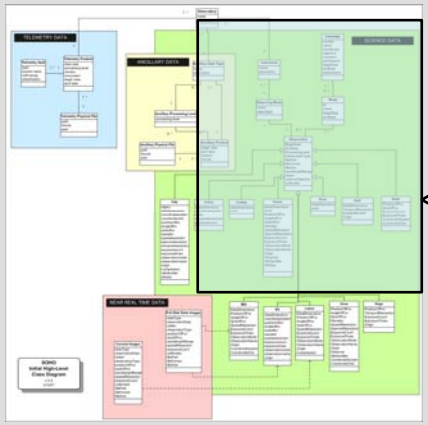
The SOHO Science Archive case

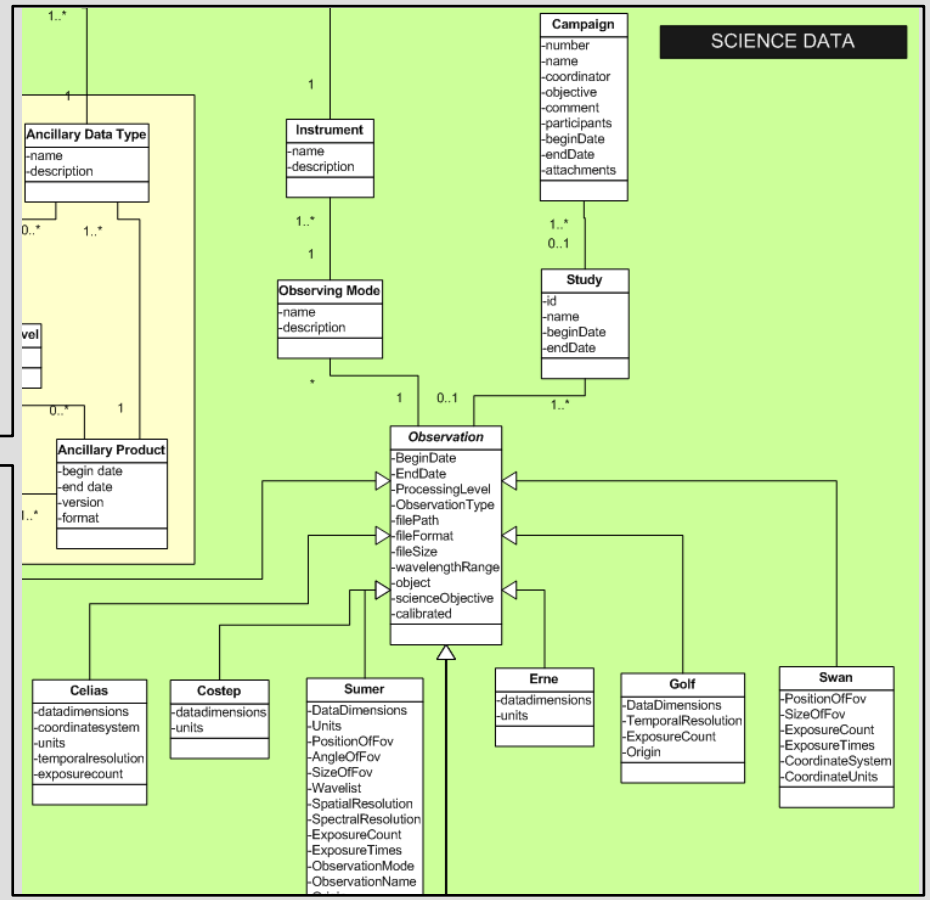
The EXOSAT Science Archive

Conclusion

SOHO Science Archive UML

SOHO Science Archive Data Model





Some numbers from SOHO

Data Type	Instrument	Data Format	Number of files
REALTIME	**EIT**	JPG	**918402**
REALTIME	**LASCO**	JPG	**765356**
REALTIME	**MDI**	JPG	**23022**
SCIENCE	**CDS**	FITS	**295519**
SCIENCE	**CELIAS**	CDF	**80632**
SCIENCE	**COSTEP**	ASCII	**30921**
SCIENCE	**EIT**	FITS	**472144**
SCIENCE	**ERNE**	ASCII	**20596**
SCIENCE	**GOLF**	FITS	**4450**
SCIENCE	**LASCO**	FITS	**658419**
SCIENCE	**MDI**	FITS	**83494**
SCIENCE	**SUMER**	FITS	**118709**
SCIENCE	**SWAN**	FITS	**9122**
SCIENCE	**UVCS**	FITS	**87226**
SCIENCE	**VIRGO**	FITS	**23765**

The SOHO Science Archive

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

General Usage (video)

Interoperability through IVOA protocols (video)

Time animator (video)

Presentation Overview

ESAC Archives History

ESAC Archives evolution

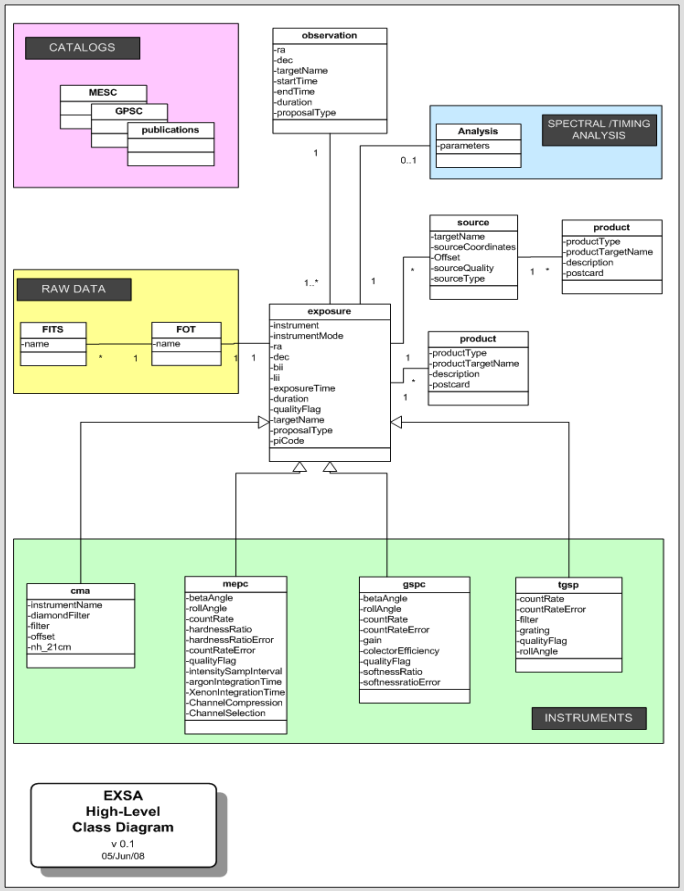
The “ABSI” concept

The SOHO Science Archive case

The EXOSAT Science Archive

Conclusion

The EXOSAT case



Different problem:

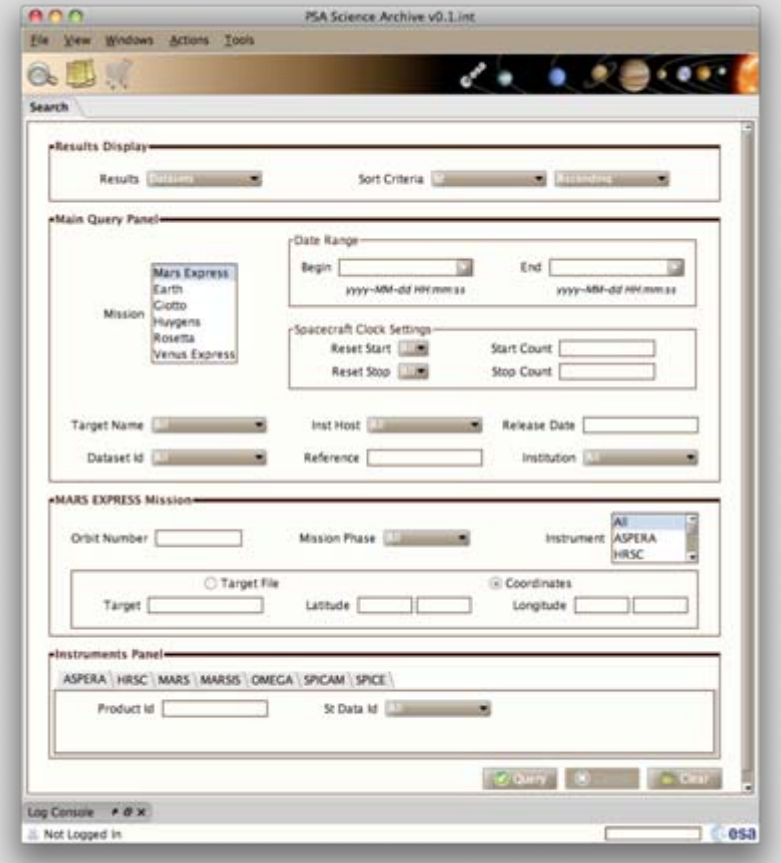
- not many instruments
- not a live project
- but facing standard issue in astronomy:
  - slow queries in spherical coordinates

The EXOSAT Science Archive

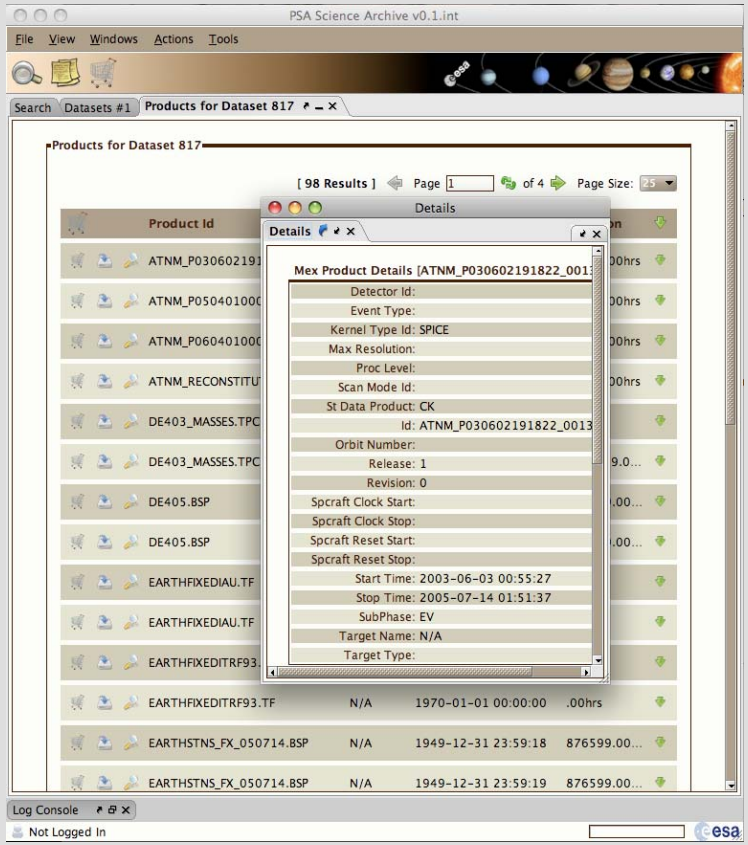
http://www.rssd.esa.int/index.php?project=EXOSAT&page=archive

General usage (video)

Adapting existing archives to new technology



Re-engineered Planetary Science Archive prototype (**PSA**)



Presentation Overview

ESAC Archives History

ESAC Archives evolution

The “ABSI” concept

The SOHO Science Archive case

The EXOSAT Science Archive

Conclusion

Conclusions

- State of the art technology ever changing, need to adapt to new trend
- Many options available: time is needed to investigate different options
- Flexibility is paramount when making decisions: will have to re-assess technologies in some time
- ABSI concept proven useful: modularity is key issue
- First ABSI produced archives: SOHO and EXOSAT
- SOHO released to public community on 28 Sep 2009
- EXOSAT released to public community on 1 Dec 2009
- Will adapt existing ESA archives to new technology
- Archives can be accessed at:
  - http://soho.esac.esa.int/data/archive/index_ssa.html
  - http://www.rssd.esa.int/index.php?project=EXOSAT&page=archive
- Science Archives Team pages