Building and Operating 15 Archives over 15 years:

Challenges and Lessons Learned

Christophe Arviset [(1)], Iñaki Ortiz [(1)], Pedro Osuna [(1)], Jesus Salgado [(1)]

[(1)] ESA-ESAC

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

EMail: Christophe.Arviset@esa.int, Inaki.Ortiz@esa.int, Pedro.Osuna@esa.int, Jesus.Salgado@esa.int

ABSTRACT

ESA's European Space Astronomy Centre (ESAC), near Madrid, Spain, currently hosts most of ESA space based missions' scientific archives[1], in planetary, astronomy and solar physics. All these science archives have been developed and operated by a dedicated Science Archives Team (SAT) at ESAC, enabling common and efficient design, development, operations and maintenance of the archives software systems. This also ensures long term preservation and availability of such science archives, as a sustainable service to the science community.

Over the last 15 years, archive development and operations have faced many challenges that will be described in this paper, such as supporting archives through various mission phases (development, in-orbit operations, post-operations and archival phase) and associated funding schemes, providing services to the Science Community and to the ESA Science Operations Centres, data processing and reprocessing strategy and resulting data versioning and retention policy, handling of proprietary and public data, needs for various archive access interfaces and interoperability with other archives and tools, archive technology evolution through time, commonality across archives while supporting requirements from different communities, development team and schedule management for many archive projects in parallel, archive usage metrics and reporting, long term data and software preservation, …

Keywords: Archives, lessons learned.

BUILDING AND OPERATING 15 SCIENTIFIC ARCHIVES

ESAC evolving towards a multi mission Science Data Centre

In the past, ESA's policy was to delegate the responsibility of its scientific archives to the scientific community. With the advent of the World Wide Web and the increased number of science missions in the late 90s, ESA needed to ensure the long term preservation of its scientific data holdings through powerful and easy to use on-line science archives. This started with the Infrared Space Observatory (ISO) Data Archive at ESA's centre in Villafranca del Castillo, near Madrid, released in late 1998 soon after the end of the mission in-orbit operations. Building on this success and the increase of Science Operations activities there, more archives were developed and operated, in particular the XMM-Newton Science Archive in 2002 and the Integral Science Data Archive in 2005. In parallel, the Planetary Science Archive

[1] All Science Archives at ESAC can be found at http://archives.esac.esa.int/

was released in 2004 as a single archive for all ESA's planetary missions, first with Giotto, and later on with Mars Express (2005), Huygens (2006), Venus Express (2009), Rosetta and Smart-1 (2010). Herschel was launched in May 2009 and was the first mission with its archive fully functional to support science operations from day 1.

After this first generation of ESA scientific archives, it was time for a technology update which took place for the Soho and Exosat Science Archives (2009), as well as for the Planck Interim Archive, also released in 2009, for consortium use only. Migration to this new archive technology is now taking place for all archives developed prior to 2009.

ESAC (European Space Astronomy Centre) is then hosting most of ESA's space science missions' archives becoming naturally the ESA Science Data Centre. The saga of Science Archives developed and operated at ESAC will continue with the final archives for Cluster and Ulysses (from ESTEC) and the Hubble archive (from ST-ECF, Garching, Germany) being now brought into the multi mission archive infrastructure at ESAC. Furthermore, development of future missions' archives is ongoing (Gaia, LisaPF, …) to ensure archive availability as soon as the in-orbit operations starts.

Common archive technical framework

All science archives at ESAC share the same technical framework, compatible with the OAIS standard. Their flexible multi tier architecture (in Figure 1) offers the modular separation from the various archive functions.

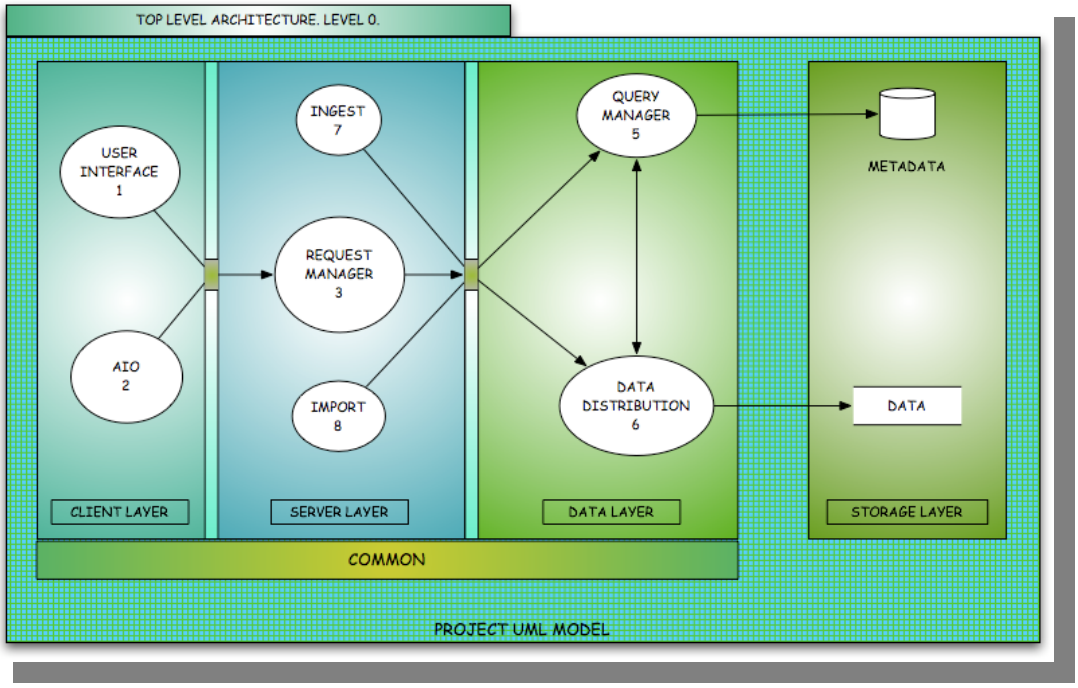

Figure 1: ESAC Science Archives Architecture

The Storage Layer consists of on-line data repositories on hard disks, with data volumes ranging from a few dozens of GB (for the old missions) to 100TB (Herschel, HST, Cluster). Keeping data on hard disks facilitates greatly the archive operations and migration process. Metadata is systematically extracted from the data into a relational database, such as Sybase or Oracle, and more recently into open source databases like PostgreSQL.

The Data Layer represents the software systems (Query Manager and Data Distribution) that access the Storage Layer. This provides independency from the way the data and metadata are stored, enabling evolution towards new storage and database systems.

The Client Layer offers the front-end accesses to the archives. The main access can be a graphical user interface (through rich internet client or a thin web client, see Figure 2) with powerful and friendly search, visualization, selection and download facilities. An alternative access is through web service called AIO (Archive Inter-Operability system) that offers a scriptable interface to the archive. Interoperability with other archives can be built on top of the AIO, translating this web services into interoperability standards (from IVOA or IPDA).

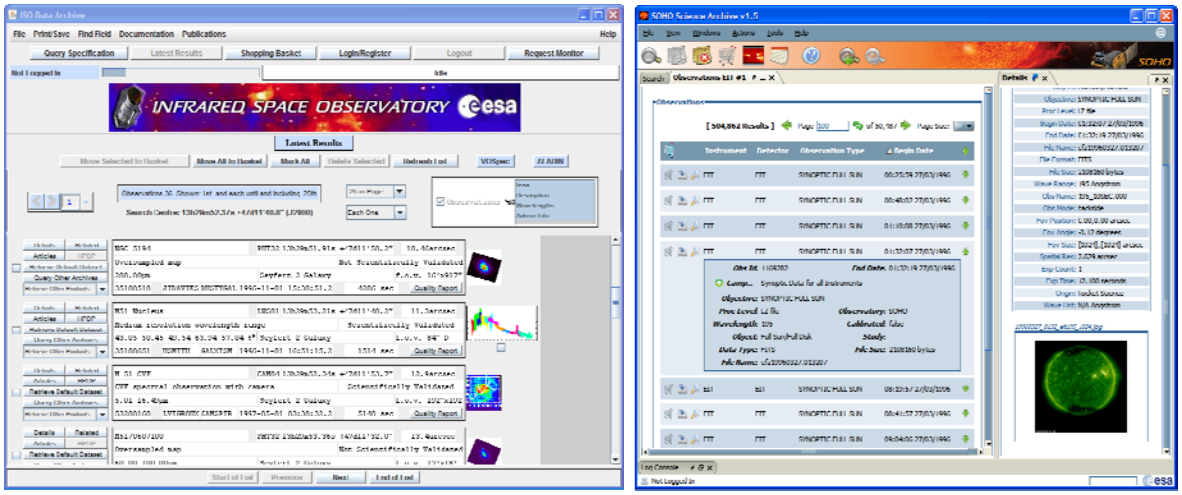

Figure 2: ISO and Soho Archives Graphical User Interfaces

The Server Layer can be seen as the "engine" of the archive, connecting the front-end layer to the data layers. Together with some Common libraries, it ensures proper access to the metadata and data (respecting the corresponding project proprietary access period), logging and usage reporting functions as well as other archive administration functions.

Having a common archive technical framework across various archives enable faster and cheaper archive development cycles, more reliable software as the code is being re-used, and knowledge transfer from project to project and people to people. Furthermore, corrections and improvements done for one archive project is being feedback to other archive projects as they are sharing this common framework. This allows also efficient and cheap long term operations of the "old" archives (for which no more project funds are available), as they are using similar technology and operations concepts as the more "recent" archives.

Once in production, ESAC Science Archives are working "autonomously", without the need for any archive operator. Through their standard web browser, end users can search the archives, visualize their items of interest, and select them for immediate download (thanks to all data being on-line on disks). All data are being distributed through the internet via standard FTP protocol.

Common ESAC Science Archives Team

All ESAC Science Archives are developed and are operated in-house by a single team. The same people are in charge of the design, implementation, deployment, testing and operations of the archive. With no need of archive operator, the archive software developers are in charge of monitoring that all archive processes (data import and associated metadata extraction and ingestion, data distribution) are running smoothly. Over the years, the Science Archive Team has increased from a few people to 18 persons to date, who are supporting the 15 archives at ESAC, This represents significant savings, taking into account the diversity of ESA missions and associated datasets. Funding for this team comes from the Science Operations Team project for which an archive is needed. Funding depends of the phase of the project, the complexity and the volume of the mission data. Some limited funds also exist to support archives in the long term as well as cross archive technology development and update.

Initially, when there were only a few archives to support, the team was structured by functions (database and data distribution experts, user interface experts) where people were providing a fully horizontal support to several missions in parallel. With the increase of the number missions, it became too complex for people to support so many archives in parallel, with often conflicting and moving milestones due to project external constraints. Team members have then been assigned to a particular archive project, in group of one to three persons, depending of the project archive workload. People can then be fully dedicated to one archive, and stay closer to the science operations team with more flexibility to adapt to new project priority and schedules, should be need arise, not being affected by other tasks related to an other archive project. Through the years and varying requests from the project, people will be re-assigned to work from one archive to another, with a smooth transition as using the same archive technical framework. Furthermore, a core archive team of a few people remains to ensure the consistency of the technical archive framework, maintain and develop the archive knowledge in the long term and support the archives for which no more project funds are available.

LESSONS LEARNED

Many "customers" for the same Archive

Scientific Archives' main customer is the general Scientific Community, exploiting the public data holdings to perform their science, with the final aim to publish scientific papers in the literature. Hence, data must be made available as soon as possible into the archive, with the highest level of quality, and it is recognized that the quality will increase with the knowledge of the on-board instruments, improved calibration and associated data processing software resulting in subsequent data reprocessing campaigns. The general scientist need well calibrated data, well documented and easily accessible through a simple and user friendly user interface offering powerful query and download facilities, based on general scientific parameters.

Furthermore, quite often data holdings are subject to a proprietary period (usually one year after the data have been processed) towards the instrument teams, key programmes or directly towards the observer who's been granted telescope time for observatory missions. These users are usually more knowledgeable about the mission, the instruments specificities and require a more expert access to the archive. Additionally, data access and distribution should be done through a secure manner to ensure the proprietary period.

The Science Operation Teams are other very important customers of the archives. They need to have privilege, fast and efficient access to the metadata and to the data, coupled with important processing capabilities so they can perform monitoring and calibration tasks. They are experts on the instruments and on the data so they don't need fancy user interfaces but are usually better served with a machine interface or web services which give them the power to run directly complex analysis tasks on the data. In some cases, the archive might even be used for the various types of processing both for inputs (of raw data) and for outputs (for processed science data), putting the science archive at the heart of the Science Ground Segment (see Figure 3).

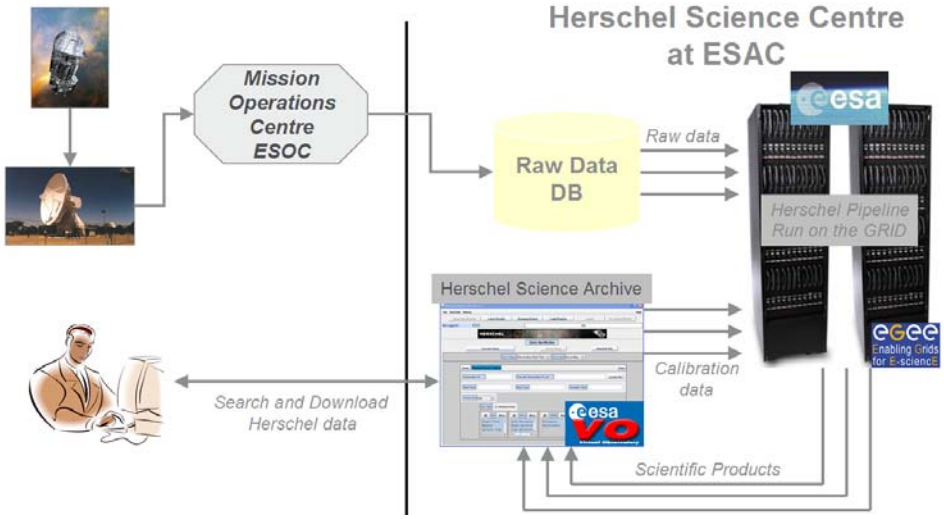

Figure 3: Role of the Archive in the Herschel Science Centre

All these groups might have very different user requirements, which can even sometimes be orthogonal to each others. Similarly the implementation priorities can vary from our group to another and it is difficult for the archive development and operations team to determine which are the more important ones.

Here comes the critical role of the Archive Scientist. As a scientific expert of the field, part of the Science Operations Team, he/she is in charge to compiling the user requirements from the various archive "customers", determine the implementation priority and interface daily with the Science Archives development and operation team to make sure that requirements are well understood and that the final archive software will meet all users' expectations. He/she is an active archive user, uses regularly the archive to search and download data to perform his/her science. This also includes the important role of the archive "ambassador" inside the project team and outside in the community.

Supporting various missions phases

One of the specificity of archives project is their long lifetime and the difficulty to make this understood to mission project team. Although the archive will be final legacy of the space mission, it is rarely considered with enough importance, especially in the earlier mission phases, as the archive is often seen as a "JBOD" ("Just a Bunch Of Disks") at the very end of the chain of the ground segment systems. Nonetheless, as demonstrated by the OAIS model, an archive is much more than a simple data repository and requires significant system and software engineering to ensure its usefulness and long term preservation through time.

During the mission development phase, prior to launch, the archive is clearly not a priority; the data products and processing algorithm are often not yet defined as the instruments still need to be built. But from launch onwards, the archive needs to be operational to ingest and distribute data, hence the need to start the archive development process well before the in-orbit phases, even without a full definition of the data products to be archived. During the operations phase, the archive must be able to accommodate regular changes and improvements on the data processing software and associated reprocessing campaigns. During these phases, project priorities are very focuses towards their own needs and requirements and they fail to see the global and long term view.

In the post-operations phase, all mission knowledge and data needs to be preserved into the archive with the appropriate level of metadata and documentation. Here again, reprocessing campaigns are key to ensure the best final legacy archive. A new requirement often comes which is to ingest high level data products coming from external teams and make then available into the archive. This poses specific difficulties as the data format and description is not always compatible to what comes out of the standard data processing pipeline. Overall, adequate preparation and funding during this phase are essential to the success of the next phase.

The most challenging phase will start after the end of the project lifetime, where the archive still needs to be preserved in the long term, although the mission people have moved to other projects and no more dedicated project funds are available.

All phases have their own specificities with regard to archive development and use, and experience has shown that the most successful archives are the ones which have contemplated all aspects during each phase. In this context, the Herschel Science Ground Segment adopted very successfully the concept of "smooth transition between phases" for all software subsystems. In particular, the Herschel Science Archive had been developed and used for all scenarios (including various types of reprocessing campaigns) during the development phase and was fully operational by launch in May 2009 (see Figure 3).

Data processing and reprocessing strategy

Data processing pipeline consists of converting the raw data coming down from the spacecraft into processed science data that can be archives and distributed to the science users, so they can perform their science and write scientific publications. Significant expertise on the instrument is required to develop such data processing software. Although that requires more engineering and development work, the best scenario remains when routine "*systematic processing*" takes place daily as soon as the data is received from the spacecraft so it can be made available very quickly (potentially subject to a proprietary period) within a few hours after reception.

Through the mission operations, increased knowledge on the instrument is gained which results in improved data processing pipeline, hence the data initially processed after reception is becoming "obsolete". This imposes the need to perform reprocessing campaigns of all data taken to date with the latest version of the pipeline software. This "*bulk reprocessing*" exercise requires careful planning of cpu and personal resources and can take up to several weeks or several months, depending of the volume of data. Once finished, all resulting updated data products are ingested into the archive and made available on-line to the archive users. But after a few more months, the pipeline processing software will have improved again and the "updated" data becomes again "obsolete". Because of their complexity and duration, reprocessing campaigns don't happen too often.

To overcome this shortcoming, archive can also offer an "*on-the-fly reprocessing*" capability. In complement to the on-line version of the data products, the end users can choose to request the raw data to be reprocessed on-the-fly with the latest version of the data processing pipeline available. This will take a bit more time (depending on the processing time and the number on parallel on-the-fly reprocessing requests) but will enable the users to always have access to the best version of the science data.

Although all processing scenario (systematic, bulk, on-the-fly) can be considered independently and generate their own specific costs, experience has shown that the most effective combination consists of:

- Systematic processing on a daily basis, as soon as the raw data is received, including archiving of the resulted science data and extraction of the corresponding metadata
- On-the-fly reprocessing capability offered from the archive, but without archiving of the resulted science data into the archive
- Regularly (every one or two years), complete bulk reprocessing of all the raw data into a new updated on-line science data set, with the extraction of the corresponding (updated) metadata

Each bulk reprocessing will generate a complete new set of all the science data holdings that will be archived, although experience has shown that there is only the need to provide access to the latest version of the data only.

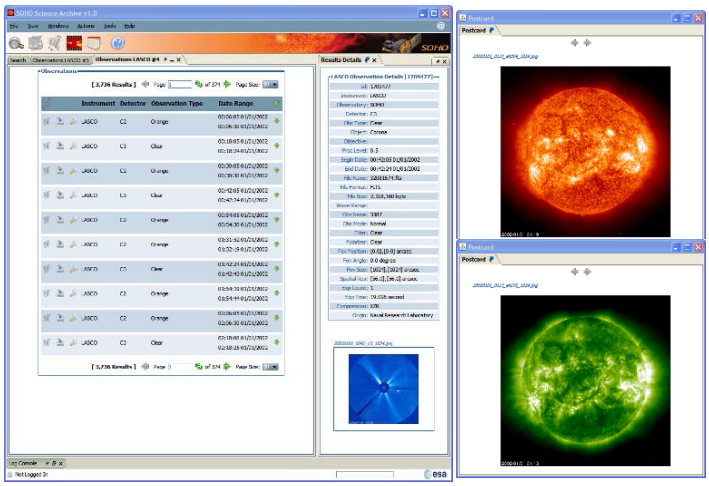

Figure 4: Rich Internet Application for the Soho Science Archive

Multiple channels for archive access

As mentioned earlier, scientific archives have various types of "customers" who have different requirements depending of the focus of their work. Some people like to have GUI with various inputs parameters organized by areas; others prefer to have a simplistic search functionality ("Google-like") offering natural language querying facilities. Users with more IT abilities will prefer a scriptable interface they can include in their own programs. Over the years, we have realized that there won't be a "one fits all" solution, and to provide the required services to the widest range of users, we had to offer multiple channels to the archives, in a non-exhaustive list so far:

- Rich Internet Application (Figure 4), offering powerful and sophisticated query and download facilities, linked to data visualization or analysis tools, interoperability with other archives and application.
- Standard thin web client, with simple query parameters and download facilities
- Scriptable or machine interface to the metadata and to the data directly, easily coupled with other programs and applications
- Map based client for some planetary mission (Figure 5)
- Access through Google Sky or World Wide Telescope, more geared towards public outreach
- Access through new portable devices (smart phones, tablets), also more geared towards public outreach

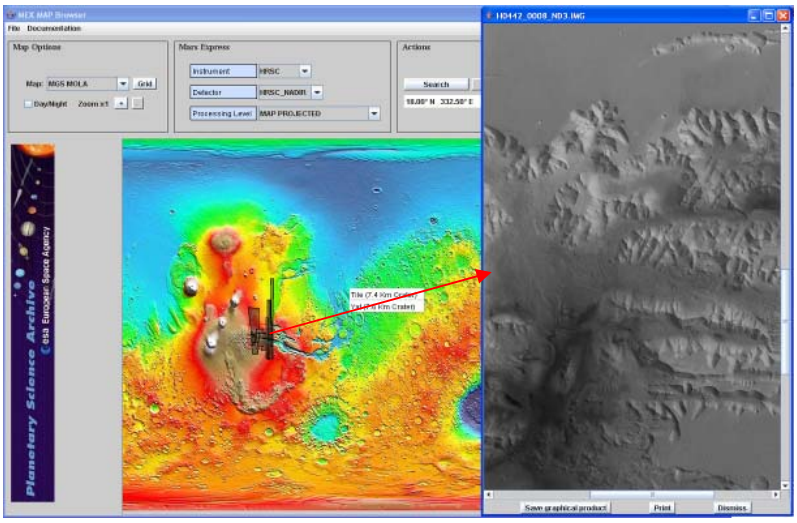

Figure 5: Map Based interface for Mars Express

Archive technology evolution through time

Archives have by definition a long lifetime in a period where IT technologies are evolving very fast. The mid 90s saw the advent of the World Wide Web which revolutionized the way the scientific community expects to have access to the science data. All archives must now be on-line with web based interfaces and immediate data access capabilities (within the limit of the network). Newer technologies such as Web 2.0 and the cloud may offer new ways of archiving, curating and distributing science data, although it is today too early to determine exactly if and how this could take place and this could be the purpose of a completely separate paper.

When released in 1998, the ISO Data Archive was pioneering new technology (at the time) and was probably the first ever astronomical archive written completely in Java. Nonetheless, some of the ad-hoc solutions that had to be implemented from scratch are nowadays offered as standard solutions (such as application servers, client-server communication, only commercial relational database management systems, very poor libraries of graphical user interface components, server ports not blocked by current firewalls, etc). It became necessary to make a technology audit and decide for new technologies that would fit better current users' expectations and easier development and maintenance costs (see reference [1]). It is worth noting that the multi tier architecture remained quite stable, although each subsystem went through drastic technology migration and improvements. Other worth noting improvement was the migration from CDROM jukeboxes towards hard disks which greatly increased performance for data import, data distribution and even reprocessing campaigns. Moving away from commercial RDBMS (Sybase, Oracle) towards similar open source products (PostgreSQL) also opened new perspectives such as powerful add-ons for geometrical searches (pgSphere[2] which provides spherical data types, functions, and operators for PostgreSQL), while removing software maintenance contract costs. Adoption of interoperability standards (SAMP[3]) allowed archives and application to inter connect and share data amongst them, focusing the archives towards data search while visualization and analysis is made on a separate application (Figure 6).

[2] http://pgsphere.projects.postgresql.org/

[3] IVOA Simple Application Messaging Protocol : http://www.ivoa.net/Documents/SAMP/20101216/

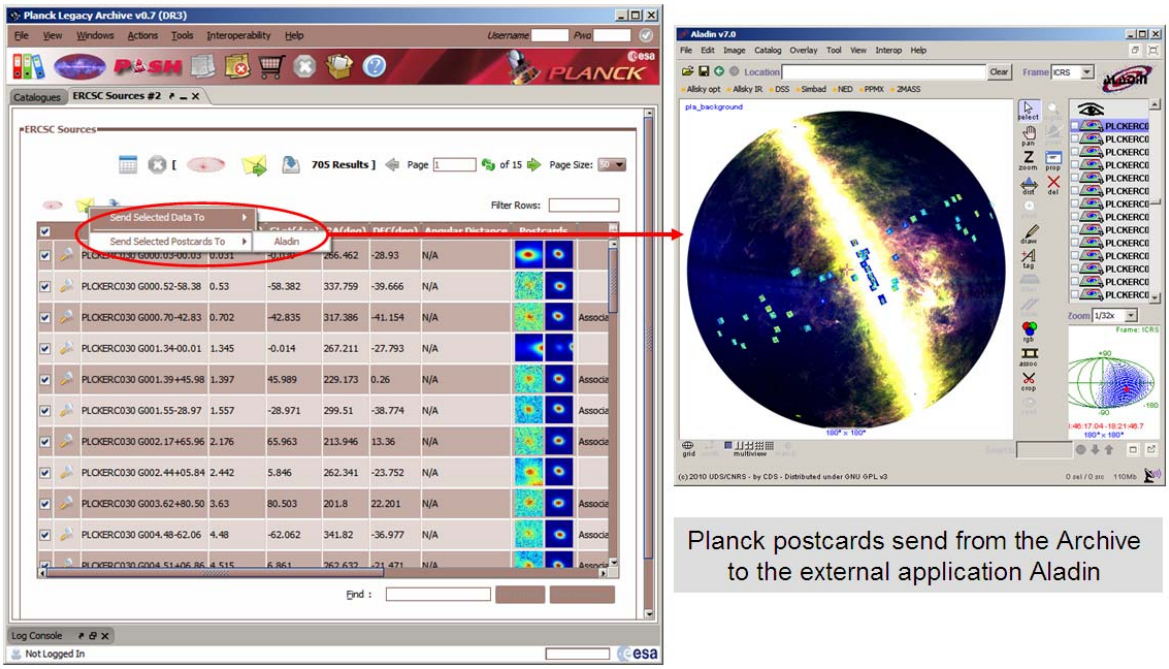

Figure 6 : Archive and Application interoperability through the SAMP protocol

Archive commonalities vs. requirements diversity

By sharing a common architecture, re-using code components and operations practices; by being developed and operated by the same Science Archives Team, Science Archives at ESAC benefit greatly of this synergy amongst each others, although they are addressed to various scientific community (astronomy, planetary and solar heliospheric). They present a similar look and feel to the end users, development, maintenance and operations costs are lower and their reliability higher as technology and knowledge is constantly being transferred from one project to another. Long term preservation is also ensured as the operations of archives without any more project funds can be performed at a very low cost since operations practices are similar to other “active” archives projects.

Nonetheless, this archive commonality is regularly put into question by projects, claiming their field of science, their data sets, and their user requirements are very specific and can not be compared to others previous projects. The archive architecture and the archive development team need to be flexible enough to accommodate this project’s specificities while maintaining similar the core of the archive, to ensure the gain in the commonalities between various projects in the short and medium term, but even more important, in the long term where no more project fund will be available for the archive operations. This definitely represents a difficult “communication” challenge for the ESAC Science Archives Team vis a vis new projects who are (correctly) focuses on their project’s priority while the ESAC Science Archives Team needs to keep the overall long term and common “picture” of all Archives, at the risk of being seen inflexible and not enough customer focus. Support from high level management is here a key to success so to set this as a strategic decision for all ESAC Science Archives.

Archive usage metrics

The success of an archive (and somewhat indirectly of the mission itself) could be measured by its usage. At various meetings, committees, conferences, one can report about the usage of a specific archive. Through time, we have tried to determine various types of metrics that could represent adequately the usage of one archive, such as: number of accesses to the graphical user interface or scriptable interface, number of queries run, number of data download requests, number of datasets downloaded, volume of data downloaded, number of new users, number of active users (downloading data at least once a month), distribution of users per country, total volume of the archive as a function of time, etc… Each of these

metrics represent a specific type of usage of the archive, and the values obtained might vary greatly from one archive to another due to the mission setup : volume of the data, organization of the datasets (many small datasets, few big datasets), size of the scientific discipline, type of mission (PI missions, observatory mission), etc.. Normalization factors could also be introduced in an attempt to "compare" one archive usage to another. In practice, we've seen over the years some factors which significantly increase the usage of the archives, in particular:

- Systematic and daily processing of the raw data into science data with immediate availability
- Usage of the archive as the prime distribution mechanism to the instrument teams and observers
- Usage of the archive by the Science Operations Team

A good example of such archive is the XMM-Newton Science Archive as shown in Figure 7.

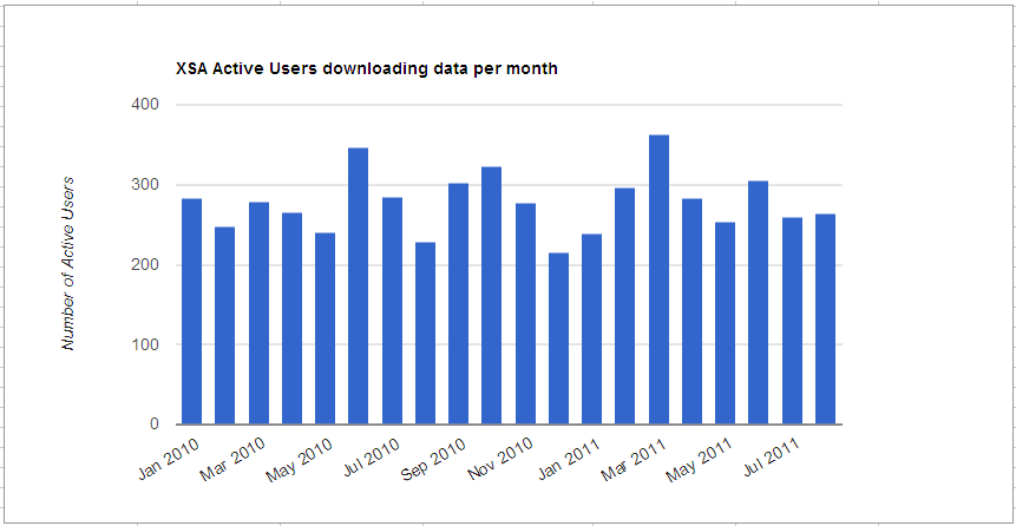

Figure 7: XMM-Newton Science Archive statistics about active users

LTDP – Long Term Data Preservation

So far only a few science missions (ISO, Hipparcos, Giotto, Smart-1) have entered their legacy archive phase, where no more project people and funds are available so where archive maintenance and operations should be done at almost no cost. This is made easier, as many other archive projects are very active, either in operations or in development. As part of the consolidation of ESAC as a Science Data Centre, it would make sense to repatriate the archives of ESA's missions that were not kept at ESA's premises in the past. This would require some investments, not so much for the IT infrastructure (as usually data volume are very low for such missions), but more difficulty to get back to the documentation to extract the corresponding knowledge and metadata and perform the development of the archive in the existing ESAC Archive technical framework, in order to ensure long term common maintenance and operations to keep costs down in the long run.

Exploitation of science data has no limit and very often reprocessing of the data (from various levels) is required to better understand the physical phenomena beyond the observations. Preserving only the data might not be enough so there is the need of preserving as well the science data processing software that had been used to initially process the raw data so reprocessing capabilities can be offered to the science users. But the science data processing software is usually attached to specific operating systems platform which rapidly become obsolete due to the constant evolution of the IT technology. Hence, one could envisage recovering all existing science data processing software for old ESA science missions, determining if they can still be run with current computers and eventually porting them to modern operating systems platforms so reprocessing of the raw data into science data is made possible.

Furthermore, based on new computing paradigms, we should offer these science data processing capabilities as "software as a service", where these software can be run by external users without forcing them to install any particular software which might require complicated (or even simple) installation procedure on the user's computer, with requirements such as specific operating systems, additional third party software and libraries, compilers, etc… Usage of GRID or Cloud technology could be offered to end user to enable them transparent access to science data processing software, through the design and provision of long term data processing infrastructure and capabilities for ESA science missions.

With the current list of missions in operations, more and more will enter their legacy archive phase in the coming years and a proper funding scheme for these LTDP activities will need to be found. As this matter is clearly shared with other directorates within ESA, the Earth Observation, the Science and Robotic Exploration and the Human Space Operations directorates are preparing jointly a proposal to address all LTDP activities at ESA for the coming decade.

CONCLUSION

Since the mid 90s, ESAC has been hosting the science archives of most of ESA's science missions. This consolidation is continuing to include all past and current archives in astronomy, planetary and solar heliospheric into a full ESA Science Data Centre (http://archives.esac.esa.int/). Although addressing the specificities of all projects, commonality is ensured through a common archive technical framework and through the ESAC Science Archives Team who is in charge of the design, development, deployment, operations and maintenance of such archives through all the mission phases.

Science Archives are considered a strategic asset at ESAC, with the aim of addressing independently and globally the main four main areas: 1) long term data preservation, 2) science and user support, 3) project science operations support and 4) archive technology to ensure the best archive service to all types of users.

REFERENCES

[1] – P.Osuna et al, ESA New Generation Science Archives: SOHO and EXOSAT, ADASS XIX Conference, 2010

ACKNOWLEDGEMENTS

The authors want to thank here all the members of the ESAC Science Archives Team, as well as the Archive Scientists from within the ESA science missions who have all participated extensively to this activity.