RISA/VM-SAS: the XMM-Newton way to preserve scientific analysis capabilities in the long term

Carlos Gabriel, Aitor Ibarra Ibaibarriaga

ESA-ESAC

PO Box 78, 28691 Villanueva de la Cañada, Madrid, Spain
EMail: Carlos.Gabriel@esa.int

ABSTRACT

The X-ray Space Observatory XMM-Newton, now in operation for more than 11 years, produces huge amounts of data, reduced interactively by observers using the distributed XMM-Newton Scientific Analysis System (SAS), as the primary analysis package. At the same time, the Pipeline Processing System (PPS), a pipeline incarnation of the SAS, reduces data to calibrated event lists, images, spectra, light curves, etc, according to the best available knowledge of calibration and methodology for curation in the XMM-Newton Science Archive (XSA).

Due to the distributed nature of the SAS, the package has to evolve naturally with changing operating systems, new platforms, new flavours and updated libraries, in order to maximize scientific return. This means high maintenance costs which are unlikely to be supported after the end of the lifetime of XMM-Newton (~ 2016-2019).

XMM-Newton data will not be superseded by new or better observations for many years and their full scientific exploitation could require repeatedly more refined analysis compared to that stored in the XSA. Our answer to the challenge of maintaining full analysis capabilities of XMM-Newton data for a decade or longer after the end of operations is a combination of:

- RISA, the Remote Interface for Science Analysis, a web service around the SAS based on a client / server application, developed to be used in GRID environments, but easily adaptable to any system architecture such as cluster or cloud computing,

- VM-SAS, SAS embedded into a virtual machine, which can run under any operating system.

We will discuss the need for long-term analysis capabilities, as well as RISA and VM-SAS as urgent and relevant complementary solutions. Further advantages of this approach will also be presented.

Keywords: SAS, RISA, XMM-Newton, analysis capabilities preservation

I - INTRODUCTION

Interactive data analysis, tailored both to the scientific needs and to the characteristics of the single observation to be analysed, is a fundamental part of the successful scientific exploitation of data. A lot of intelligence and knowledge can be put in automatic reduction and processing of data for deriving final calibrated products, but when it comes to extract the last bits of information it may be necessary to interact with the data in certain steps of the reduction, or even reprocess it fully using a somewhat changed scheme. Furthermore, new analysis techniques, new calibration algorithms, could be crucial for solving scientific questions. And those new "items" may be developed long after the mission originating the data has ceased being operative. In this sense, our current[1] approach of developing interactive

[1] We cannot speak here about traditional approach, since it is just in the last decades that large astronomy projects, and not all of them, consider the provision of specific analysis packages to the community as their responsibility.

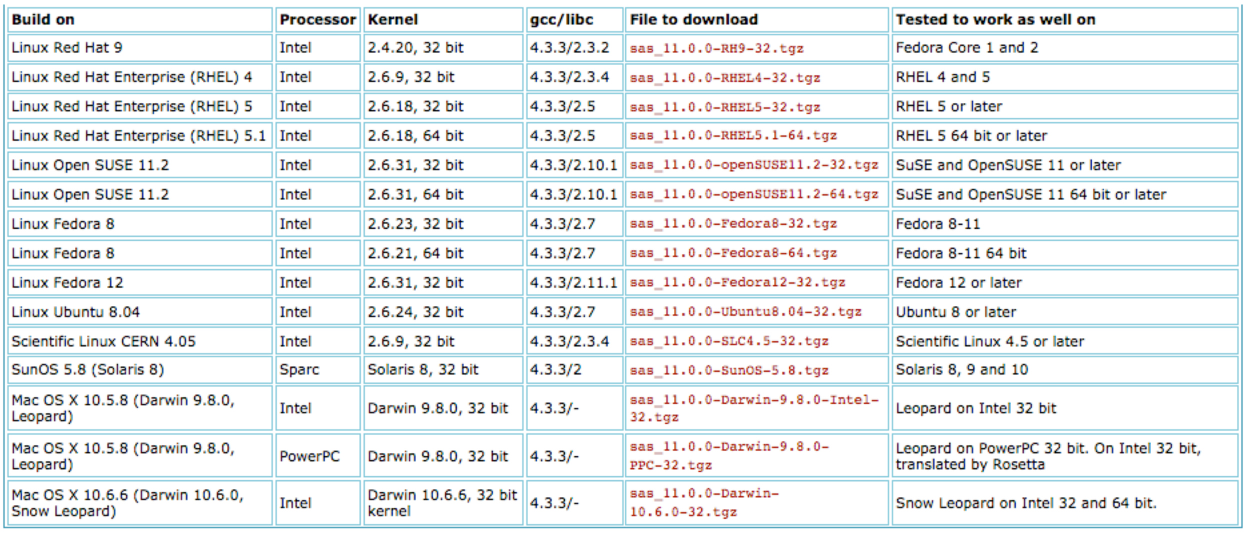
analysis tools shortly before or around the beginning of operations, maintaining and upgrading them during the lifetime of the project, and then leaving behind a more or less frozen version at the end of it, together with data processed at "final" stage, shows a clear deficiency. The difficulties imposed to someone trying to reassess the data with a new approach or with a view angle that was not in the original scope of the "final" processing is extremely hard to overcome. In the case of a large mission, like XMM-Newton, whose data will not be "superseded" by new or better observations for many years, this leaves a hole ("superseded" here in quotes, since of course observations can never become obsolete, taking into account that they represent measurements at a different epoch - essential in the case of variable objects, such as, to a large extent, X-ray sources). Again, full scientific exploitation could require repeatedly more refined analysis since even very long-term projects (≥10 years) show fundamental upgrades in their calibration or algorithmic treatment of data after such a long.

Furthermore, a fundamental problem arises for those packages sensitive to the evolution of operating systems and libraries (which S/W is not sensitive at all?). Fully freezing a package is a chimera, unless the whole environment is frozen alongside it, and this is in principle incompatible with a distributed analysis package, as the XMM-Newton Science Analysis System (SAS).

II - SAS, INTERACTIVE ANALYSIS AND PIPELINE SOFTWARE

The Science Analysis System (SAS) [1] is the package used for interactive and pipeline data reduction of all XMM-Newton data. It is a freely distributed analysis system, written and maintained by a large group of scientific software developers and instrument and calibration specialists. It is used by a large community of around 1500 astronomers and has been used by almost every one of the 2800 refereed scientific publications written so far using data from the XMM-Newton mission [2].

Such a large community can only be reached by distributed software if this runs on almost every important operating system and flavour. Furthermore, it has to evolve continuously with the operating systems, and this translates to high maintenance costs. Due to this and to the continuous further development of the package, we build and integrate SAS every night on machines running several Linux and MacOS versions, both in 32bit and 64bit. Together they constitute the officially supported versions. From one SAS version to the next SAS version (normally released with a frequency of one per year) we migrate several of them. The actual version, SAS 11.0, has been released in February 2011 in fifteen different flavours, including Linux, MacOs and also Solaris (which was announced at the time of that release to be fully discontinued). Fig. 1 shows the list of different binaries released, with other information, eg. kernel type and also information about which other platforms are covered by those binaries.



Build on	Processor	Kernel	gcc/libc	File to download	Tested to work as well on
Linux Red Hat 9	Intel	2.4.20, 32 bit	4.3.3/2.3.2	sas_11.0.0-RH9-32.tgz	Fedora Core 1 and 2
Linux Red Hat Enterprise (RHEL) 4	Intel	2.6.9, 32 bit	4.3.3/2.3.4	sas_11.0.0-RHEL4-32.tgz	RHEL 4 and 5
Linux Red Hat Enterprise (RHEL) 5	Intel	2.6.18, 32 bit	4.3.3/2.5	sas_11.0.0-RHEL5-32.tgz	RHEL 5 or later
Linux Red Hat Enterprise (RHEL) 5.1	Intel	2.6.18, 64 bit	4.3.3/2.5	sas_11.0.0-RHEL5.1-64.tgz	RHEL 5 64 bit or later
Linux Open SUSE 11.2	Intel	2.6.31, 32 bit	4.3.3/2.10.1	sas_11.0.0-openSUSE11.2-32.tgz	SuSE and OpenSUSE 11 or later
Linux Open SUSE 11.2	Intel	2.6.31, 64 bit	4.3.3/2.10.1	sas_11.0.0-openSUSE11.2-64.tgz	SuSE and OpenSUSE 11 64 bit or later
Linux Fedora 8	Intel	2.6.23, 32 bit	4.3.3/2.7	sas_11.0.0-Fedora8-32.tgz	Fedora 8-11
Linux Fedora 8	Intel	2.6.21, 64 bit	4.3.3/2.7	sas_11.0.0-Fedora8-64.tgz	Fedora 8-11 64 bit
Linux Fedora 12	Intel	2.6.31, 32 bit	4.3.3/2.11.1	sas_11.0.0-Fedora12-32.tgz	Fedora 12 or later
Linux Ubuntu 8.04	Intel	2.6.24, 32 bit	4.3.3/2.7	sas_11.0.0-Ubuntu8.04-32.tgz	Ubuntu 8 or later
Scientific Linux CERN 4.05	Intel	2.6.9, 32 bit	4.3.3/2.3.4	sas_11.0.0-SLC4.5-32.tgz	Scientific Linux 4.5 or later
SunOS 5.8 (Solaris 8)	Sparc	Solaris 8, 32 bit	4.3.3/2	sas_11.0.0-SunOS-5.8.tgz	Solaris 8, 9 and 10
Mac OS X 10.5.8 (Darwin 9.8.0, Leopard)	Intel	Darwin 9.8.0, 32 bit	4.3.3/-	sas_11.0.0-Darwin-9.8.0-Intel-32.tgz	Leopard on Intel 32 bit
Mac OS X 10.5.8 (Darwin 9.8.0, Leopard)	PowerPC	Darwin 9.8.0, 32 bit	4.3.3/-	sas_11.0.0-Darwin-9.8.0-PPC-32.tgz	Leopard on PowerPC 32 bit. On Intel 32 bit, translated by Rosetta
Mac OS X 10.6.6 (Darwin 10.6.0, Snow Leopard)	Intel	Darwin 10.6.6, 32 bit kernel	4.3.3/-	sas_11.0.0-Darwin-10.6.0-32.tgz	Snow Leopard on Intel 32 and 64 bit.

Figure 1 - SAS 11 binaries released

With this large number of different platforms, we cover the necessities of the vast majority of potential SAS users. The disadvantage of this approach is that the large integration efforts cannot be maintained in the very long term, especially not once the mission is over.

The Processing Pipeline System (PPS) is a subset of the SAS. It is basically a set of Perl scripts, running SAS tasks with fixed parameters for the creation of a variety of data products. The PPS products for individual pointed observations include summary information, cleaned event lists, position and brightness of detected sources, high-level data products (eg. images, spectra and time series), catalogue cross-correlation information and quality information. Slew observations, data taken during slews between pointed observations, also get reduced to images by a separate branch of the pipeline system.

The PPS products populate the XMM-Newton Science Archive (XSA), used for distributing them first to the proposers and after one year for making them public.

II.1 Lifetime of SAS and PPS

The general experience shows that the usage of scientific data goes on for a long time after the mission / experiment producing them stops operating. Depending on several factors, this time can range from months to decades. The main factors are volume and quality of data, level of exploitation, and also accessibility and reliability. But also the ability to handle data in an interactive way will influence the level of confidence the potential users have in them and can also extend the time they are able to extract the best out of the data.

In the case of XMM-Newton we estimate that potential time of SAS capabilities usage after the mission ended to be one to two decades, based on all the factors mentioned. Concerning the pipeline system it will probably be essential to reprocess all the data obtained by the mission once it is over, using the same software and calibration, and also to obtain the final catalogue of serendipitous sources, which could take approximately one year.

III - THE REMOTE INTERFACE FOR SCIENCE ANALYSIS (RISA)

RISA is a client/server application, providing a web-based interface to the XMM-Newton data and data processing system. It wraps, in an easy and flexible way, the functionalities of the XMM-Newton Science Analysis System (SAS) and allows the user to perform XMM-Newton data analysis without the need to install any software. It works asynchronously such that a user chooses the datasets (s)he wants to analyse, defines the flow of tasks that (s)he wishes to perform, submits them and receives the answer later, either in an e-mail or by polling a job status web-page. The SAS tasks are executed in a GRID environment, using GridWay as the Grid meta-scheduler.

RISA allows scientists to discover, download and reduce data on-the-fly, using all XMM-Newton SAS capabilities (including the parameter interface and the image selection expressions). It has been coded in Java, using AJAX and SOAP technologies, and taking into account a main Virtual Observatory (VO) paradigm: "move results instead of data". VO protocols are used, such as SIAP (Simple Image Access Protocol) for communication with the XMM-Newton archive, or SAMP (Simple Application Management Protocol), used to communicate with VO tools like ds9, Aladin, VOSpec, etc.

Starting from a web-start application (see Fig. 2), the system allows the user to search for any XMM-Newton data (pointing or slew observation) using SIAP protocols and the CDS name resolver services. After processing the SIAP response, RISA client creates a main window listing all the XMM-Newton observations available according to the request. The user can then create the SAS workflow(s) and, by submission the client will serialize the information and send it to the RISA server. This will dispatch the job(s) to the Grid through DRMAA OGF standard (GridWay implementation). Each node in the Grid makes a request to the XMM-Newton Science Archive (XSA) to retrieve the XMM-Newton data set, corresponding to a given observation ID, using the AIOClient application, a tool provided by ESAC archives to automatically download data from the archive.

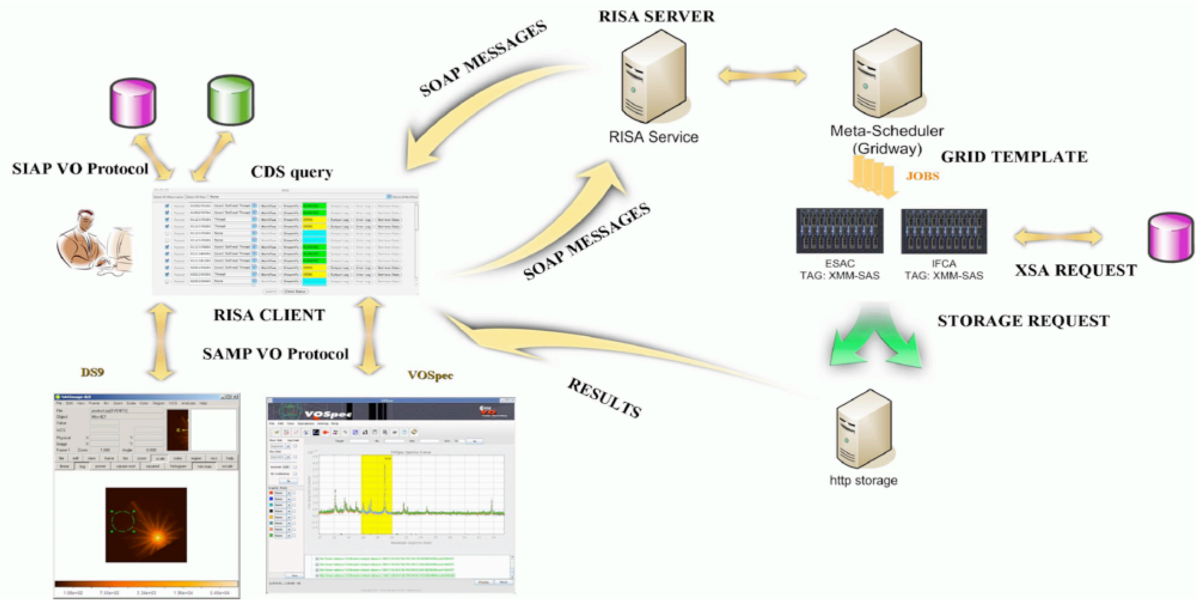


Figure 2 - RISA architecture scheme

The user can create fully configurable tailored workflows, or select pre-defined workflows that automatically produce XMM-Newton images, spectra, light curves, and/or source lists. Either way the whole parameter space of all the single elements of the data reduction are accessible through context sensitive graphical interfaces. Once the data processing has finished, GridWay takes the results to the Storage Element automatically. The RISA service knows then that the observations have been processed and informs the client ("job DONE"). The results can be displayed with VOSpec, ds9 or Aladin, reached directly through the SAMP protocol, i.e., without having to download the data. The RISA client is able to work starting from raw XMM-Newton data or from pipeline pre-processed files. Finally, the user can retrieve the data when the jobs have finished.

III.1 ADVANTAGES OF THE RISA APPROACH

RISA offers several advantages both for users and for providers compared to the plain, traditional use of SAS. On the users side the necessity for downloading S/W, data and calibration data disappears. Upgrades of the package or of the calibration are done on the server's side, fully transparent to the users. Processing in semi-batch mode large amounts of data is then made very easy.

On the system side the advantage of reducing the integration to a single platform is evident. Furthermore, there is no need of continuous evolution anymore with evolving operating systems, libraries, compilers. The system can remain frozen for long periods of time, reducing the maintenance costs to a minimum.

Due to its nature, RISA is also a perfect complement for performing "on-the-fly-reprocessing", an added capability to the XMM-Newton archive for reducing a data set using the most up-to-date SAS and calibration available, but also giving all the flexibility to optimize the analysis.

Another important point to be taken into account is the scalability of a system which is going to be required by a fluctuating number of users, especially after the XMM-Newton mission is over. Performing the data reduction in a Grid environment provides such scalability by default. Using several clusters geographically distributed as different nodes of a Grid (see [3]) even increases the reliability. RISA's architecture is such, that operating it in the Cloud would also be trivial to implement.

While we expect to increase the level of interactivity within the system (see IV-1 Future of RISA), the fact that the workflows are dispatched to the Grid for execution makes RISA not especially suited as an interactive tool, but more as a system prepared for creating mini-pipelines. The use of own scripts by a user is in principle incompatible with the use of the Grid. Therefore RISA cannot interact with other data analysis packages (eg. IDL, IRAF), something used often by advanced SAS users through scripts combining calls to different packages. This has been seen as a shortcoming of a system like RISA. To further give users the possibility of using their own scripts and interacting with other local installed systems, we see the SAS Virtual Machine (see next section) as the ideal complement to RISA.

III.2 THE SAS VIRTUAL MACHINE, A COMPLEMENT TO RISA

A virtual machine, as a tightly isolated software container that can run its own operating system and applications as a physical computer, offers several advantages for long term maintenance, on top of the recognized advantages like compatibility, isolation, encapsulation and hardware independence.

We have started to distribute SAS virtual machines back in 2007, at that time exclusively for MS Windows users. The first version contained a Linux Fedora Core 4 virtual machine including SAS 7.0 (built on Red Hat Enterprise Linux 4) and all the external software tools required to work with SAS: ds9, Grace, Heasoft, etc. This way we could compensate the lack of a native version of SAS for MS Windows users.

Together with our last release (SAS 11.0) we offer for the first time two universal SAS Virtual Machines, one with a Linux 32 bit and another with a Linux 64 bit kernel. They can run on Windows, MacOS X and Linux. The technology used is VMWare, well known for several years. Users do not require any license thanks to their free VMWare player products. The SAS-VM already contains, together with SAS 11, all the external packages required by SAS as well as the calibration files.

The main disadvantage of using a VM is the reduced level of resources compared to the host machine. The exponential growing of computing technology of course largely compensates this disadvantage from year to year.

The advantage from the S/W provider's side is that integration, testing and distribution can be reduced to a single platform and also in this case the possibility of freezing the analysis package for long periods of time is possible, due to the "virtual longevity" of an OS in opposition to its poor short term real life[2].

[2] The newest version of VMware fusion, offers under others a Red Hat VM (17 years old) and a Windows 3 VM (19 years old).

IV - RISA, WHERE WE ARE, WHERE DO WE GO

RISA has been internally released as Beta 1.0 version in June 2011 for testing purposes, and to get the feedback of a larger group of people about its usage. Representing a rather different approach to data analysis as the plain SAS, we need to gauge how comfortable users of the system can be, what they do miss, what they do wish to see incorporated.

The version includes full analysis capabilities with full access to all parameters, eg. the same granularity as SAS, reachable step by step through the construction of their own workflows. A configurable "pipeline" is also integrated in the system through a SAS metatask, "xmmextractor", which by default produces the entire set of products according to the exposures present in the observation, eg. event files, images, low and high resolution spectra, lists of detected sources, response matrices, etc. In this beta release the parameters are taken with their default values and cannot be modified, but in the next release it will allow to fully configure the pipeline.

Documentation is written as a series of "threads", ie. short cookbooks addressing how to reduce the data step by step to a certain product. It can be found together with the web start application under the URL: http://scilsn11.esac.esa.int/risaws/

IV-1 - THE FUTURE OF RISA

A number of improvements are planned to be implemented in RISA in future releases. They address diverse categories:

- Processing
  - full configuration of the pre-defined pipeline will be possible through a graphical interface
  - own data (eg. event files) uploading will make possible a level of interactivity within the package, through stop / start mechanisms from several levels of reduction
  - combination of data from different observations
  - light RISA services will be incorporated (mini-pipelines for fast response)
- Data persistency
  - a login mechanism will permit to check the status of previous sessions
  - in case of service failures, it will be possible to recover submitted workflows
- Data storage / Security
  - data privacy will be guaranteed
  - output data will go to a distributed storage system (VOSpace, Cloud) instead of being locally stored

On top of that, the RISA services will be integrated into the new XMM-Newton Science Archive, which is under development, for on-the-fly reprocessing of data directly running from the archive interface.

RISA is running on the Grid, but can be easily adapted to run on Cloud architectures. Furthermore, the system is not specifically built for running SAS, but is able to allocate workflows reducing data from other observatories. We plan to soon embed EXOSAT data analysis into RISA, showing that this is possible, and making the data corresponding to that mission fully accessible again.

V - CONCLUSION

RISA, the Remote Interface for Science Analysis, which makes it possible to run SAS through fully configurable web service workflows, enabling observers to access and analyse data making use of all of the existing SAS functionalities without any installation/download of software/data, represents a way to ensure that XMM-Newton data analysis capabilities are preserved for a long time after the end of the mission operations. In addition to further distribution of a SAS Virtual Machine, we expect to cover all data reduction necessities for a maximization of the scientific exploitation of the XMM-Newton data for decades.

REFERENCES

[1] - Gabriel C. et al. 2004, in Astronomical Data Analysis Software and Systems XIII, ed. by F. Ocshenbein, M.G. Allen and D. Egret (San Francisco, CA: ASP), Vol.314 of ASP Conf. Ser.,759

[2] - Jansen F. et al. 2001, A&A, 365, L1

[3] - Ibarra A. et al. 2006, in Astronomical Data Analysis Software and Systems XV, ed. by C. Gabriel, C. Arviset, D. Ponz and E. Solano (San Francisco, CA: ASP), Vol.351 of ASP Conf. Ser., 520