SIPAD-NG: a multi field system for archiving and accessing space data. Feedback on ten years of development and operational use

Aurélie Bellucci (1), Dominique Heulet (1)

(1) Centre National d’Etude Spatiales (CNES)
18, Avenue Edouard Belin, 31401 Toulouse Cedex 9, France
EMail: aurelie.bellucci@cnes.fr, dominique.heulet@cnes.fr

ABSTRACT

SIPAD-NG (“Information System for Data Preservation and Access – New Generation”) is a generic software providing three of the main OAIS functionalities: “data ingest”, “data management” and “data access” (the “data storage” function is provided by online storage or by the CNES STAF facility). SIPAD-NG development began in 2003. Thus, the paper reports on ten years of system evolutions and operational use by different space data centres or mission centres. Both its data model and its architectural design have proved the ability of SIPAD-NG to adapt to the needs of several projects and thematic: plasma physics (CDPP archive), operational oceanography (MERCATOR), altimetry (AVISO), microgravity experiments (CADMOS), clouds and aerosols (ICARE), ocean humidity and salinity (SMOS), launchers simulation data (MINOS). The paper presents the methodology and tools developed around SIPAD-NG to help data centre teams in configuration and use of the system. It also presents the evolutions of the “data access” function to follow the constant progress of web technologies and interoperability standards as well as the solutions adopted to avoid major impact on “data ingest” and “data management” functions.

Keywords: Data preservation, OAIS, Data access

INTRODUCTION

The SIPAD-NG project began 10 years ago in 2003 from the feedback of former SIPAD system. The original architecture and functionalities have grown following the needs of an increasing number of projects using SIPAD-NG and addressing various topics. This paper presents the main steps of this evolution and the methods drawn up along years.

ORIGIN and CREATION of SIPAD-NG

The development of SIPAD-NG started in 2003 as the former system, SIPAD, was reaching its limits. Thus, SIPAD-NG specifications were based on SIPAD feedback. Software architecture and technologies have changed in order to easily adapt the system to various scientific topics. Indeed, the aim of the SIPAD-NG system is to avoid recurrent development of data access systems in Mission Centers or Data Centers funded by CNES. In particular, SIPAD-NG is intended to be used by scientific data centers whose long-term mission is to regularly expand the list of data available to a user community.

SIPAD-NG has been designed with a highly modular architecture (see Figure 1). Provided that interfaces are not modified, it is possible to change a component without affecting the rest of the system. The architecture of SIPAD-NG is composed of:

- **Six basics services** which implements the functionalities of any data access system. This is the core of the system. An instance of SIPAD-NG (a set of services) can manage one or more projects.
- **consultation**: search for data by browsing the catalogue and applying selection criteria
- **user management**: centralization of all the information about users, access rights and quotas
- **command**: extract data from archive, eventually processing data with specific tools (added-value services) and deliver the data to the users via FTP or media
- **userworkspace**: provide to users a personal workspace from which they can retrieve the ordered data
- **ingestion**: enrich the catalog (data, documents, graph) using meta-data provided in XML files
- **administration**: administration and supervision of the system.

- **a database layer** using Oracle Relational Database Management System. This part is comprised of
  - a small database in charge of managing the instance
  - a data catalogue for each project managed by the instance.

- **client applications** such as
  - user and administration web servers (Apache/Tomcat servers with Struts and Java)
  - monitoring client, delivery client, statistic client, …

The data presented by an instance of SIPAD-NG are archived
- **on disk** (online archive)
- **by the CNES STAF¹ service** (near line archive). STAF provides a preservation service for data files using tape libraries (StorageTek). This service takes in charge the regular copy of archived files on new devices on behalf of spatial projects teams.

The first SIPAD system was used by two projects: CDPP² and Mercator (Oceanography). Those two projects were the first to use the SIPAD-NG system. For CDPP, the migration to the new data model induced a long work to add new metadata into the catalogue before opening the service to users. For Mercator, the migration implied a complete reorganization of the catalogue leading to two data-access systems to answer the need of the scientific community.

---

¹ STAF: Service de Transfert et d’Archivage des Fichiers
² CDPP: Centre de Données de la Physique des Plasma - French data center for Plasma Physics
ACQUISITION and ARCHIVING

The SSALTO\(^3\) multimission ground segment encompasses ground support facilities to control CNES altimetry missions (JASON 1&2, SARAL, …), to process data and to provide user services and expert altimetry support. In 2008, the SSALTO project decided to develop a new data access system to replace the former Aviso catalog. The Aviso catalog wasn’t meeting the needs of the project, especially to constitute a reliable archive of altimetry data.

With the six original services presented before, SIPAD-NG was able to fulfill the requirements for data-access. To meet the need to constitute a centralized archive for SSALTO data, we develop two new services: the acquisition and archiving services.

- **acquisition**: This service is in charge to collect data files and transfer them to a local online archive. If necessary, this service produces the XML files containing the metadata describing those data. Those XML files will later be given to the ‘ingest’ service to update the catalogue. Specific XML files called ‘supplies’ indicate how to manage the collected data: collect location, archive location, dataset receiving this type of data, methods to recover metadata from filenames or inside files using specific plugins,…

- **archiving**: This service is in charge to transfer data files from the local online archive to the STAF service to constitute a perennial archive. Once transferred to STAF, the files in the local online archive are deleted after a retention period defined for each type of data.

The creation of those two services induced the addition of a new data base, specific to those functionalities. However, an instance of SIPAD-NG can only manage only one project using the acquisition and archiving functionalities. It also induced the creation of new clients to start acquisition and archiving tasks, to update data catalogue with XML files produced during acquisition and to purge the local online archive. The architecture of SIPAD-NG including those new functions is presented in Figure 2.

The SSALTO SIPAD-NG opened in 2009 after a tedious migration work from Aviso catalogue and the production of supplies for the various products managed by the SSALTO project.

**SIPAD-NG and the OAIS model**

With those eight services, the SIPAD-NG system is compliant with the main functionalities of the OAIS model (Open Archival Information System) for long-term archives of numerical data and their managing systems. Indeed, the six basic services of SIPAD-NG fulfill the ‘data management’ and ‘access’ functions of the OAIS model. The acquisition and archiving services fulfill the ‘ingest’ function of the OAIS model. The storage function is provided by STAF service (see above) or by projects online archives. The connections between the OAIS model and the SIPAD-NG architecture are presented in Figure 3.

---

\(^3\) SSALTO : Segment Sol multimissions d'ALTimétrie, d'Orbitographie et de localisation précise
Figure 2: SIPAD-NG architecture with acquisition and archiving functions

Figure 3: OAIS model and SIPAD-NG components
OTHER PROJECTS USING SIPAD-NG

CADMOS project

Acquisition and archiving services were developed as generic components of the SIPAD-NG system. Thus, they were quickly used for other projects than SSALTO. In 2010, CADMOS⁴ expressed the need for a reliable archive along with a data access system for results of experiments carried out on-board the International Space Station (ISS). Two data access systems have been developed to separate the results of experiments addressing different topics:

- **Physio**: physiological data resulting from experiments on astronauts on-board ISS. Confidentiality is one of the critical point of this project
- **DECLIC⁵**: Results from this material physics experiment are returned to Earth on a RHD (remote hard disk) during a return flight of Soyouz or SpaceX spaceships.

This is an example of a SIPAD-NG instance managing two projects with their own data model, data catalogue and web interfaces but sharing the use of SIPAD-NG services.

For DECLIC experiment, a specific set of tools was developed to elaborate a complete procedure from the reception of data until they are made available through the SIPAD-NG system. First, a tool relying on EAST⁶ data description decommutes data received on disk and re-organize data in zip files. The acquisition and archiving services are then used to add those data to the catalogue and archive them in the STAF.

For Physio, data are delivered as ZIP files ready to be archived. Furthermore, each dataset only contains one ZIP files. Thus, the acquisition and archiving services don’t meet the needs for this project and a specific small tool was developed for it. Indeed, acquisition and archiving functions are well fitted for projects for which new data are regularly delivered to be distributed in various datasets.

MINOS project

At the same time, the DLA⁷ decided to constitute an archive of simulation results related to space launch systems. However, the need of this community (MINOS project) was slightly different from other projects. For the MINOS project, users are in charge of adding their data into the system, whereas for other projects this task is under the responsibility of the administration team (also in charge of managing users, access rights, availability of services, …) Thus, we had to develop a user interface to allow classic access to data of the catalogue as well as addition of data into the catalogue. Such an interface must include a filling form in which the user indicates the metadata related to the data he wants to add. The interface must also be able to upload files. It was difficult to add those functions in the classic user interface while preserving the ergonomics of the web site. This was the opportunity to develop a new type of user interface using RIA (Rich Internet Application) technologies.

The RIA Minos server was dedicated to this project. It offers the needed functions (browsing the data catalogue, adding new datasets and data …) but not all the functionalities available in the default user interface used by other projects. For instance, in the first RIA Minos server, the user workspace didn’t exist. The only way to recover data was direct download, thus with a size-limit on recoverable data.

The MINOS web server was a first step towards the full use of RIA technologies in a default user interface including all the functionalities available in the classic interface. This is the last evolution of SIPAD-NG described later in this paper.

---

⁴ CADMOS : Centre d’Aide au Développement des activités en Micropesanteur et des Opérations Spatiales
⁵ DECLIC : Dispositif pour l’Etude de la Croissance et des Liquides Critiques
⁶ EAST : Enhanced Ada SubseT : standard using ADA language to precisely describe the format of data files
⁷ DLA : Division des LAnceurs – CNES division for space launch systems
Feedback on ten years of use

After 10 years of use, we can say that SIPAD-NG is a well-tried system which has proved its ability to adapt to various scientific and technical topics without changing the core of the system but just configuring services and client according to project requirements. The following table presents the list of projects using the SIPAD-NG system with the amount of data available.

<table>
<thead>
<tr>
<th>Hosting entities</th>
<th>CNES</th>
<th>Lille university</th>
<th>IFREMER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Projects</td>
<td>CDPP</td>
<td>Mercator</td>
<td>SSALTO</td>
</tr>
<tr>
<td>Thematic</td>
<td>Plasma Physics</td>
<td>Oceanography</td>
<td>Atmosphere</td>
</tr>
<tr>
<td>Status</td>
<td>Operational</td>
<td>Operational</td>
<td>Operational</td>
</tr>
<tr>
<td>Nb of datasets</td>
<td>681</td>
<td>191</td>
<td>355</td>
</tr>
<tr>
<td>Nb of objects</td>
<td>2 642 063</td>
<td>487 000</td>
<td>3 532 233</td>
</tr>
<tr>
<td>Nb of quicklooks</td>
<td>573 480</td>
<td>1</td>
<td>1 056 544</td>
</tr>
</tbody>
</table>

Table 1: List and usage of SIPAD-NG systems

METHODS TO ADAPT SIPAD-NG TO A PROJECT

After 10 years of development and adaptation of SIPAD-NG system for various projects, it appears that the method used is nearly the same for every project.

Data Model

First step is to create a prototype. With people in charge of the project, we examine the type of data which will be accessed and define the way the data will be displayed to users. That means defining with them collections and datasets relevant to the subject and the users of the scientific community. A data access system in which the organization of data is not understood by the target audience will not be used. We also defined with them the selection criteria that will be available for each dataset to retrieve the wanted data.

With this information, we create a dictionary (an XSD schema) specific to the project. This dictionary contains meta-entities forming the data model of SIPAD-NG. Each meta-entity contains meta-attributes mandatory for any SIPAD-NG and meta-entities specific to the project, especially for datasets and data objects. For instance, the selection criteria are used to define the meta-attributes of the data-objects. After a few iterations with the project team, the data model is ready.

Customization of User interface

The next step is to adapt the default user interface to the project. Indeed, the appearance of this web-site can be easily customized by modifying the colors, police and logo in css files and jsp. Usually, we give to this web-site an appearance consistent with the other web-sites of the project, especially for the choice of colors and images. The content of the welcome page is defined in an html page. As a consequence, to display a temporary message on the welcome page, for maintenance announcement for instance, the administrator of the project just needs to modify an html page. A text-description can be added for each data-set. Those descriptions are provided by scientific teams but changing the datasets description page on the web server requires the use of the ingestion service of SIPAD-NG. Thus, this operation is reserved to development team or trained administrators. Once more, this customization work is done in collaboration with the project team so that the prototype corresponds to their requirements.

This method is presented as a schema in Figure 4.
Tools to fill the catalogue

With the data model and the customized user interface, the data access system is now ready. The catalogue must now be filled with data. Various methods can be used for that task. If new data are delivered regularly to be distributed in existing datasets, the most efficient way to manage the data is to use the acquisition and archiving functions. That’s what is done for the SSALTO and DECLIC projects. However, there are projects for whose data are already archived in STAF (SERAD project) or stored on specific disks (Smos). In those cases, we usually developed some specific tools (usually shell scripts) to help project team to add data in the catalogue. For instance, for the Serad project, we developed a script which scans the content of the STAF, creates the XML files describing the data files, and ingests those descriptors files so that data are visible and available through the SIPAD-NG Serad server. For Smos, a Java tool was developed to compare every day the content of the SIPAD data catalogue with the content of the online archive. This tool produces the appropriate descriptors files to add new data into SIPAD-NG catalogue, update modified data or remove data deleted from the archive. A shell script is then used to ingest all those descriptors files.

Along the 10 years of use of SIPAD-NG, several small tools have been developed. They constitute a library called SipadLib. The specific tools developed to answers the needs of our various project usually used SipadLib’s tools.

PERSPECTIVES

RIA technologies and REST interfaces

The SIPAD-NG system is in constant technical evolution. Since summer 2013, the system includes a default user interface developed using RIA technologies. This evolution results from the feedback of the first MMI of this type developed for MINOS project. A default user interface is now available for any new
The CDPP project is the first project to benefit from this new interface. Then, the specific RIA interface developed for MINOS project will be replaced by the default RIA interface, customized with MINOS appearance and functionalities. Progressively, every existing project will have its own RIA user interface. The administration interfaces doesn’t change, the classic model is still used.

For an efficient use of RIA technologies, REST interfaces have been developed for the services that constitute the core of SIPAD-NG system. The core of the services is not changed, but two interfaces are now available for communication between services and client applications: RMI and REST. Progressively, the RMI interfaces will disappear.

**Oracle 11 and other database management system**

SIPAD-NG system requires Oracle for the database management system. Presently, Oracle 10 is used but upgrade to Oracle 11 is already in preparation.

The choice to use Oracle was done ten years ago because SIPAD-NG was developed for large communities able to fund this choice. However, in order to decrease the cost of the system, a transition to PostgreSQL has been studied. This evolution is possible without modifying the data model which is the core of the system. It should be one of the main changes planned for the years to come.

**CONCLUSION**

After ten years of use, SIPAD-NG has proved its performance and its ability to follow the requirements of various projects using the system. Since the first version, SIPAD-NG has undergone many changes. There were important additions in the initial architecture (acquisition and archiving functions) and some components have been completely modified (command service in 2010). We had to cope with performance issues for projects with much more data than expected by specifications. However, the initial architecture (separation of services, client applications …) is flexible enough to meet the needs of various scientific communities and is able to develop without affecting all the components. The data model is suitable for more subjects than what was initially imagined, such as the technical data from MINOS projects. It also appears that developments required by a project are often useful for other projects (acquisition and archiving re-used for DECLIC…). Therefore, we have a robust system used by about ten projects with a global cost lower than the developing costs for specific data-access system for each of those projects. Thus, the initial target set ten years ago has been reached.