THE PRESERVATION AND VALORISATION OF SCIENTIFIC DATA BY THE 'CENTRE DE DONNÉES DE LA PHYSIQUE DES PLASMAS' (CDPP, PLASMA PHYSICS DATA CENTRE)

Michel NONON-LATAPIE [1], Isabelle BESSON [2], Michel GANGLOFF [3]

[1] Centre National d'Etudes Spatiales
Bpi 1502, 18 avenue Edouard Belin,
31401 Toulouse Cedex 4
michel.nonon@cnes.fr

[2] CS Systèmes d'Information
ZAC de la Grande Plaine, Rue Brindejonc des Moulinais, BP 5872,
31506 Toulouse Cedex 5
Isabelle.Besson@c-s.fr

[3] Centre d'Etudes Spatiales des Rayonnements
9 avenue du Colonel Roche, BP 4346,
31028 Toulouse Cedex 4
michel.gangloff@cesr.fr

The CDPP results from a joint initiative of the CNRS (Centre National de la Recherche Scientifique) and the CNES (Centre National d'Etudes Spatiales). Its principal objectives are to ensure the long term preservation of data relevant to the physics of naturally occurring plasmas, to render this data easily accessible, and to encourage its analysis. The data is produced by instruments, in space or on the ground, which study the ionised regions of space near the Earth and elsewhere in the solar system.
The principal users of this data centre are space scientists, wherever they are located.

This data centre is located in Toulouse (France), and it uses a computer system which is accessible via the Internet (http://cdpp.cesr.fr/english/index.html). This system offers several services : firstly the possibility to search for and retrieve scientific data, but also access to the "metadata" archived in association with this data, as the relevant documentation and quicklook data (graphical representations). Several tools are available to help the user to search for data.

The CDPP has been accessible since October 1999. Since then its data holding and the services offered have been steadily augmented and developed.

After a brief presentation of the objectives, the organisation, and the services currently offered by the CDPP, this paper will concentrate on :

- the system architecture (based on the ISO "Reference Model for an Open Archival Information System")
- the data model
- the standards used to format and describe the archived data

We will also present how the computer system, first developped to meet CDPP requirements, has been since reused and adapted for other operational data servers.

We will finally present lessons learned concerning the operation, maintenance and evolution of the current system, as well as new technical developments under study.

1. PRESENTATION OF THE CDPP

The "Centre de Données de la Physique des Plasmas" (CDPP, Plasma Physics Data Centre) is the result of a collaboration between the "Centre National de la Recherche Scientifique" (CNRS, French National Scientific Research Centre), the "Institut National des Sciences de l'Univers" (INSU, French National Institute for the Sciences of the Universe), and the "Centre National d'Etudes Spatiales" (CNES, French Space Agency).

The main mission of this Thematic Centre is to ensure long term preservation (over several decades), accessibility, and valorisation of data acquired by French experiments or by experiments involving French participation in the Natural Plasma Physics field. It provides a service for the international scientific community. It is located in Toulouse (France), on two main sites, one of which is on CNES facilities, while the other is in a CNRS laboratory, the "Centre d'Etude Spatiale des Rayonnements" (CESR).

A team composed of personnel from the CESR and CNES is responsible for the following activities to accomplish these missions:

- coordination and interface with the scientific teams responsible for the data and the information associated with this data, preparation of the archive plans with these teams,
- support in preparing data and metadata to be archived,
- operational maintenance and administration of an information system dedicated to the CDPP, the "Système d'Information, de Préservation et d'Accès aux Données" (SIPAD, System for Preservation and Access to Data and Information),
- definition, prototyping and implementation of value added services for archived data,
- contacts with the scientific community: presentations in various seminars and meetings, six-monthly bulletins, etc.

2. THE SIPAD

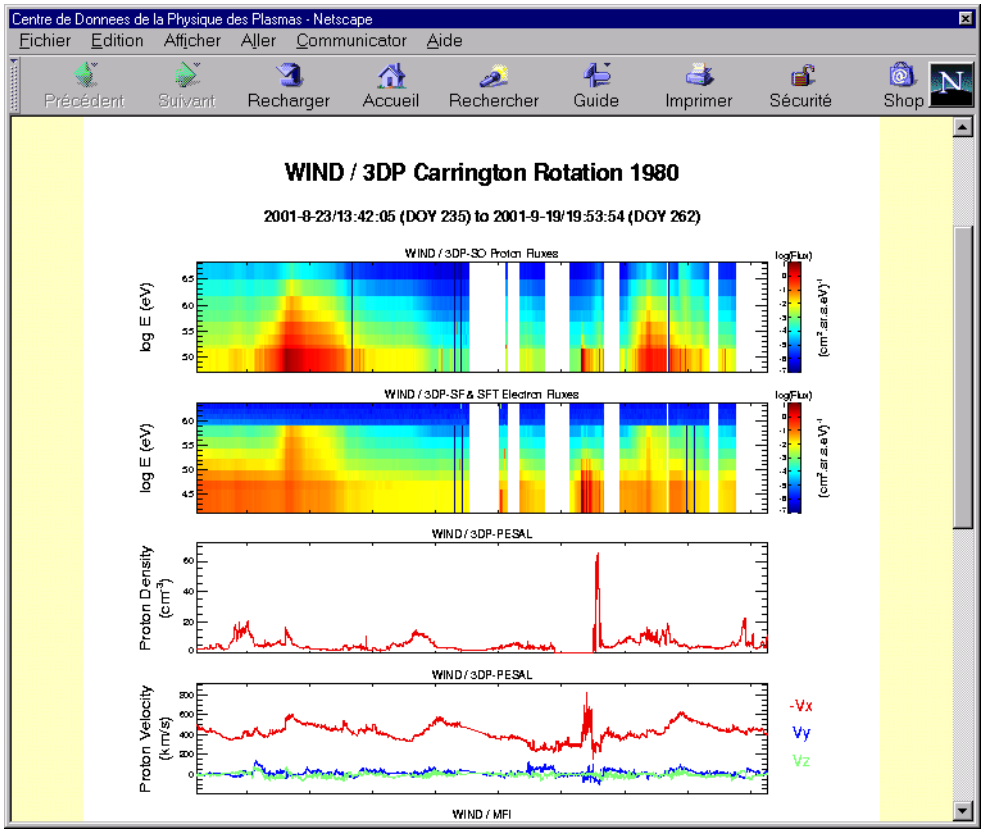
2.1 PRESENTATION OF SERVICES OFFERED BY THE SIPAD

The "Système d'Information, de Préservation et d'Accès aux Données" (SIPAD) is a computer system originally developed to meet CDPP requirements, i.e. archiving of and access to Plasma Physics data. This data is essentially time-dependent series of measurements made by satellite on-board instruments or ground instruments. This is accompanied by "metadata", for understanding and using data. For example the data is accompanied by documentation describing the mission (objectives, scientific teams, etc.), the experiment (measurement type and principles, etc.), and also syntax and semantics for each data file archived.

This system has since been adapted and re-used for other projects as will be described in more detail in § 2.4.

Any user of this system uses a Web browser to:

- access data, and its documentation, using several research tools:
    - overview of all data archived in the CDPP,
    - navigation within the mission/experiment/data set hierarchy,
    - selection by keyword(s),
    - selection of time period(s) by direct entry, by viewing browse data (predefined graphics, see example below), or by using event tables,



- order data:
    - by selecting files in the archive according to time-dependent or other criteria (orbit number for example),
    - by applying different transformations as necessary, on the archive files selected:
        - eliminating from the archive files any data not strictly corresponding to the time-dependent criterion selected, or parameters not selected by the user, in order to reduce the size of the files to be retrieved and to better carry out the user request,
        - (future) more specific transformations, i.e. for example converting raw data into physical values, making transformations between different coordinate systems, or changing data formats,
    - by selecting a means of delivery:
        - availability on CDPP disk space (with multiple compression options),
        - FTP delivery on a remote machine,
        - delivery on CD-ROM or magnetic media (DAT, Exabyte); CD-ROMs have a specific CDPP serigraphy.

SIPAD also provides pedagogical information on plasma physics, a scientific discipline which is often not well known or not covered in the media, and general information designed for the scientific community: significant events (seminars, etc), useful Web addresses, etc.

SIPAD too provides access to MAGLIB, a portable, well documented software library dedicated to the discipline.

The part of the SIPAD dedicated to data and metadata access was developed by the "CS Systèmes d'Information" company who is also responsible for its corrective and evolutive maintenance from its entry into service.

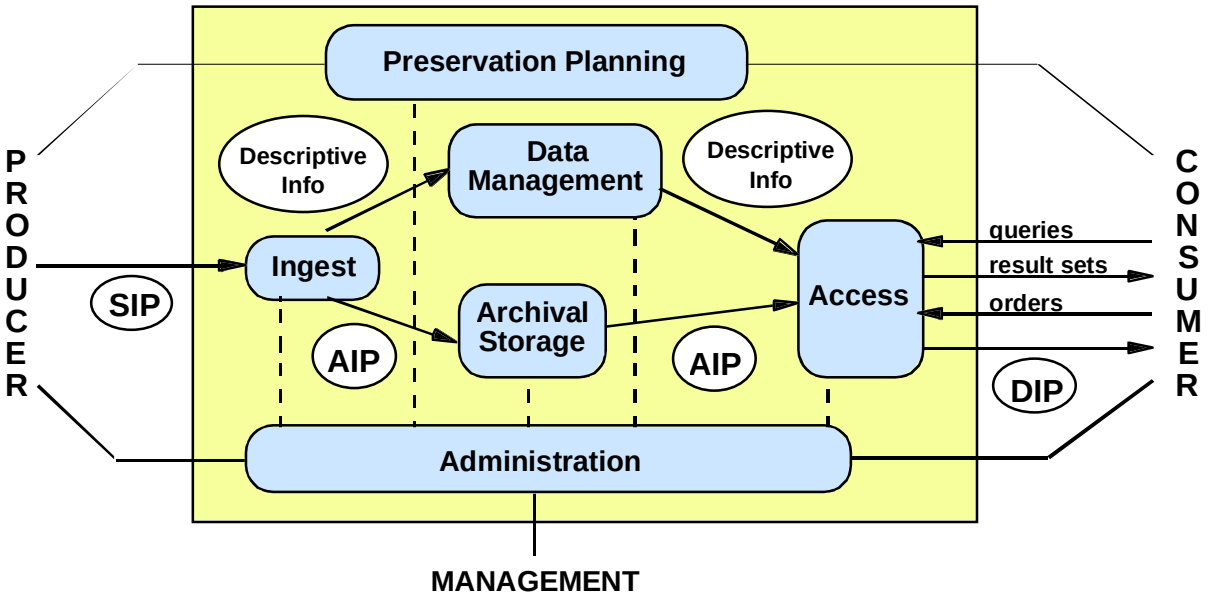
SIPAD (and therefore CDPP services) can be accessed via Internet, since October 1999, at the following address http://cdpp.cesr.fr.

About a hundred users are registered.
Approximately 200 sets of data and "quicklooks" are currently available, which represents around 350000 accessible files.
These figures increase regularly (as new data acquisitions and registrations are made continuously).

2.2 SIPAD SOFTWARE ARCHITECTURE

The SIPAD system is deployed on several machines, in the CESR and in CNES (Computer Centre in Toulouse).

Its architecture is based on the functional model of the ISO standard "Reference Model for an Open Archival Information System (OAIS)" (standard undergoing finalisation, ISO/DIS 14721.2).



The storage function in particular comes under the responsibility of a CNES Computer Centre facility, the "Service de Transfert et d'Archivage de Fichiers" (STAF, File Archiving and Transfer Service – refer to Anne Jean-Antoine Piccolo's paper on this Service), which makes media management and file storage management on these media transparent for its users. This is one of the main functions ("Archival Storage" function) defined philosophically in the OAIS model, and put in practice for the SIPAD architecture.

The SDRC company Metaphase product (this product is an object-oriented technical data management software) was selected to manage the SIPAD database: metadata and file references (data, document elements, graphic representations, etc.) archived on disk or on the STAF, and also information on users and their data orders, etc. Thus this base is the core of the system (OAIS "Data Management" function).

2.3 THE MANAGED DATA MODEL

A thematic centre database must have the capacity to acquire information and data from different missions, delivered by different teams at various times.
Furthermore the SIPAD, designed for plasma physics requirements, is also used for other scientific disciplines.
The acquisition function for this type of system (the OAIS "Ingest" function referred to in the previous paragraph) must be both generic (multiple suppliers), evolutive (new types of information must be envisaged depending on the data and services which become available), and rigorous (systematic input checks, so as not to discover errors after information is managed in the database).

To meet these apparently contradictory objectives, a modelling phase was planned to define the data model which the SIPAD had to manage and also, in correspondence with this model, a dictionary of the entities which may be delivered to the SIPAD. This dictionary drives the acquisition function: it is updated as necessary (add a new entity or add a new possible value for an existing entity, modify entity cardinality, etc.), and the acquisition function makes checks according to what is expected in the dictionary.

To illustrate this principle, here is a real example, from the current CDPP SIPAD dictionary, referring to the description of a document to be delivered and made available on the server.
A document must be delivered to the SIPAD in compliance with the following dictionary definition:

```
BEGIN_GROUP = ENTITY_DEFINITION;
        NAME = DOCUMENT_DESCRIPTION;
        MEANING  = " Description d'un document";
        COMPONENT = (DOCUMENT_IDENTIFIER ,1..1);
        COMPONENT = (TITLE ,1..1);
        COMPONENT = (AUTHOR ,0..1);
        COMPONENT = (ORIGIN ,0..1);
        COMPONENT = (DOCUMENT_TYPE ,1..1);
        COMPONENT = (ENGLISH_CONTENT ,0..1);
        COMPONENT = (FRENCH_CONTENT ,0..1);
        COMPONENT = (GLU_TAG ,0..1);
        COMPONENT = (ABSTRACT_FILE_NAME ,0..1);
        COMPONENT = (ISSUE_REVISION ,0..1);
        COMPONENT = (LAST_VERSION_DATE,0..1);
        COMPONENT = (SENDER_SERVICE ,0..1);
        COMPONENT = (MAINTENANCE_CONTACT ,0..1);
        COMPONENT = (MAINTENANCE_GLU_TAG ,0..1);
        COMPONENT = (PAGE_NUMBER ,0..1);
        COMPONENT = (FILE_NAME ,0..N);
END_GROUP = ENTITY_DEFINITION;
```

Which gives, for the description of a specific document delivered to SIPAD:

```
BEGIN_GROUP =  DOCUMENT_DESCRIPTION;
  DOCUMENT_IDENTIFIER=PLAS-DIF-INTER_MEMO-00268-LPC;
  TITLE="INTERBALL Auroral satellite : AC Electric and Magnetic Waves (50 Hz - 240 kHz) High Resolution Snapshots : Raw
Data";
  AUTHOR="Rauch,J-L.";
  DOCUMENT_TYPE=DIF;
  ISSUE_REVISION="Issue 01- Rev 00";
  LAST_VERSION_DATE=2001-04-22;
  SENDER_SERVICE=SIPAD;
  FILE_NAME=00268.htm;
 END_GROUP =  DOCUMENT_DESCRIPTION;
```

2.4 REUSING THE SIPAD

The SIPAD was developed to meet CDPP requirements. However part of the software requirements and some design and implementation choices, like the description of all the information delivered by a dictionary, allowed it to be relatively easily reused in two non-CDPP contexts.

It was first re-used to meet ETHER project requirements, in collaboration with the Pierre-Simon Laplace Institute (IPSL), in the atmospheric chemistry thematic domain. This SIPAD re-use required the addition of specific features, such as the geographic criterion (selecting files in the archive from latitudes and longitudes chosen by the user), and setting up adapted management of source software, while avoiding the duplication of common software.

It was re-used for the second time to meet Multi mission Access Service (SAM) requirements whose aim is to make data from any CNES project available on the Internet, on condition that the data is archived in the STAF and described according to the SIPAD input interfaces. The SAM currently mainly hosts and makes available data from two projects:

- CLUSTER (plasma physics field, this is a major European Space Agency mission, composed of 4 satellites currently in orbit around the earth), for the duration of the mission,

- MERCATOR (operational oceanography field, to analyse and predict ocean conditions around the globe: temperature, salinity, sea level, etc.), in the context of a prototype for access to Mercator products archived in CNES (input and output products of the models used, and research and development data).

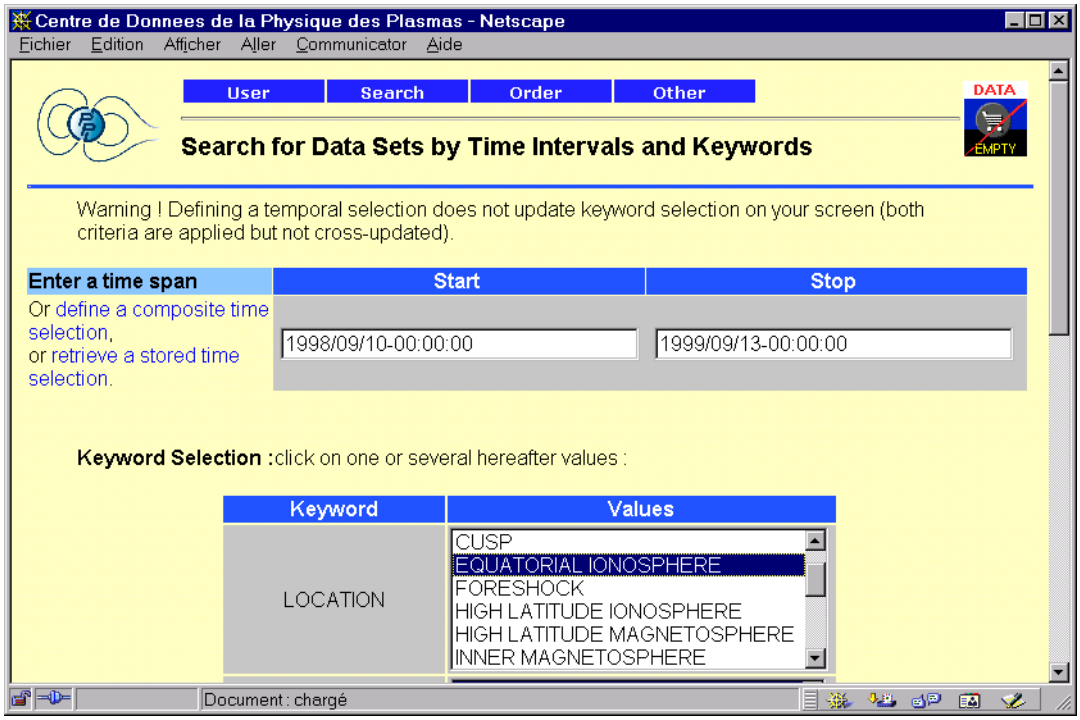
To meet SAM requirements, the SIPAD was re-used by integrating several evolutions such as the automation of large volume data acquisitions, and the "transparency" of use of a genuine multi project service: each Internet user thinks he is only accessing the server of "his" project (specific interface look and terms used, etc.).

3. VALORISATION OF DATA BY THE CDPP

Above and beyond responsibility for its long-term archiving, the data available in the CDPP is valorised by:

- the associated information archived in addition to experiment data: auxiliary data (orbit), quick look images (standardised resolutions for comparing different missions), document elements, specific event tables; this additional information facilitates the selection and analysis of archived data, often by scientists who were not involved in the original experiment,

- the services and tools proposed for consulting and retrieving this information and data: multiple data selection (see example below), ability to apply on-line transformations to archived data, ability to compare quick look images from different experiments or missions, etc.; the centre’s perspectives for evolution are mainly concerned with this added value field (cf conclusion),



- the active participation of the CDPP in international interoperability activities: briefly, the issue is that one day the user will perceive the information system as a global, coherent, unique system, whereas it is really composed of several geographically separate centres, each of which archives data from different missions or experiments, in different formats, offering different value-added services. Participating in a network of interoperable centres makes each centre legitimate, and avoids data duplication, while encouraging access and therefore making this data well-known and used.

4. MAIN LESSONS LEARNED

Following three years in operation, a first evaluation of our feedback may be made. This paper presents the main lessons learned from this feedback, actually more lessons were learned than are presented here, but it would take too long to present them all.

The first lesson is that an information system like the SIPAD is subject to quasi permanent evolutions: these evolutions are responses to new requirements expressed by users, new services defined by the team in charge of the CDPP, and also justified by the evolution of its computer environment and the technologies used.

These evolutions are mandatory for a system like the SIPAD, with an unlimited life duration and with users having multiple and various use profiles.

The (Web) man-machine interface is a part of the system which is particularly subject to a significant number of evolutions to improve its ergonomics and the use of current or future services.

This constant evolution is not generally found on a data processing system for example, for which initial specifications are globally frozen for the processing operation duration.

This constant evolution requires sufficient resources, both financial and human, and a contractual relationship adapted in cases (as on the SIPAD) where the system maintenance is entrusted to an external company.

The second lesson is related to architecture limits and technological choices made (in 1997/1998). We note the major impact on the system of the omnipresent Metaphase product, which simultaneously:

- causes performance problems (inadequate time responses), particularly because the product excessively masks access to the information managed (and makes fine-tune intervention on adjustments and optimisations associated with this access difficult, which reflects on performance),
- makes necessary the use of specific programming interfaces or mechanisms, above and beyond its "data management" function (which reflects also on performances),
- does not offer bottom-up compatibility between its different versions, which represents an additional migration cost (change in programming interfaces).

The third lesson, associated with the first two, is the crucial need to have the most modular possible computer system to facilitate its evolutions (technique, cost, schedule, etc.), control and optimize performances, and limit the impact of evolutions of one part. This statement is quite close to the aforementioned OAIS objectives, but binds to make more. As such, emerging new technologies, associated with the use of the XML (eXtensible Markup Language), seem to have a great potential.

The fourth lesson is that re-using and making this type of system generic and evolutive involve a cost. Not only must these be "thought out" from the start, but they must not impose implementation choices which adversely affect performances or which excessively extend the non regression tests required in the evolutive maintenance phase.

Finally the fifth lesson, less technical than the previous ones, is that it is necessary to have sufficient availability of the teams, shared between the project, operation, maintenance, and technical prospecting activities: it is difficult to manage interruptions and priorities associated with the operational maintenance of operating services, and generally technical prospective is the activity which is shelved indefinitely … whereas we have to continue to make these systems evolve to meet new requirements, offer new services for archived data and benefit from constant technological advances in the field.

5. CURRENT AND FUTURE WORK

Above and beyond guaranteeing the preservation of and access to data, new user requests concern:

- Value-added services proposed for this data: graphical displays, delivery formats (different from the archive format), transformations between coordinate systems, conditional extractions, etc. , with additional interactivity (possible user choices, better access time to information),
- interoperability capacities with other data servers (in both directions: consult a remote server or be consulted by a remote server).

CDPP team current work is partly dedicated to these subjects: user defined graphical displays, deliveries in CDF (Common Data Format) and ASCII formats, prototype for simultaneous access to different centres (to search for data sets according to different keywords), etc.

In parallel, definition and modelling tasks for a future access system were begun, designed to replace the current SIPAD in time (within 3 years ?). Apart from the usual phases concerning the expression of user needs and the system specifications to meet them, these activities require us to perform:

- studies of emerging XML technologies: Web services (SOAP, etc.), etc. , and of how they may contribute to our type of system: design into the most autonomous services possible, which may be called up internally by the system, or externally (from another system, which may be

geographically remote), "plug in" approach to add new tools, etc. ; these studies lead us to prototype some services and the interfaces between these services,

- exhaustive feedback for first generation SIPAD, both concerning what works well and what needs to be improved; current SIPAD acting as a sort of prototype for its successor, with this prototype having the advantage of being extremely representative (diversity in the types of information managed, volume of current data bases, feedback on the operating and administration functions, feedback on effective adaptation/re-use, etc.).

The aim is to improve current negative points, such as performance and efforts in making the system evolve, while retaining the positive points and of course retrieving in the new system all the existing metadata and data available through current SIPAD.

The SIPAD has the ambition of being a real evolved high-performance information access service, as the STAF is currently the CNES data file long term storage service.
This service will be part of the effective and active CNES support for the national and international scientific community, and will be used to further valorise the data collected, avoiding becoming a data "cemetery", as other initiatives did in the past.