IAS
Institut d'Astrophysique Spatiale
Orsay

-

IDOC guidelines for Archive long term preservation

IDOC-OD-006

Préparation

	Nom et Fonction	Date
Rédacteurs	Gilles Poulleau Martine Chane-Yook	Janvier 2018
Vérificateur		
Approbateur		

Liste de diffusion

Nom	Fonction	Société

Evolutions

Edition	Date	Modifications
1.0	02/03/2016	1st draft
		translated in english
2.0	23/01/2018	Entirely rewritten
2.1	23/02/2020	Adressing comments
2.2	23/05/2021	Global strategy

SOMMAIRE

1 SCOPE OF THE DOCUMENT

This document is related to the « IDOC-OD-008 IDOC Guidelines for new services » and "IDOC-OD-004 IDOC Guidelines for Data integration" which needs to be handled first.

This document is a guideline to a customer seeking to maintain over time the availability and use of data at least as far as its initial implementation is concerned. It therefore deals with the long-term preservation of the datasets hosted within IDOC and addresses the issues related to the digital preservation.

2 CONTEXT OF THIS DOCUMENT AND RECOMMENDATIONS FOR USE

This document is based on « IDOC-EX-001 IDOC executive summary », which describes IDOC (Integrated Data & Operation Center, https://idoc.ias.u-psud.fr/), which combines mission satellite operations and a spatial data center.

It is also described in this document « heads-up», in addition to the aspects of steering, strategy and implementation within IDOC, the global approach to take into account any new demand.

The objective of this document is to insure the long-term preservation of datasets implemented as part of « IDOC-OD-008 IDOC Guidelines for new services » and « IDOC-OD-004 IDOC Guidelines for Data integration » document.

For this it is necessary:

- to retain the dataset over the long term,
- to retain accessibility,
- to preserve the intelligibility and integrity

This means insuring :

- monitoring of: hardwares, softwares, user designated communities and their needs,
- management of the migrations which can be of several natures: hardware/software migration, digital migration of data with or without transformation (e.g. change of supports, disks, or change of format and structure of data)

Each new archive project enters into a process that leads to the identification of the specific points that need to be addressed in the implementation considerations described below and in the associated questions. These clearly established points make it possible to build an archive for these new data that meets the expectations of the scientific community and respects the FAIR criteria.

This archive then joins the pool of archives for which IDOC ensures the stable operation and plans the evolutions that will ensure the continuity of the availability.

3 APPLICABLE DOCUMENTS

	Référence	Titre
AD1	IDOC-LI-000	IDOC item list

4 REFERENCE DOCUMENTS

	Référence	Titre
	Référence	**Titre**
RD1	IDOC-EX-001	IDOC description executive summary
RD2	IDOC-OD-002	IDOC Risk analysis and management
RD3	IDOC-OD-003	IDOC General principles applicable to project design
RD4	IDOC-OD-004	IDOC Guidelines for Data integration
RD5	IDOC-OD-005	IDOC Guidelines for Pipeline Data Production
RD6	IDOC-OD-006	IDOC Guidelines for Archive long term preservation
RD7	IDOC-OD-007	IDOC Guidelines for Instrument operations
RD8	IDOC-OD-008	IDOC Guidelines for new services
RD9	IDOC-INF-009	IDOC Guidelines for Dataset dissemination
RD10	IDOC-INF-010	IDOC Organigramme
RD11		REGARDS – A generic CATALOG ACCESS SYSTEM AND data VALORIZATION tool

5 GENERAL STRATEGY FOR LONG TERM ARCHIVES PRESERVATION

Long term preservation of digital documents has three main objectives:

- **Preserve the information**
- **Make it accessible**
- **Preserve intelligibility.**

these three objectives aim to perpetuate not only the data as such but above all their capacity to be used effectively by the user communities.
Let's detail these objectives:

Preserve the information over the years?

That is the most obvious function expected of a repository. It must ensure that the record is always available on the storage medium, and that it maintains its integrity.

Make it accessible?

This means that you can find the document on the storage medium and retrieve its contents for use from any workstation that is normally available to users of that data.

Preserve the intelligibility of the document?

It is a question of ensuring that the document is certainly readable but above all that its content is intelligible to the user and that the semantics carried by this content is well preserved.

Note: Secure backup (or storage) only takes into account the first two of the three objectives listed above and only in the short and medium term.

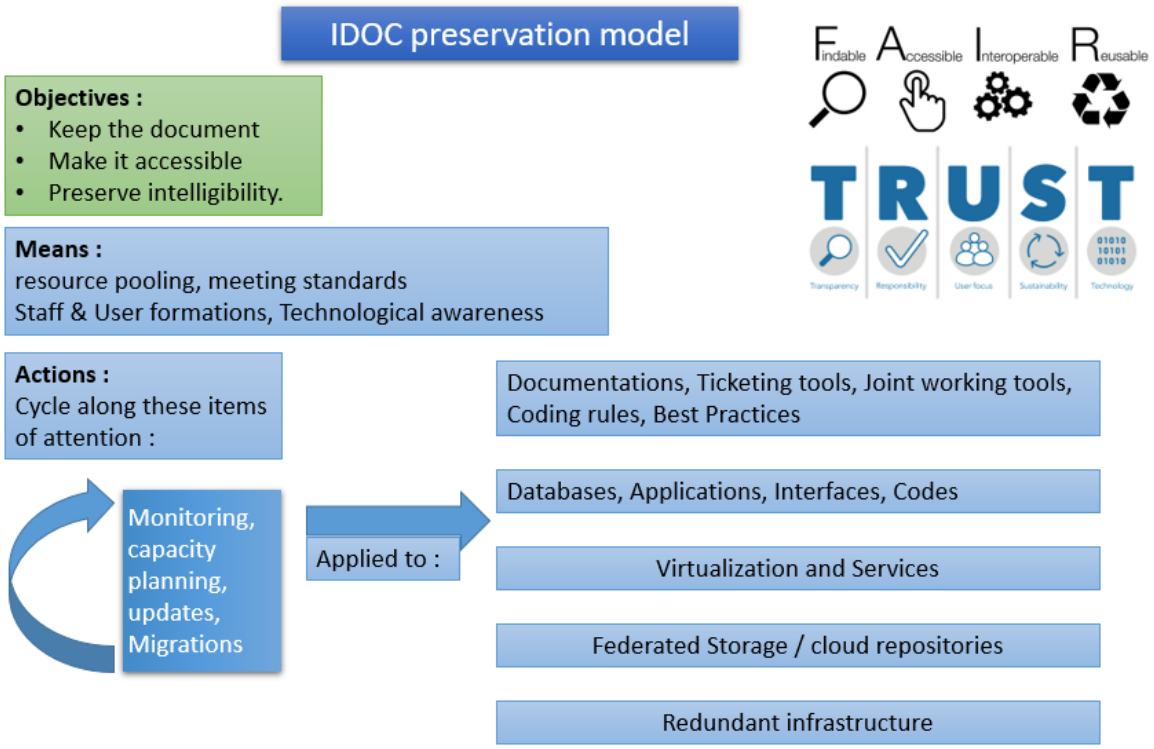
Ensuring that all three objectives are met means that it is necessary to validate over time that the tools, interfaces, descriptions, etc., which are the environment of the data and allow its use, retain their relevance for the understanding and effective use of the data.

To build and repeat this validation over time and at regular intervals, IDOC interacts with the scientific teams behind the data to describe this environment. This interactive procedure is described in the following chapters, and the result will lead to identify how to mitigate the four main risks that a dataset inevitably must face:

- Hardware obsolescence,
- Software obsolescence,
- File format obsolescence,
- Loss of the meaning of the content.

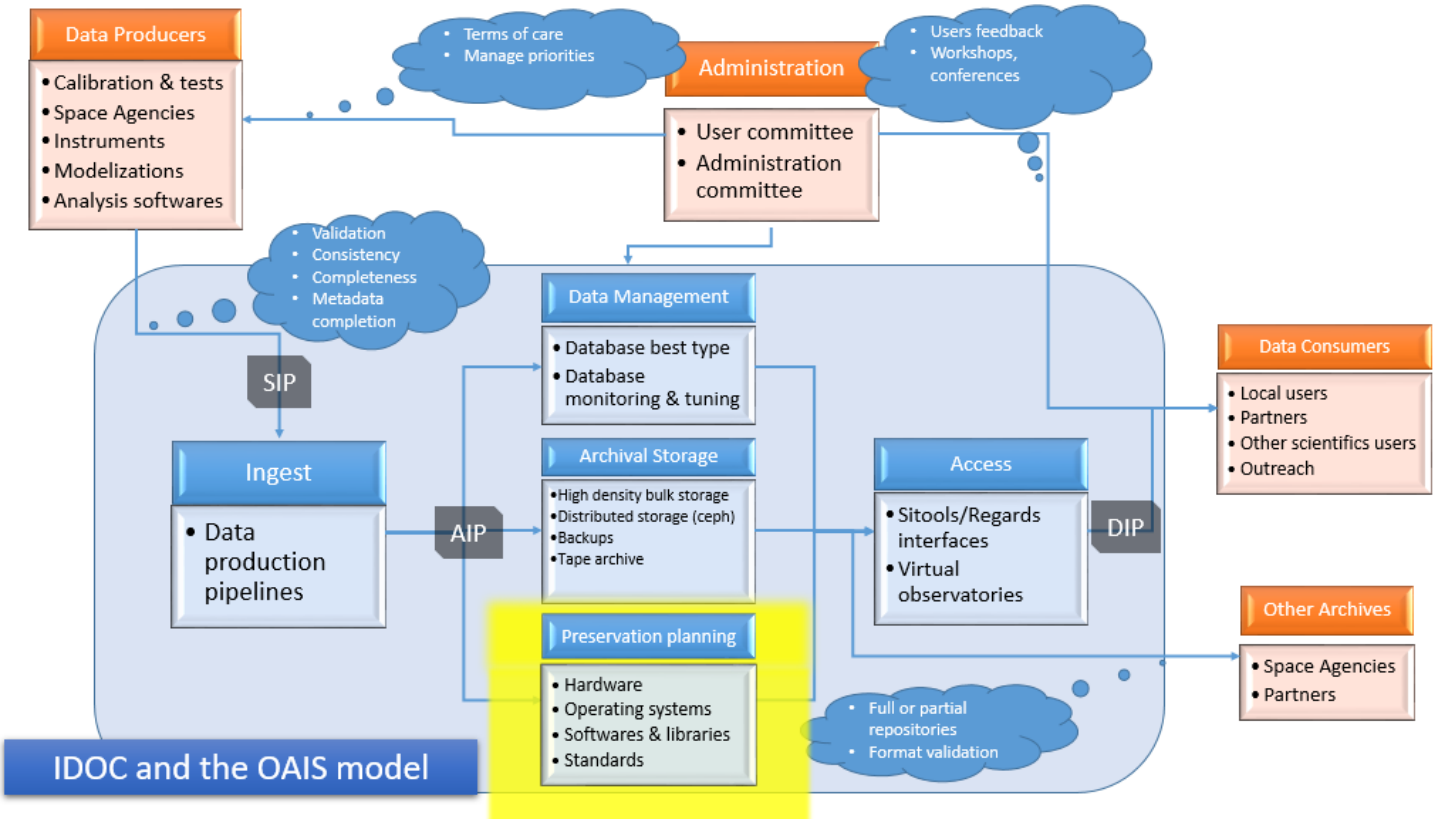
This will allow to determine the specific points of attention of this datatset that will join the usual points to which IDOC knows to take in attentive consideration.
Over time, these points of attention are validated in a cyclical way, and this scheme allow to keep the data intelligible.

This strategy is summarised in the following diagram:



6 APPROACH TO INTEGRATING A NEW ARCHIVE INTO THE LONG-TERM PRESERVATION OF ARCHIVES.

The main points to be validated and checked on the expected duration of preservation are described in the following paragraphs. The integration approach will therefore aim to answer each of the following chapters under the headings « associated question » of each of these chapters.
For a new dataset, it is important to describe which points of these descriptions can be made critical due to specific dataset specificities.
This document describes the yellow underlined part of the IDOC application of the OAIS model in the figure below.



7 PROCEDURE TO PREPARE

The creation of a Archive long term preservation service must be based on initial knowledge of the following elements that must be provided by those responsible for the dataset to be preserved who must therefore provide responses to this questionnaire :

7.1 USER COMMUNITY

7.1.1 Monitoring of user communities

What are the evolutions of user communities: number, centers of interest, tools,..?

7.1.2 Should the dataset be adapted to this evolution

7.2 ELEMENTS OF THE DATASET TO BE PERPETUATED

7.2.1 General considerations

The elements of the dataset to be perpetuated are to be specified, for example, in the case where successive versions have been constructed.

7.2.2 Related issues

Describe precisely what may not be sustained.
All the public concerned have been warned of the detail of this non-perpetuation.

7.2.3 Particularities of conservation of the datasets according to their level

See “IDOC-OD-004 IDOC Guidelines for Data integration” paragraph of the same name.

8 IDOC GENERAL CHECKPOINTS FOR ARCHIVE LONG TERM PRESERVATION

8.1 SUSTAINABILITY OF THE FORMAT OF THE DATA USED AND ASSOCIATED TOOLS

IDOC is careful to provide answers to the following questions at least in the user and steering committees:
What are the future evolutions due to the formats used?
What other emerging formats might be more relevant?
What are the upcoming evolutions of the tools related to these data formats (manipulation, conversion, integration with languages, analysers...)
If one of these issues raises the need for an evolution, the user committee or at least the thematic manager is consulted and a migration is then described and carried out.

8.2 SUSTAINABILITY OF THE HARDWARE TECHNOLOGIES USED

8.2.1 Reliability and durability of storage means

8.2.1.1 General considerations

IDOC, which operates active archives that require more immediate access than those available through tape-based technologies, virtually all of its storage is performed on disk. To adapt the needs to the costs, three types of configurations are used.

- High capacity storage at the lowest cost : the best price per terabyte is sought for disk bays
- Storage high capacity, high performance, high availability
- Distributed storage : scalable capacity for performance and availability (CEPH type solution)

8.2.1.2 Related issues

What are the future evolutions of the means of storage?

- CEPH distributed storage for all types of storage: access, redundancy, backup, long-term archive.
- Storage media flash memory

Which of these developments would be relevant in the short to medium term for an archive, several, all?
Future hardware technologies that emerge will be studied to identify which potential will make a change profitable.

8.3 MIGRATION FROM A STORAGE PLATFORM TO A NEWER PLATFORM

8.3.1 General considerations

Such a migration is organized nominally at IDOC according to the following scheme:

- Receiving new equipment
- Tests and formatting
- Data transfer, old active source, validation of tranfer ensuring integrity and authenticity (use of checksum)
- Access tests for new equipment

- Addition of new equipment to the monitoring and control system
- Transfer of the last modified data, old source inactive
- Toggle accesses and other data flows (backups, NFS mounts,..)
- Activating the new source

8.3.2 Related issues

What are the special precautions to be taken when migrating a dataset (e.g. flow rates, latency,..)?
What is the maximal unavailability time allowed during the toggle?
Should we privilege a particular period for this toggle?

8.4 Reliability and durability of hardware architectures (processors, ...) and operating systems

8.4.1 General considerations

IDOC implements virtualization to overcome a first level of hardware dependency. This virtualization also makes it possible through the high availability of the virtualization platform distributed over 3 sites to avoid a second level of dependency on hardware (failures, unavailability,..).
The computing platform is fully standardized (operating system, compilers, libraries,..)

8.4.2 Related issues

Which evolutions of the virtualization platform will benefit its use in IDOC ?
Would another virtualization platform be more appropriate?

8.5 Migrating a hardware architecture and operating system to a newer platform

8.5.1 General considerations

These migrations can be extremely cumbersome and involve the availability of language or format converters, advanced compilers, etc, and sometimes lead to complete rewrites of softwares.
Such a migration is organized nominally at IDOC according to the following scheme:

- Installing of a new test architecture
- Identification of unavailable software in the new infrastructure and choice of replacement
- Identification of new software versions requiring changes to previously configurations, settings or adaptations.
- Implementation of the entire software infrastructure updated on the new platform
- Tests of the new platform
- Adding the new platform to the monitoring and control system
- Toggle infrastructure on new platform
- Activation of the new infrastructure

The standardized calculation machines are updated very regularly thus damping the shock effect of spaced migrations.

8.5.2 Related issues

Is it possible to supply copies of the current hardware architecture and maintain them with the associated skills for the duration of a dataset?

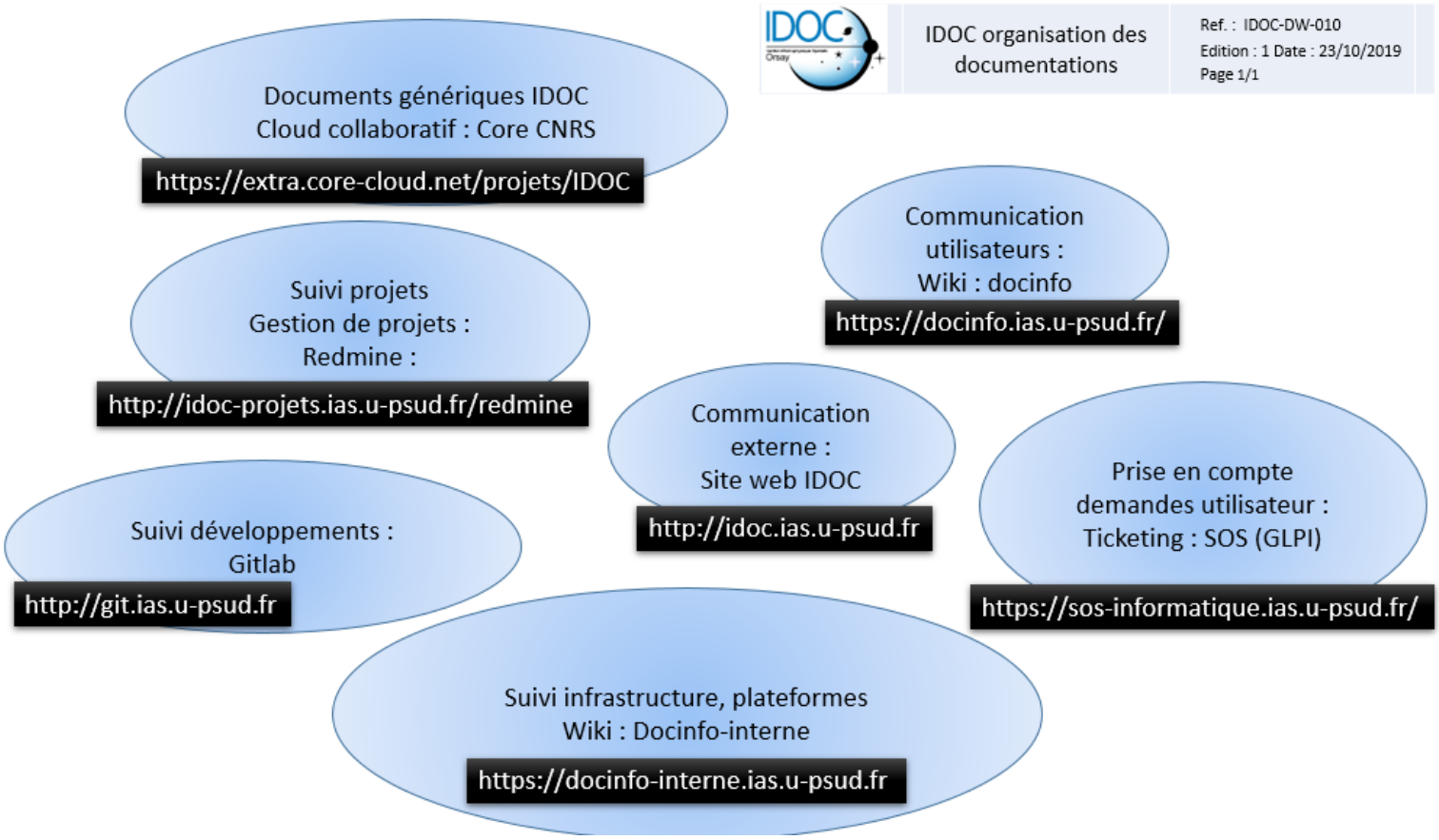
Can a virtual machine (or other type of software emulation) fully answer to the activity continuity requirements?
Is this also an opportunity to update the format, the form of access, the general availability of the relevant datasets?
What are the special precautions to take when migrating an infrastructure (access permissions linked to an address or a context, accounts, passwords,..) ?
What computing softwares can not support updating their environment ?

9 SUSTAINABILITY OF THE SOFTWARE TECHNOLOGIES USED

9.1 DOCUMENTATION TRACKING TOOLS

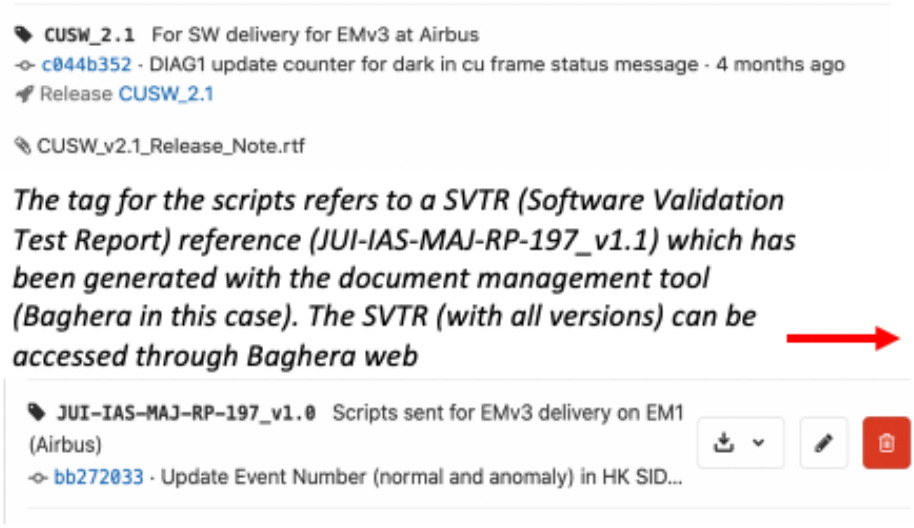
9.1.1 General considerations

IDOC deploys several types of document tracking, depending on the types of documentation and the intended audience.



All SW sources, associated scripts tests, reports and any related documentation are under configuration control. An example below with the MAJIS project (IAS instrument for the ESA JUICE mission).

Software source tagged with release note (gitlab)



9.1.2 Related issues

Are all types of documentation required for a new project covered by the panel proposed by IDOC ?
In the event that a project imposes its documentary system, what is the additional cost to IDOC of this support?

9.2 MIGRATION OF DOCUMENT TRACKING TOOLS

9.2.1 General considerations

The evolutions of the tools in this field are quite rapid and this renewal is generally accompanied by new functionalities whose character can become indispensable. Staff are also inclined to use the latest tools for their ergonomics and the constraints of the old tools then become a brake on their proper use.

9.2.2 Related issues

Features associated with considering migrating or evolving (collaborative work, project management,..) ?
Tools for migrating old documents of the same type to the new tool (Note : these tools rarely exist) ?

9.3 DATA INTERPRETATION TOOLS

9.3.1 General considerations

If the development recommendations described in IDOC documents such as « *IDOC-OD-003 IDOC General principles applicable to project design*» are followed , the maintenance of the sotware used will be greatly simplified. Nevertheless, beyond these recommendations, the general mode of operation of data processing can not guarantee the durability of the tools used (even a highly used open source library

can evolve without backward compability, requiring rewriting of calls, redesign of the operation mode of the software)

9.3.2 Related issues

Are the dependencies of the developed software identified (external and local developments, languages and their compilers, libraries, ..)?

9.4 MIGRATION OF DATA INTERPRETATION TOOLS

9.4.1 General considerations

The usual motivation for migration is more often the addition of features or the adoption of new interpretation techniques rather than intrinsic technical obsolescence.

9.4.2 Related issues

Is there another more recent/more used tool, possibly from another discipline and also/better suited to answer to the needs ?

9.5 DATA INTERFACES, WEB TECHNOLOGIES

9.5.1 General considerations

IDOC strategy is based on the use of the Sitools/REGARDS data access framework.
This framework is supported by CNES for its development and maintenance.

9.5.2 Related issues

Does the framework provide all the expected features for a new interface?
Is CNES commitment sustainable ?

9.6 MIGRATION DATA ACCESS INTERFACES

9.6.1 General considerations

CNES has planned the development of the successor of Sitools, on behalf of REGARDS. IDOC is largely involved in designing functionalities and monitoring the development of REGARDS. The migrations of the old access interfaces must be carried out in order to ensure technological coherence with the most recent interfaces. This strategy allows IDOC staff to focus their skills on the latest software components for better efficiency.

9.6.2 Related issues

Are the future problems of IDOC well taken into account in the future developments of REGARDS?
Is CNES strategy to promote REGARDS adequate and relevant?

10 STATISTICS

IDOC maintains on its data access interfaces a service for accounting for these accesses for statistical purposes. It makes it possible to evaluate among others the number, the volume and the geographical origin of the requests.
On a yearly basis the complete logs are deleted. Only the statistical elements allowing the monitoring of trends are kept.