IDOC

OSUPS

Data Center

IDOC-DATA Instructions for Data Preservation

IDOC-OD-006

Préparation

	Nom et Fonction	Date
Rédacteurs	Gilles Poulleau	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
3.0	23/08/2022	Reorganization

SOMMAIRE

1 SCOPE OF THE DOCUMENT

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

This document describes how IDOC-DATA operates to maintain over time the availability and use of data at least as far as its initial implementation is concerned. It therefore deals not only with short or medium term but also with the long-term preservation of the datasets hosted within IDOC-DATA and addresses the issues related to the digital preservation.

Preliminary note: predicting the availability of data in the long term involves a certain degree of uncertainty because the underlying technologies involved are constantly changing and this change is, from experience, unpredictable beyond the short term.

For example, the choice of a future data storage method in an era that has at has moved from the file cabinet to the five-inch floppy disk and the cloud in the space of 35 years.

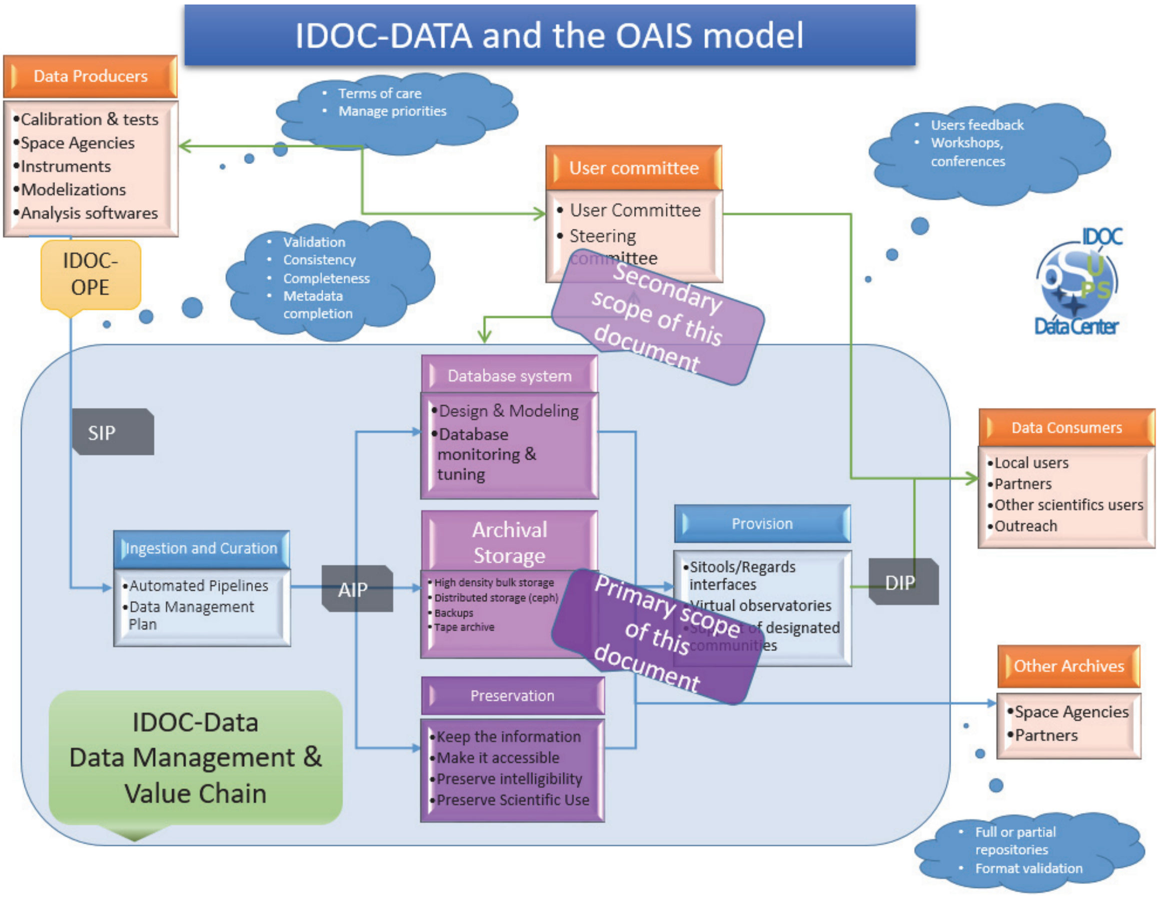
2 CONTEXT OF THIS DOCUMENT AND RECOMMENDATIONS FOR USE

This document is based on « IDOC-EX-001 IDOC executive summary », which describes IDOC (Integrated Data & Operation Center, https://idoc.osups.universite-paris-saclay.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 consider any new demand.

The objective of this document is to ensure the short, medium and long-term preservation of datasets implemented as part of « IDOC-INS-008 IDOC Instructions for new services » and « IDOC-INS-004 IDOC-DATA Instructions for Data integration » document.

This document describes how to build and maintain the highlighted part of the IDOC application of the OAIS model in the figure below:



IDOC-DATA and the OAIS model

Data Producers
- Calibration & tests
- Space Agencies
- Instruments
- Modelizations
- Analysis softwares

- Terms of care
- Manage priorities

User committee
- User Committee
- Steering committee

- Users feedback
- Workshops, conferences

IDOC-OPE

- Validation
- Consistency
- Completeness
- Metadata completion

Secondary scope of this document

SIP

Database system
- Design & Modeling
- Database monitoring & tuning

Data Consumers
- Local users
- Partners
- Other scientifics users
- Outreach

Ingestion and Curation
- Automated Pipelines
- Data Management Plan

AIP

Archival Storage
- High density bulk storage
- Distributed storage (ceph)
- Backups
- Tape archive

Provision
- Sitools/Regards interfaces
- Virtual observatories
- Support of designated communities

DIP

Primary scope of this document

Other Archives
- Space Agencies
- Partners

IDOC-Data Data Management & Value Chain

Preservation
- Keep the information
- Make it accessible
- Preserve intelligibility
- Preserve Scientific Use

- Full or partial repositories
- Format validation

3 REFERENCE DOCUMENTS

Acronym	Reference of the document	Document full name
RD1	IDOC-EX-001	IDOC-EX-001 IDOC executive summary
RD2	IDOC-OD-002	IDOC-OD-002 IDOC Risk analysis and management
RD3	IDOC-INS-003	IDOC-INS-003 IDOC Instructions applicable to project design
RD4	IDOC-INS-004	IDOC-INS-004 IDOC-DATA Instructions for Data Ingestion and Curation
RD5	IDOC-INS-005	IDOC-INS-005 IDOC-OPE Instructions for Ground Segments
RD6	IDOC-INS-006	IDOC-INS-006 IDOC-DATA Instructions for Data Preservation
RD7	IDOC-INS-007	IDOC-INS-007 IDOC-OPE Instructions for Instrument Operations
RD8	IDOC-INS-008	IDOC-INS-008 IDOC Instructions for Services

RD9	IDOC-INS-009	IDOC-INS-009 IDOC-DATA Instructions for Data Provision
RD10	IDOC-INF-010	IDOC-INF-010 IDOC Organigrammes
RD11	IDOC-DW-011	IDOC-DW-011 Diverses schemas for documentation
RD12	IDOC-INS-012	IDOC-INS-012 IDOC instructions for architecture and coding practices
RD16	IDOC-EX-016	IDOC-EX-016 OSUPS Schéma Stratégique Numérique
RD17	IDOC-OD-017	IDOC-OD-017 Services offerts par IDOC
RD30	IDOC-HO-030	IDOC-HO-030 Presentation IDOC-public-english
RD31	IDOC-HO-031	IDOC-HO-031 Presentation IDOC Français

4 GENERAL STRATEGY FOR DATA PRESERVATION

4.1 PRESERVATION OBJECTIVES

Over time preservation of digital documents has four main objectives:

- **Preserve the information**
- **Preserve intelligibility.**
- **Make it accessible**
- **Make it usable in a scientific way**

These four 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:

4.1.1 Preserve the information over the years? or Impact of Ingestion and Curation on the "Database system " and the "Archival Storage" (secondary scope)

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.

IDOC-DATA answer:

As stated in [RD2], a constant flow of actions is triggered to ensure the stability over time of the infrastructures, their evolution and the migration of data without loss of information by ensuring their integrity and continuous availability. The challenge that IDOC-DATA is responding to is the following:
A new data storage paradigm may emerge in due course (and probably will). emerge in the next few years or decades (and probably will) but it is unlikely to replace the digital paradigm.
In short, data infrastructures will retain the ability to store data in a so-called "digital" form of coding for a long time to come. Conversely, it is certain that the media or structures on which all digital information is stored will be obsolete within a decade or two.

- But the tools to transfer from a storage technology n to a technology n+1 have always existed, but IDOC-DATA has to make sure that it never leaves data in a version n-1.

Although the process is often less straightforward, the same considerations apply in relation to databases and their process of migration and evolution as their underlying technologies change.

4.1.2 Preserve the intelligibility of the information?

The aim is to ensure that the data and assiocated documents are certainly readable but above all that their contents are intelligible to the user and that the semantics carried by this content is well preserved.

IDOC-DATA answer:

This information at IDOC-DATA is always coded in standardised formats.
The knowledge of the formats of these preserved data is itself information, and is also represented digitally. It is therefore be intrinsically preserved by the same means as the data itself.same means as those used for the data itself.

- As for the storage itself, the evolution of formats is accompanied by tools allowing the translation of a format becoming obsolete to a fashionable one, and as for the storage, IDOC-DATA has to make sure to carry out these translations at the right time.

Among all the processes that will lead to the preservation of the data, one of them will possibly lead to the renewal of the ingestion and curation process during the future life of the dataset.

4.1.3 Make it accessible?

Data provision preservation must for this specific matter comply with:

- Stability: ability to have the same result along time,
- Referencing: the location of the information stay predictable
- Certified origin: informations are produced by successive certified processes,
- Context: each bit of information has a context allowing its understanding.

IDOC-DATA answer:

At IDOC-DATA, these points are addressed by considering that a new interface for accessing the data must be designed and implemented. Experience shows that the process is no longer one of evolution, but of designing new interfaces for the data when obsolescence is detected, given the technological leaps involved. In this sense, document [RD9] is therefore applicable again.

Note : The greatest attention is paid to the contribution of scientists who are aware of the latest or future developments in the practices of the community concerned.

4.1.4 Make it usable in a scientific way

Finally, the data stream and it's semantic information available via any means of remote selection and retrieval may require other elements to be scientifically interpreted or manipulated to extract higher level information. It can be:

- specific software beyond the usual data format tools,
- or the data alone from this primary stream may not be sufficient for a scientifically correct use of the data. (Examples include the addition of instrumental context data, acquisition environment data, instrument calibration data).

For all these data and software, it is also necessary to maintain the continuity in time of their availability.

IDOC-DATA answer:

Note: As far as software is concerned, the conversion of a program into a modern language or platform version can be very complex and expensive and require skills that are sometimes scarce. However, virtualisation environments allow for a significant extension of operations that meet the requirements.

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.

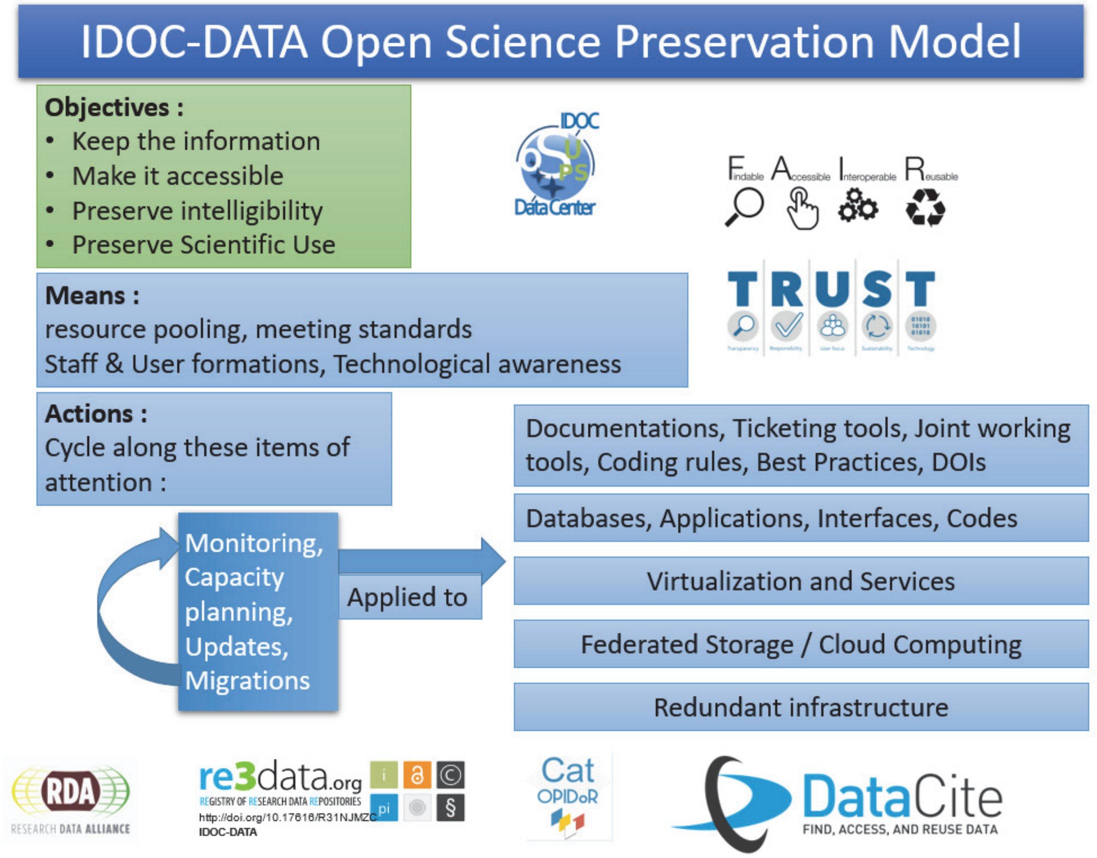
4.2 GENERAL STATEMENT : CYCLING PROCESSES THE PRESERVATION PHASE

Ensuring that all four 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.

Applying this strategy identifies how to mitigate the four main risks either globally, by category of data and then at the level of a particular dataset.

If for the first two perimeters IDOC-DATA can rely on its internal know-how, for the aspect that concerns a particular dataset, IDOC-DATA interacts at regular intervals with the scientific community behind the data to describe this environment (see annex) with the aim of adapting the evolution of the means of this preservation to not only technical but also scientific realities. IDOC-DATA's final objective is to ensure permanent access to the intellectual content of the data, and not to a simple flow of bytes.

All the aspects described thus follow a continuous and supervised cycle as described in the following schema.



IDOC-DATA Open Science Preservation Model

Objectives :

- Keep the information
- Make it accessible
- Preserve intelligibility
- Preserve Scientific Use

Means :
resource pooling, meeting standards
Staff & User formations, Technological awareness

Actions :
Cycle along these items of attention :

Monitoring, Capacity planning, Updates, Migrations

Applied to

Documentations, Ticketing tools, Joint working tools, Coding rules, Best Practices, DOIs

Databases, Applications, Interfaces, Codes

Virtualization and Services

Federated Storage / Cloud Computing

Redundant infrastructure

5 LONG TERM PRESERVATION: CONTINGENCY MEASURES IN CASE OF IDOC-DATA IS UNABLE TO FULFIL ITS OBLIGATIONS

Firstly, it is important to specify that, as things stand, the CNRS and the Université Paris Saclay strongly support IDOC-DATA and are committed to its approach.

In addition, CNES, the French national space agency, has recognized IDOC-DATA for a certain number of data archiving missions which have been the subject of agreements linked to the space missions from which the data originated or specifically for archiving purposes at regional or national level.

All this context tends to reassure on the sustainability of IDOC-DATA either in its current structure or if this structure should change, the willingness of the guardians and contributors of IDOC-DATA to perpetuate the actions towards the hosted data. This transfer could be conducted to a new entity or by distributing these services in existing structures and adapted (the french CINES for example, already certified by CTS).

If nevertheless, it happened that the services could not be reassigned by the trustees, the international nature of data hosted and made available by IDOC-DATA would allow to keep their access for most of them.

Thus, for the solar theme, most of the data are already available through virtual observatories and hosted in another global location (SDO data for example)

For the ESA missions, all the produced datasets stabilized by the instrument teams must be transferred to the ESA data center, which guarantees permanent access to these data. https://www.cosmos.esa.int/web/esdc.

For instance, regarding the planetary theme, the most important data from IDOC-DATA are regularly added to the Planetary Science Archive (PSA) of the European Space Agency (ESA).

What would be lost on the data described above would be the redundancy of the accesses associated with a degradation of the response times of these services. Some original ergonomic aspects of the interfaces today offered by IDOC-DATA could be compromised but would probably be compensated for as the interfaces of the other centers are renewed and their technology evolves.

Moreover, some of the services implemented at IDOC-DATA are the result of the designation of the center as winner of a call for tenders aiming at the implementation of services (e.g. Space Situational Awareness of the European Space Agency). As these services are subject to a recurrent follow-up, there is no doubt that the donor would repeat its call for tender and would then designate a successor to IDOC-DATA.

As for the rest of the datasets not concerned by the contexts described above, their volumetry is less important than those already preserved today. It is likely that a call to the international data centers concerned by these themes would allow a very significant part of these data to be taken up and made available to the communities. IDOC-DATA's current involvement in initiatives such as the virtual observatories makes it possible to create links that would then prove useful.

6 ANNEX

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

The main points to be 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.

6.2 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 :

6.2.1 User community

6.2.2 Monitoring of user communities

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

6.2.3 Should the dataset be adapted to this evolution

6.3 ELEMENTS OF THE DATASET TO BE PERPETUATED

6.3.1 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.

6.3.2 Related issues

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

6.3.3 Particularities of conservation of the datasets according to their level

See "IDOC-INS-004 IDOC-DATA Instructions for Data integration" paragraph of the same name.

6.4 IDOC GENERAL CHECKPOINTS FOR ARCHIVE LONG TERM PRESERVATION

6.4.1 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.

6.4.2 Sustainability of the hardware technologies used

6.4.2.1 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)

6.4.2.2 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.

6.4.3 Migration from a storage platform to a newer platform

6.4.3.1 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

6.4.3.2 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?

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

6.4.4.1 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,..)

6.4.4.2 Related issues

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

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

6.4.5.1 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.

6.4.5.2 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,..) ?

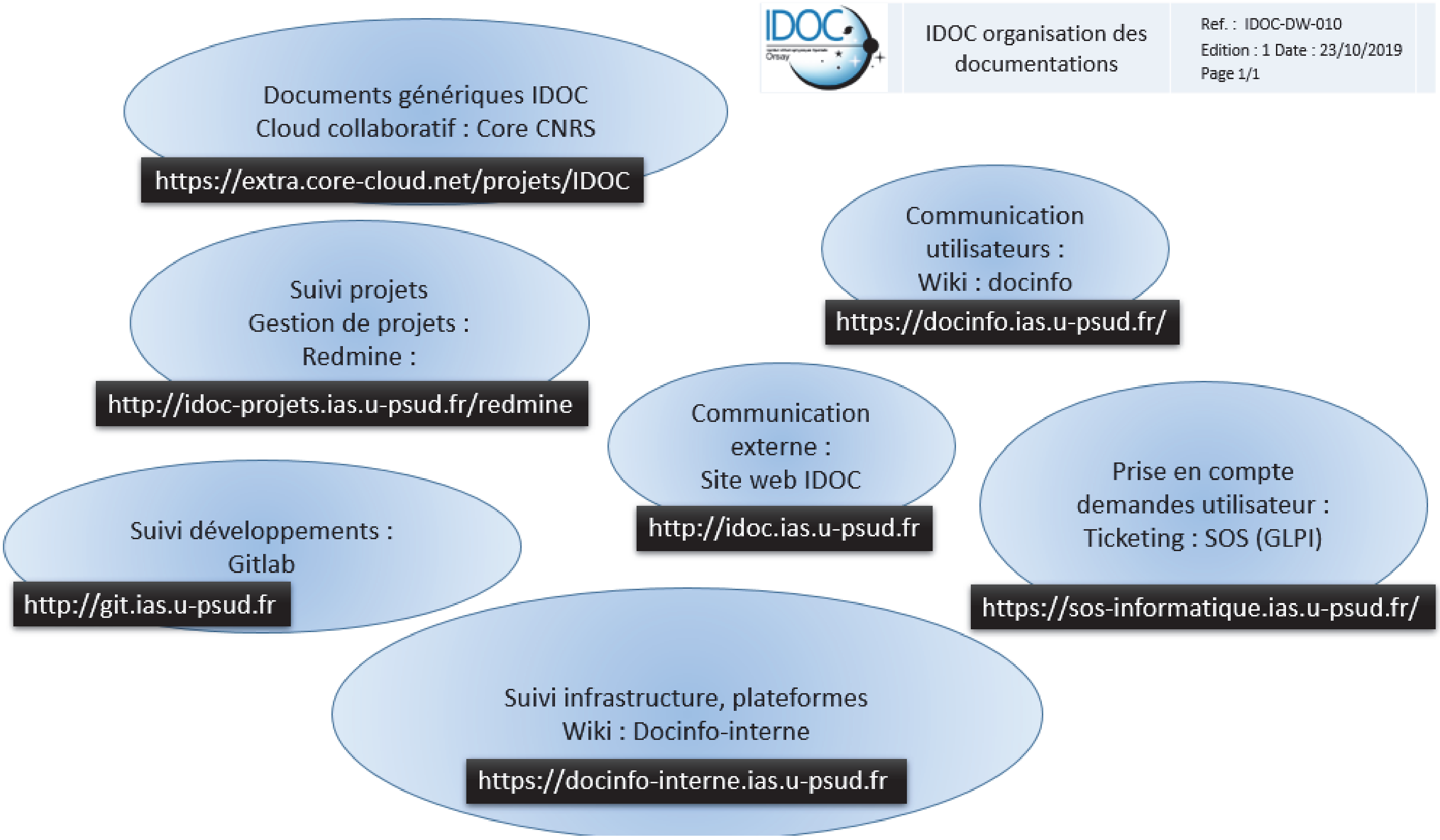
What computing softwares can not support updating their environment ?

6.5 SUSTAINABILITY OF THE SOFTWARE TECHNOLOGIES USED

6.5.1 Documentation tracking tools

6.5.1.1 General considerations

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



IDOC organisation des documentations

Ref. : IDOC-DW-010
Edition : 1 Date : 23/10/2019
Page 1/1

Documents génériques IDOC
Cloud collaboratif : Core CNRS
https://extra.core-cloud.net/projets/IDOC

Communication utilisateurs :
Wiki : docinfo
https://docinfo.ias.u-psud.fr/

Suivi projets
Gestion de projets :
Redmine :
http://idoc-projets.ias.u-psud.fr/redmine

Communication externe :
Site web IDOC
http://idoc.ias.u-psud.fr

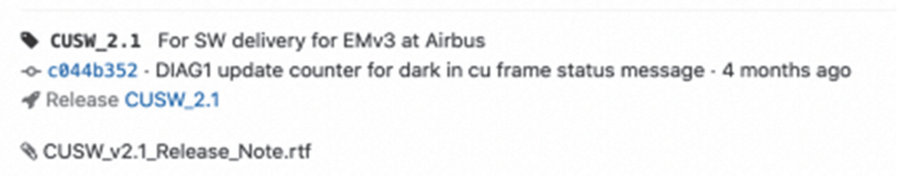
Prise en compte demandes utilisateur :
Ticketing : SOS (GLPI)
https://sos-informatique.ias.u-psud.fr/

Suivi développements :
Gitlab
http://git.ias.u-psud.fr

Suivi infrastructure, plateformes
Wiki : Docinfo-interne
https://docinfo-interne.ias.u-psud.fr

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).

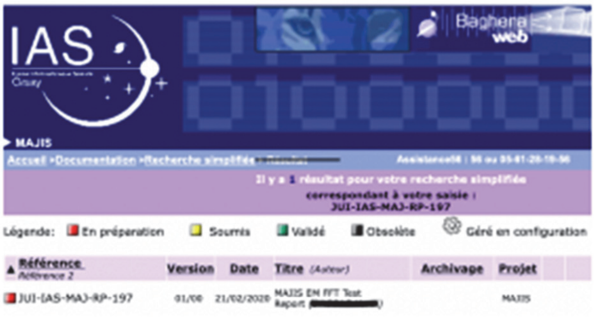
Software source tagged with release note (gitlab)



CUSW_2.1 For SW delivery for EMv3 at Airbus
c044b352 · DIAG1 update counter for dark in cu frame status message · 4 months ago
Release CUSW_2.1

CUSW_v2.1_Release_Note.rtf

The tag for the scripts refers to a SVTR (Software Validation Test Report) reference (JUI-IAS-MAJ-RP-197_v1.1) which has been generated with the document management tool (Baghera in this case). The SVTR (with all versions) can be accessed through Baghera web



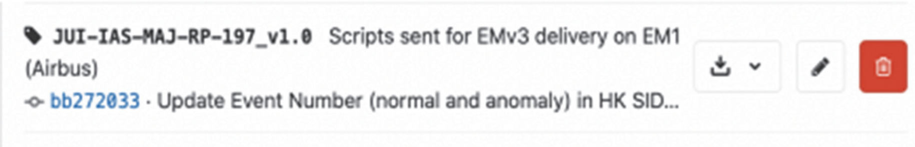
MAJIS

Accueil >Documentation >Recherche simplifiée [redacted] Assistance : 56 ou 05-61-28-19-56

Il y a 1 résultat pour votre recherche simplifiée correspondant à votre saisie : JUI-IAS-MAJ-RP-197

Légende: 🟥 En préparation 🟨 Soumis 🟩 Validé ⬛ Obsolète ⚙ Géré en configuration

Référence Référence 2	Version	Date	Titre (Auteur)	Archivage	Projet
🟥 JUI-IAS-MAJ-RP-197	01/00	21/02/2020	MAJIS EM FFT Test Report ([redacted])		MAJIS



JUI-IAS-MAJ-RP-197_v1.0 Scripts sent for EMv3 delivery on EM1 (Airbus)
bb272033 · Update Event Number (normal and anomaly) in HK SID...

6.5.1.2 *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?

6.5.2 Migration of document tracking tools

6.5.2.1 *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.

6.5.2.2 *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) ?

6.5.3 Data interpretation tools

6.5.3.1 *General considerations*

If the development recommendations described in IDOC documents such as « *IDOC-INS-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)

6.5.3.2 Related issues

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

6.5.4 Migration of data interpretation tools

6.5.4.1 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.

6.5.4.2 Related issues

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

6.5.5 Migration of data interfaces, web technologies

6.5.5.1 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.

6.5.5.2 Related issues

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

6.5.6 Migration of data access interfaces

6.5.6.1 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.

6.5.6.2 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?

6.6 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.