IDOC-DATA Instructions for Data ingestion and curation

IDOC-OD-004

Préparation

	Nom et Fonction	Date
Rédacteurs	Gilles Poulleau	Aout 2022
Vérificateur	Marian Douspis	
Approbateur	Prénom Nom, ***fonction***	05/04/2016

Liste de diffusion

Nom	Fonction	Société

Evolutions

Edition	Date	Modifications
1.0	23/01/2016	1st draft in french
2.0	01/02/2017	translated in english
2.1	19/04/2017	Entirely rewritten
2.2	19/03/2021	Update
3.0	03/08/2022	IDOC-DATA

SOMMAIRE

1 Scope of the document 5
2 Reference documents 5
3 General Instructions applicable to data INGESTION AND curation 6
3.1 Each Dataset is officialised within IDOC 6
3.2 FAIR principles 6
3.3 Data Integrity and authenticity 7
3.4 Ingestion 7
3.5 Curation 7
3.6 Impact of Ingestion and Curation on the “Database system “ and the “Archival Storage” (secondary scope) 8
3.7 Cycling processes of ingestion and curation 9
3.8 IDOC-DATA practices given the data processing level 9
4 Production of Data managements plans 11
5 Annexes 12
5.1 Annexe : Procedure to prepare : Context of the dataset (DMP-1) 12
5.1.1 How to produce a DMP 12
5.1.2 Context within IDOC 12
5.1.3 Description of the dataset 12
5.1.4 Level of requested curation 13
5.1.5 Data access 13
5.1.6 Ethics and confidentiality 13
5.1.7 Quality of the dataset 13
5.1.8 Dataset discovery 13
5.1.9 Dataset reuse 14
5.2 Annexe : Procedure to prepare : Dataset enumeration form (DMP-2) 14
5.2.1 External dataset content 14
5.2.2 Property of the dataset 14
5.2.3 Size and time requirements 14
5.2.4 Criticality, availability and expected security of the dataset 14
5.2.5 Lifecycle of the dataset 14
5.3 Annexe :Order of magnitude of costs according to Dataset environment : Implementation matrices 15
5.3.1 « Availability / Performance » matrix 15
5.3.2 « Backup / Historization » matrix 15
5.3.3 Example of approximate cost calculation using implementation matrices 16
5.4 Annexe : Supplementary notes 16

5.4.1 Procedure of new data integration in IDOC 16

1 SCOPE OF THE DOCUMENT

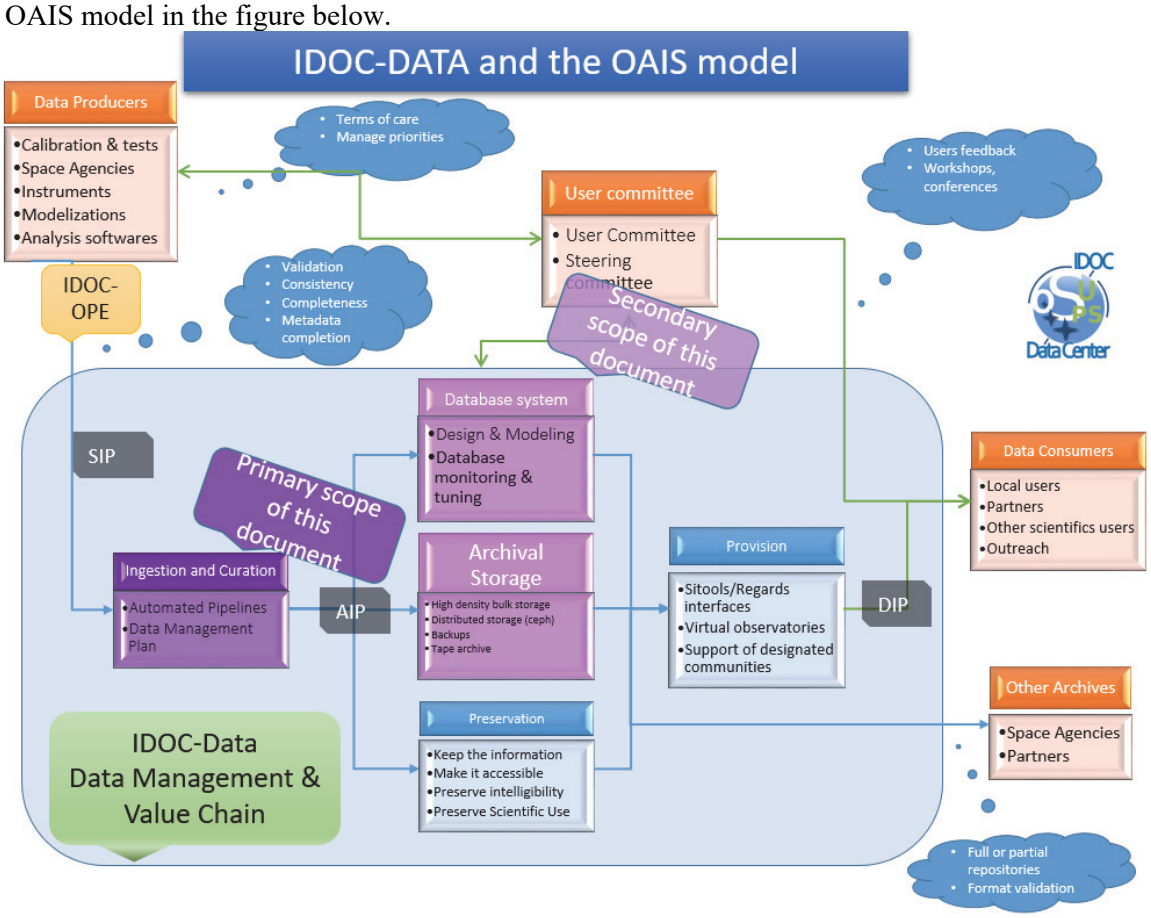
This document is related to the « IDOC-INS-008 IDOC Instructions for new services » which needs to be handled first. This present document is applicable in the specific case of a decision-maker requesting the integration of an external dataset in the IDC infrastructure. The document explains and details the needed inputs for building and deliver this dataset service.

2 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

3 GENERAL INSTRUCTIONS APPLICABLE TO DATA INGESTION AND CURATION

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



3.1 EACH DATASET IS OFFICIALISED WITHIN IDOC

In order to integrate IDOC-DATA, a new dataset requires approval from the steering committee which will make its decision according to the usefulness for the communities gathered around IDOC-DATA. After approval, it is examined both scientifically and technically.

The technical aspects of the ingestion have to be checked and as this external dataset might not be complete, organized, documented, or can presents other insufficiencies for long-term preservation (metadata, format...), processes of curation are applied to remove or mitigate these defects. This is done by following the following instructions.

3.2 FAIR PRINCIPLES

IDOC-DATA has been actively involved since the beginning in processes that aim to improve the infrastructure supporting the reuse of scholarly data. Therefore, when a diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that now refer to as the FAIR Data Principles, IDOC-DATA makes every effort to comply with these recommendations.

For example, from the first implementations of the data access interfaces, the ability of machines to automatically find and use the data, has been implemented.

3.3 DATA INTEGRITY AND AUTHENTICITY

Whatever the service IDOC-DATA implements, and independently of the possible enhancement on the datasets, the initial dataset is kept unchanged. This means that all level of dataset management assume that potential added value and curation are only made on copies of the modified originals parts allowing the data to be restored to its original state if the need arises.

Moreover, IDOC-DATA's dataset management ensures integrity and authenticity during the processes of ingest, archival storage, and data access: changes to data and metadata are documented and the relationship of the dataset with the original data is maintained.

3.4 INGESTION

The channels through which data is ingested must be secure, and the available contacts at the source of the data identified.
If they are to be used for an extended period of time (e.g. in the case of an ongoing mission) ingestion processes are automated as much as possible and the instructions to be followed are those described in the document IDOC-INS-005 IDOC-OPE Instructions for Ground Segments for designing and operating data pipelines.

If the ingestion process is a one-off or if the expected frequency is very low, then the overall process can simply be documented very precisely.
In all cases, programs for reading, transforming and sending data to the structure defined during the design of the "database system" (see below) must be fully developed, tested and validated.
Most of the implementation takes into account the information obtained during the drafting of the data management plan for the data set.

It may happen that between the target data format and the different data sources, an intermediate "pivot" format fully intelligible to both parties (producer and IDOC-DATA) is useful to implement to simplify the ingestion operations towards appropriate database or ather storage .

3.5 CURATION

Data curation activities at IDOC-DATA always include:

- **Contextualizing:** In contextualizing, metadata (e.g. relevant sources and attributions) is added to the dataset. The purpose is to show the context regarding how and why the data was generated. In this process the associated metadata are always dependent on the context, the current state of the art, of the associated scientific communities. The data set up at IDOC-DATA must conform to these community uses or respect the rules imposed by the space agencies for the storage and dissemination of ingested data. Compliance with this intangible rule means that the dataset can be included in the relevant virtual observatories without any problems.
- **Citing the data:** Contributors who have significantly contributed to the genesis of the data are easily accessible in the global metadata of the dataset. Third party users are encouraged to use these appropriate attributions when citing the data.
- **Validating and adding metadata:** Information about a dataset is structured in a machine-readable form for search and retrieval purposes.

- **Validation of data:** Local experts within OSUPS (or partners) with appropriate qualifications and knowledge of the subject as relevant creators or users of this type of data review the dataset. It is performed to confirm the accuracy of the data.

All these points are essential to build data access interfaces that respect the FAIR approach of IDOC-DATA.
They will allow these interfaces to have:

- stability (the data flow through the interface is reproducible over time)
- ergonomics and user-friendliness,
- and the referencing of the data used

As the data to be ingested are in the initial state often only accompanied by a fraction of the above elements, a process must be set up to fill in the gaps.
During this process the appropriate curation is determined and applied in order to meet the ergonomic standards of IDOC-DATA portals.
A document is built with the data provider before hosting and distributing a new dataset. It serves as a reference for the implementation of a new data service at IDOC-DATA. It explains and details the needed inputs for building and deliver this dataset service
Building the appropriate curation process involves iterative interaction between the client and the IDOC-DATA Technical leader until all questions are answered unambiguously. This document can subsequently be used as a Data Management Plan (DMP) as it contains all the required elements. It will then be implemented, and the service finally given to the customer.
An annex below describes more in detail this process.

3.6 Impact of Ingestion and Curation on the "Database system " and the "Archival Storage" (secondary scope)

The "database system" is the sum of the repository structure and its appropriate receptacles (file structures, hierarchical data structures, databases, storage types, technology and mediums). It's design must be carefully conducted taking into account the type of data to be ingested, the level of curation applied, and the future uses - and the load they will represent - of these data.

Depending on the dataset and how it will be used, the form of the structures created to store and describe the data (organisation and types of files, databases, etc.) are designed. Typically, the target system will depend on a number of factors including:

- the expected number of records (database size);
- the expected number of users (database clients);
- the expected duration of the registry (length of data storage);
- The type of data being stored (data type);
- and the duration of the data storage after the registry project is complete.

IDOC-DATA has the expertise to implement the most appropriate type of databases for the context of the dataset being implemented, both for the data itself and for the metadata (i.e., SQL or NoSQL).

Regarding the storage infrastructure (including redundancies and various backups), or more generally all the infrastructures concerned by the dataset support, IDOC-DATA's pooling strategy of all such resources will unambiguously apply to any new dataset. Given the level of stability and availability expected for some of the data operated by IDOC-DATA and the versatility of the solutions chosen, the platform will always be in a position to build the required level of reliability;

3.7 CYCLING PROCESSES OF INGESTION AND CURATION

Among the processes that will lead to the preservation of the data, one of them will probably lead to the renewal of this ingestion and curation process during the future life of the dataset. The IDOC-DATA Instructions for Data Preservation document [RD6] lists the topics checked and validated for implementing a data preservation service at IDOC and this cycling process. In the [RD6] "Annex" (see the document) is described the expected implementation of the preservation of one particular dataset. The dataset parameters are regularly revised (minimum every three years) to ensure the cyclic aspect of the monitoring. Along time this « curative » strategy also allows to enrich the content.

3.8 IDOC-DATA PRACTICES GIVEN THE DATA PROCESSING LEVEL

Data products at IDOC-DATA are processed at various levels ranging from Level 0 to Level 4. Level 0 products are raw data from the instrument. The higher the levels, the more processed are the data towards a possible scientific goal. The next table summarizes the usual processing levels.

Data Level	Description
Level 0	Complete data of the instrument reconstructed from telemetry flow from the satellite. In most cases, the control center will remove communication artifacts from this flow (e.g. synchronization frames, communication header, duplications,..)
Level 1A	Reconstructed, unziped, sorted, time referenced data of the instrument. Adjonction of auxiliary data necessary to the technical or scientific interpretation of the data (score, relative positioning,..) as well as calibration data. The different types of data (technical, instrument channels,..) from the instrument will be processed differently and will be separated.
Level 1B	Level 1A data translated in physical quantities.
Level 2	First level of data interpretation of the instrument. The data management is not modified at this stage. (locations, fields of view, zones are the same as level 1) A knowledge of the scientific field of the instrument can recreate its own and independent Level 2. Many Level 2 can coexist.
Level 3	Second level of data interpretation of the instrument. The produced data management is adapted to their interpretation or processing.
Level 4	Third level of data interpretation. The required knowledges to produce this type of data level can be a totally different field from the initial field of the instrument.

Level 0 data are usually kept as is. IDOC-DATA has no commitment other than the provision of such data and related documents. Indeed, the understanding of these data is usually the responsability of the instrument team. The same applies to levels 1A and 1B data.
A priori, no reprocessing will be performed by IDOC-DATA on these data without the request and the support of the instrument team. Indeed, it is the only one able to provide all the elements allowing this transformation because it imposes a very fine knowledge of the instrument to be free from errors.

The other levels of data, if modified (improving the pipeline from a more advanced understanding of the interpretive elements, or through the availability of better tools in order to raise the FAIRness of the data) will see the original retained or the original programs made available for reconstruction.

At each level, the appropriate medata are generated in order to fullfill the designated community's standards.

4 PRODUCTION OF DATA MANAGEMENTS PLANS

A Data Management Plan (DMP) is a formalized document detailing the way to obtain, to insert, to document, to analyze, to circulate and to use data produced during a process or a research project in IDOC-DATA.
The DMP uses the data/document lifecycle and describes the choices realized in term of metadata norms, data base formats, methods and access security, archiving period, and costs for data management.
Special attention should be given to data coming from publications, which should stay available and open to many people as possible.
The creation of the DMP is more and more requested for calls of proposals funded by public funds, in particular european funds.
Excerpt from the european commission guide on publications and data open access in Horizon 2020 :
“ A data management plan is a document outlining how the research data collected or generated will be handled during a research project, and after it is completed, describing what data will be collected/generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved .”
The objective is thus to document how collected or generated data will be handled during a research project, and after this document is completed, how these data will be described, organized, shared, protected and preserved.
Another aspect of the digital data preservation can be described as the following 5 terms, ensuring the data integrity:

- Content: a set of formatted and sorted digital informations.
- Stability: all the informations mentioned above are supposed to allow to find the same information over the years.
- Referencing: the location of an information is predictable.
- Certified origin: informations come from a process targeted by a chain of commitments.
- Context: each information is linked to a context which allows its interpretation.

This former approach can be formalized in the DMP and enables the participant a better consideration and understanding of the challenges.

At IDOC-DATA, a DMP is therefore unique for each hosted project. It is a formal document describing:

- how to retrieve (or receive), ingest, document, analyse, circulate, and make use of the data,
- how collected data or data produced are handled, and, when completed, how these data are be described, organized, shared, protected and preserved.
- The DMP shall also describe the metadata norms, data base formats, methods and access security, archiving period, and costs for data management.

5 ANNEXES

5.1 ANNEXE : PROCEDURE TO PREPARE : CONTEXT OF THE DATASET (DMP-1)

5.1.1 How to produce a DMP

The process is for the decision-maker to interact iteratively with the IDOC Technical leader until all the questions enumerated in the following chapters are answered unambiguously. These answers will be then formalized in a Data Management Plan. This DMP will then be implemented: the resources will be allocated for implementing and performing the service; the rights and commitments of IDOC-DATA will be set.

5.1.2 Context within IDOC

Dataset name

Define the dataset name and its acronym. Pay attention to the name which will be internally to IDOC-DATA given to the project. Give a unique name to the service which will finally be given to the users.

Dataset stakeholders

For each of the above categories, appointments should be given, contact point should be defined, and a brief description shoud be given:

- Dataset producers
  - Technical team of the instrumental project
  - Spatial agencies
  - Other partners providing integrated data
- Management, ie. decision-making authorities
  - Scientific management of the instrument
  - Providers of funds
  - Other partners
- Dataset users
  - Scientific team of the instrumental project
  - Scientific partners
  - Restricted community
  - General public

5.1.3 Description of the dataset

General description

- Description of the dataset content
  Note that a dataset is not only made of the data but also all related elements: format, metadata, documents, tools, access interfaces, databases, access rights and user rules, etc. A dataset consists of a set of packets in the OAIS sense.
- Purpose of the requested service
- Origin of the dataset:
  - in the context of IDOC-DATA, the data of a dataset are generated by an instrument of a satellite or a suborbital (ground based or air-borne) experiment
  - processing levels of the dataset
- Structure of the dataset
  - Data model
  - Associated metadata

Description of the metadata

- Describe all metadata
- Are the metadata enough to describe the data, their completeness and their understanding? If not, describe the quality control check to ensure or mitigate this.
- Are the formats of the metadata dedicated or commonly used in the community?
- Is there any metadata that can be used to allow the archiving or the service to the community? Should the answer be positive, a production pipeline will have to be implemented? The implementation of this service has to follow the recommendations of [RD5]

Possibly associated data
Associated data might be associated to the dataset. By definition, those associated data are not produced by the instrument but are nevertheless essential for the use of the dataset (e.g. the pointing data). For these associated data, provide the same information as for the main dataset.

5.1.4 Level of requested curation
Whatever the service to be implemented on the dataset (new data ingestion, new service on an existing dataset, archiving a dataset), give a brief description on the level of requested curation:

- None: the data are to be used without any change
- Basic: brief checking, addition of basic metadata or documentation
- Advanced: conversion to new formats, enhancement of documentation

5.1.5 Data access
Specify if there are licences applicable to the data access. If so, indicate the licence agreements, conditions of use, and processes to ensure their management.

5.1.6 Ethics and confidentiality
Describe the processes to be implemented to ensure that the dataset properties are adequately disclosed:

- Owner/Organization of the original dataset
- Name of the mission
- Name of the project at IDOC-DATA,
- References/affiliation of the scientists or scientific collaboration having participated to the enhancement of the data,
- etc

Conversely, describe the limits of the disclosure, and the processes to implement it and to mitigate the associated risks.

5.1.7 Quality of the dataset
The dataset to be handled at IDOC-DATA might be stained by inadequate quality or completeness. Describe the procedure to be applied in order to assess this quality. Describe the mechanisms for the users to assess this quality.

5.1.8 Dataset discovery
Should the IDOC-DATA requested service allows it, the way the data can be “discovered” has to be described, including (but not only):

- Access by external robots, internal search functions
- Connections to other datasets or catalogues
- Registration and participation in virtual observatories
- Way(s) the dataset will be citated and credited
- DOIs creation (granularity to be specified)

5.1.9 Dataset reuse

Describe how to ensure dataset reusability cause either by a new processing of the dataset to be applied or the migration of the dataset for reason of changes of technology/hardware/software evolutions with time.

5.2 ANNEXE : PROCEDURE TO PREPARE : DATASET ENUMERATION FORM (DMP-2)

5.2.1 External dataset content

- Are there any software tools (eg. compression/decompression software) associated with the dataset to be ingested?
- Is the dataset conforming to standards for data organization, naming and characterization?
- Does the structure of the dataset allow interoperability with other IDC datasets?
- How is the documentation associated to the dataset structured (eg. global to the dataset versus documentations per classes of data)?
- Does the dataset have (and/or shall have) a DOI –or equivalent nomenclature?
- Which rules, procedures, automations were used to build and validate the creation of the original dataset?

5.2.2 Property of the dataset

- What are the rights on the dataset granted to IDOC-DATA:
    - transform data format?
    - modify the structure (organization, associated databases...) of the dataset?
    - add to dataset (enrichment of metadata, ..)?
    - modify or delete all or part of the dataset?

5.2.3 Size and time requirements

- Provide the useful numbers for the expected quantities and bandwidth.
- Detail the dataset transport methods.
- Give the requested timeline and durations for the dataset reception, production (if any)
- Explain the expected needs in term of time constraints for providing a basic access to the data to users.

5.2.4 Criticality, availability and expected security of the dataset

Note that answers to the following items allow to set the degree of reliability and the performance to implement the dataset, as well as access permissions.

- Is the dataset unique ?
- Is the complete loss of the dataset acceptable?
- What would be the human and material cost and the duration of its recreation in case of loss?
- How long the dataset unavailability is tolerable?
- What is the size (number of people in the community concerned or number of accesses per day for example) and the quality (specialists, non-specialists, general public,..) of the population[s] who [will] access to the data and what is the average size of these accesses ?
- Are there different accesses (authorizations, functionalities, fraction of data,..) according to the populations identified?

5.2.5 Lifecycle of the dataset

- What is the estimated working life of the dataset? (access will be made during this period)
- Will the dataset evolve over time and is it necessary to keep traces/versions of these evolutions?
- Is the dataset intended to be archived beyond its period of activity?

- Are there any aspects in the accesses to the dataset that can modify its accessibility (specific and/or proprietary softwares)?

Are there other elements essential to the scientific use of the data (semantic aspects or dictionaires to clear up ambiguities, links to other datasets useful for data interpretations,..)?

5.3 ANNEXE :ORDER OF MAGNITUDE OF COSTS ACCORDING TO DATASET ENVIRONMENT : IMPLEMENTATION MATRICES

Given the answers to the previous questions, the two following matrices are used to set the target infrastructure to the constraints and resources of the dataset:

- Human and financial resources
- Data volume, duration of data transfer between sites

5.3.1 « Availability / Performance » matrix

	1	2	3	4	5
Availability / Performance matrix implemented at IDOC	**Data always accessible, many accesses**	**Data always accessible**	**Data allowing unavailability of x hours**	**Data allowing unavailability of x days**	**Data « easily » reconstructed**
Redundancy	Automatic switch	Automatic switch	Switch after intervention	Redundancy insured by backups	Reconstruction
Flow/ capacity	High/	Standard/ important	Standard/ important	Standard/ important	Standard/ important
Cost	Very high	high	high	high	standard

5.3.2 « Backup / Historization » matrix

Backup / Historization matrix implemented at IDOC	**Critical data with need of historization or unique data**	**Critical data**	**Standard data**	**Data « easily » reconstructed**
Backup	Reliable support	Reliable support	Possible partial backup	no
Historization	Day/week/month /year	Archive by volume	no	no
Cost	Very high	high	standard	null

Examples	Messaging IDOC databases	Software development		Computer workstations

The choice of the column of the matrix most adapted to the needs and resources of the project is carried out bearing in mind the scale of cost involved.

5.3.3 Example of approximate cost calculation using implementation matrices

For an evaluated dataset, it was decided to place it in column 3 of the « Availability / Performance» matrix, resulting in a « cost » of about : 3
If for the « Backup / Historization » matrix, it is requested that only « level 2 » data representing 1/4 of the overall volume of the dataset should be historized and that this history be limited to 6 versions, then this request resuts in an additional cost of 1 + 6/4.
In total the evaluation of the two matrices results in a « cost » of 5,5.
If the cost of a « standard » storage is 100 per Tb, the total cost of a Tb of the dataset under the conditions allowing to respect the demands expressed would be approximately 550. It is understood that the volumes involved have an impact on costs, but the voluntary pooling of storage resources within IDOC limits this fluctuation.

5.4 ANNEXE : SUPPLEMENTARY NOTES

5.4.1 Procedure of new data integration in IDOC

The two chapters “Procedure to prepare” are also available in the form of a web questionnaire:
http://sondage.ias.u-psud.fr/index.php/survey/index/sid/33435/newtest/Y/lang/fr