IAS
Institut d'Astrophysique Spatiale
Orsay

-

IDOC Guidelines for Data integration

IDOC-OD-004

Préparation

	Nom et Fonction	Date
Rédacteurs	Gilles Poulleau	Mars 2021
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

SOMMAIRE

This document is IAS propriety

1 SCOPE OF THE DOCUMENT

This document is related to the « IDOC-OD-008 IDOC Guidelines 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.

This document is IAS propriety

2 APPLICABLE DOCUMENTS

	Référence	**Titre**
AD1	IDOC-LI-000	IDOC item list

3 REFERENCE DOCUMENTS

	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

This document is IAS propriety

4 GENERAL PRINCIPLES APPLICABLE TO DATASET MANAGEMENT

4.1 FAIR PRINCIPLES

IDOC 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 measureable set of principles that now refer to as the FAIR Data Principles, IDOC 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.

4.2 DATA INTEGRITY AND AUTHENTICITY

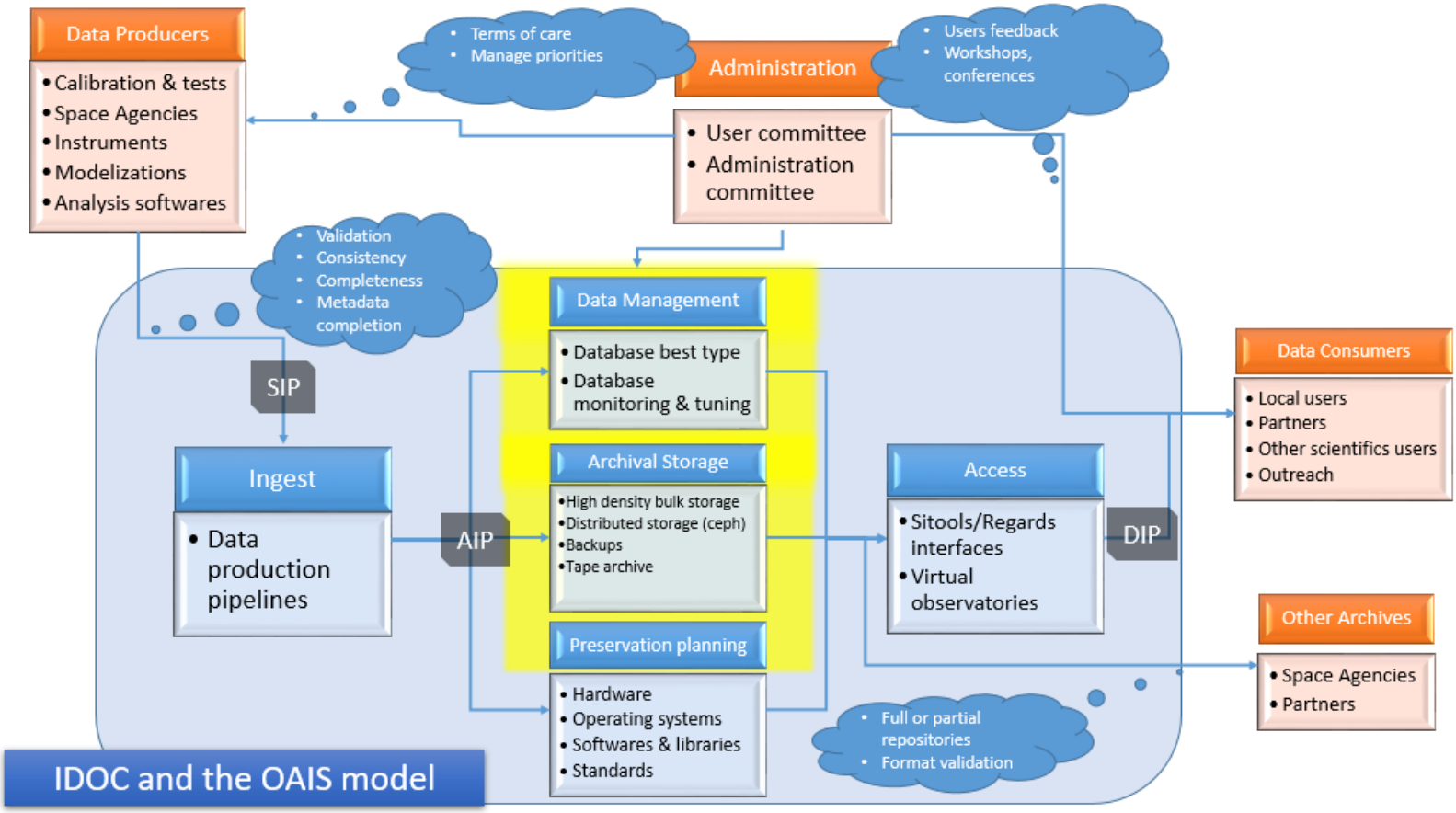
Whatever the service IDOC implements, and independently of the possible enhancement on the datasets, the initial dataset is kept unchanged. This means that all level of dataset management assume that potential added value and curation are only made on copies of those originals.

Moreover, IDOC'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.
This document describes how to build the yellow underlined part of the IDOC application of the OAIS model in the figure below.

4.3 CYCLING CURATION MODEL

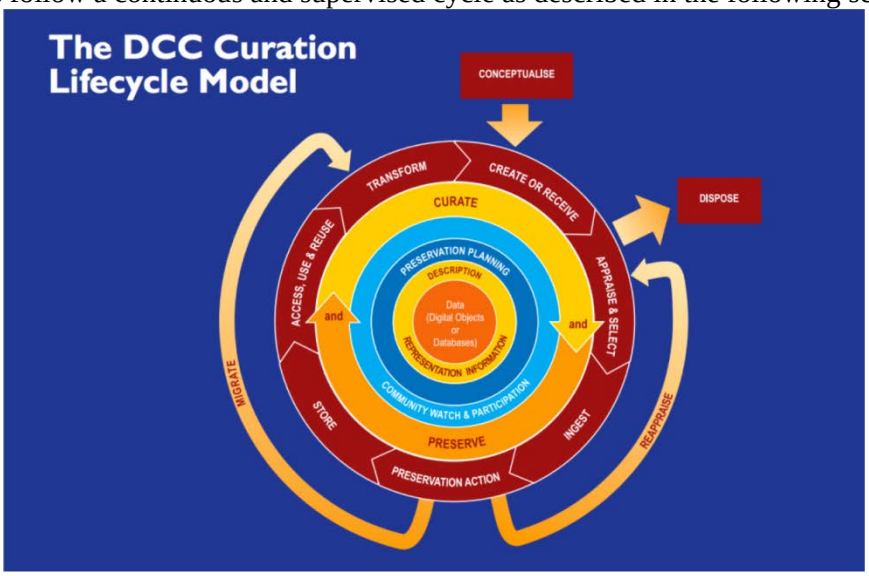
Digital content (set of formatted and organized informations) preservation must comply with:

- Stability: ability to have the same result along time,
- Referencing: the location of the information is predictable,



- Certified origin: informations are produced by successive certified processes,
- Context: each bit of information has a context allowing its understanding.

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



The IDOC guidelines document [RD6] lists the topics to be answered, checked and validated for implementing a data preservation service at IDOC. “Related questions” (see the document) allow specifying the expected implementation of the preservation. Answers to these questions must be regularly revised (minimum every two years) to ensure the cyclic aspect of the monitoring. Along time this « curative » strategy also allows to enrich the content.

4.4 IDOC PRACTICES GIVEN THE DATA PROCESSING LEVEL

Data products at IDOC 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.

This document is IAS propriety

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 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 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.

This document is IAS propriety

5 DATA MANAGEMENT PLAN

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.
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, 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.

5.1 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 will be set.

This document is IAS propriety

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

6.1 CONTEXT WITHIN IDOC

Dataset name
Define the dataset name and its acronym. Pay attention to the name which will be internally to IDOC 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

6.2 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, 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

This document is IAS propriety

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.

6.3 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

6.4 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.

6.5 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,
- 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.

6.6 QUALITY OF THE DATASET

The dataset to be handled at IDOC 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.

6.7 DATASET DISCOVERY

Should the IDOC 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)

6.8 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.

This document is IAS propriety

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

7.1 EXTERNAL DATASET CONTENT

- Is 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 has (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?

7.2 PROPERTY OF THE DATASET

- What are the rights on the dataset granted to IDOC:
  - 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?

7.3 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.

7.4 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?

7.5 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,..)?

This document is IAS propriety

8 ORDER OF MAGNITUDE OF COSTS ACCORDING TO RESPONSES : 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

8.1 « 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

8.2 « 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.

This document is IAS propriety

8.3 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.

This document is IAS propriety

9 SUPPLEMENTARY NOTES

9.1 IDOC SERVICE: DATA SET MANAGEMENT

9.1.1 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

This document is IAS propriety