Data Management Plan

IAGOS Data Centre

PUBLIC

16 June 2021

Version: 1.0

damien.boulanger@obs-mip.fr
www.iagos-data.fr

Contents

1 Introduction

In-service Aircraft for a Global Observing System (IAGOS) is one of the main European Research Infrastructure (ERI) in the environmental domain. IAGOS aims to provide long-term, regular and spatially resolved in situ observations of the atmospheric composition. IAGOS observational platforms are mobile commercial aircraft hosting IAGOS sensors that produce data based on a regular measurement during all the flights performed by the aircraft. These platforms perform measurements of reactive and greenhouse gases, cloud particles, aerosols and also meteorological variables by applying state-of-the-art in situ measurement techniques under consideration of harmonized, standardized, and quality controlled instrumentation, operation procedures and data retrieval schemes. The fleet is currently composed of 8 commercial aircraft.

The data measured by IAGOS sensors is labelled as IAGOS-CORE. The data centre also manages data acquired during the former project MOZAIC with similar sensors (labelled IAGOS-MOZAIC) and by the current project CARIBIC with similar sensors (labelled IAGOS-CARIBIC).

1.1 The mission of the IAGOS Data Centre

The mission of the IAGOS Data Centre (IAGOS-DC) is to collect, archive, produce and provide access to well documented and traceable IAGOS data products, including digital tools for data quality control, analysis, visualisation and research. As a tool for science, the highest priorities for the IAGOS-DC are to maintain and increase the availability of IAGOS observational data and data products relevant to climate and air quality research for all interested users.

The overall goal of the IAGOS-DC is to provide scientists and other user groups with free and open access to all IAGOS data, complemented with access to innovative and mature data products, together with tools for quality assurance, data analysis and research. IAGOS data products should be findable, accessible, interoperable and reusable (FAIR), and the data centre work towards fulfilling the FAIR principles. All data collected and produced by IAGOS are linked through the IAGOS Data Portal serving as the access point to all data and related information.

The IAGOS-DC is in charge of the definition of the technical specification of the IAGOS Data Portal, metadata catalogue and common machine-to-machine access interfaces. It makes sure that technical solutions are implemented in such a way that all data hosted and managed are visible and accessible through the portal.

The IAGOS-DC manages the following data products:

- Observational data sets (IAGOS-CORE, IAGOS-MOZAIC and IAGOS-CARIBIC)
- Elaborated data products, produced by the IAGOS-DC or the IAGOS partners, are composite data sets derived from IAGOS observational datasets and other sources (e.g. model outputs, external dataset, etc.)

- Near Real Time (NRT) or Real Real Time (RRT) data sets for partners (e.g Copernicus Atmosphere Monitoring Service)

The IAGOS-DC can also provide hosting for data from similar airborne projects that cannot be able to maintain this service for themselves (e.g. NOXAR program). In the future, some projects could ask for this service but it is not encouraged as it is not the main purpose of the IAGOS-DC.

1.2 Organization of IAGOS

The IAGOS ERI is led by the IAGOS-AISBL (international non-profit association under Belgian law). Its members are public institutional partners from European countries (Germany, France and UK) involving that all funding is public. IAGOS-AISBL defines the data management policy.

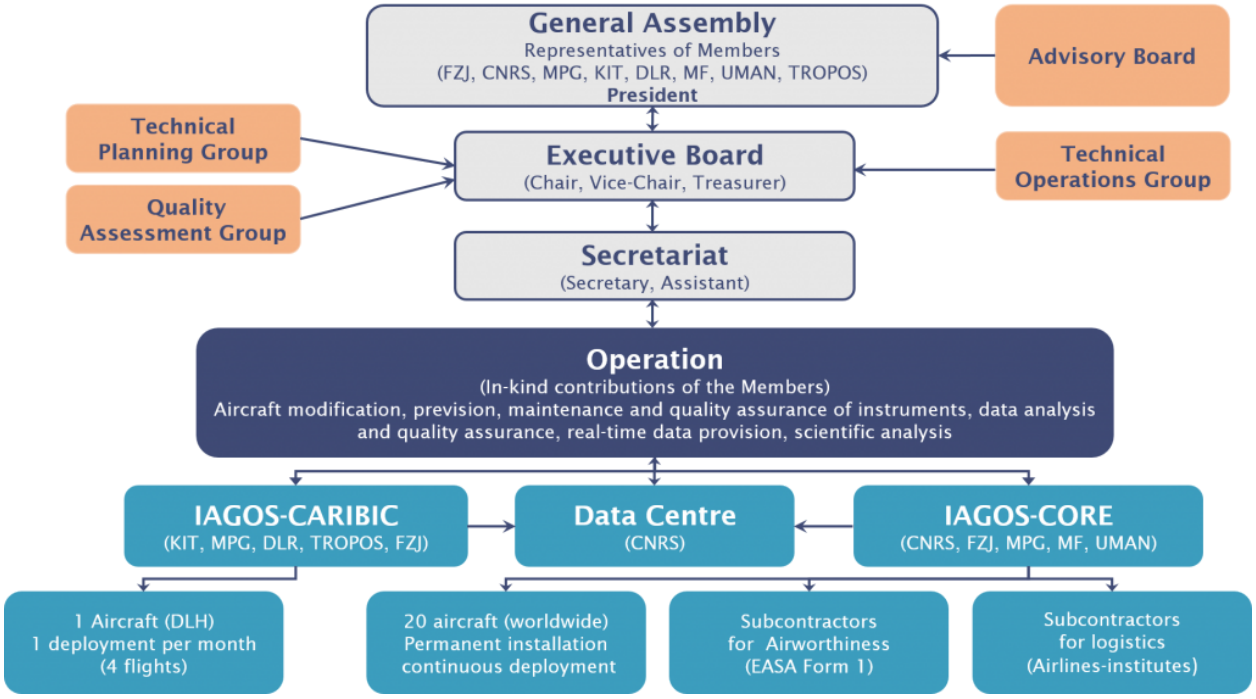


Figure1: Organization chart of IAGOS-AISBL

The members of IAGOS-AISBL are involved in the IAGOS data workflow as data providers and principal investigators of the sensors acquiring the data (see description of the instrumentation: https://www.iagos.org/iagos-core-instruments/). They have the following responsibilities regarding the data:

- **FZJ** (Forschungszentrum Jülich GmbH), Jülich, Germany is responsible for a part of the IAGOS-CORE data and part of IAGOS-CARIBIC data collected by:
  - Humidity sensor (part of Package 1)
  - Odd nitrogen (NOy) Package 2a

    - Nitrogen oxides (NO_x) Package 2b
    - Aerosol Package Package 2c (pending certification for flying on IAGOS-CORE, flying on board IAGOS-CARIBIC)
    - NO2 and aerosol extinction parameter (Package 2e, under development)
- **MPI-BGC** (Max Planck Institute for Biogeochemistry) Jena, Germany is responsible for a part of the IAGOS-CORE data and part of IAGOS-CARIBIC:
    - Greenhouse gases (CO2, CH4, CO, H2O) from Package P2d
- **KIT** (Karlsruhe Institute of Technology), Karlsruhe, Germany is responsible for parts of IAGOS-CARIBIC data:
    - O_3 measurements
    - Water vapour measurements
    - Cloud water measurements
    - VOC measurements (PTRMS)
    - H_2O isotope measurements (TDLAS)
    - CO_2, CH_4 measurements (TDLAS)
- **DLR** (Deutsches Zentrum für Luft- und Raumfahrt e.V.), Köln, Germany is responsible for a part of the IAGOS-CARIBIC data:
    - NO/NOy measurements
- **TROPOS** (Leibniz-Institut für Troposphärenforschung e.V.), Leipzig, Germany: is responsible for parts of IAGOS-CARIBIC data. Tropos hosts the world calibration center for Aerosols and organizes QA/QC workshops for IAGOS-CORE and IAGOS-CARIBIC aerosol measurements.
- **University of Manchester**, U.K. is responsible for a part of the IAGOS-CORE data:
    - Backscatter Cloud Probe (BCP)
- **Météo-France**, Toulouse, France is responsible for the implementation of Real Time data provision.
- **CNRS**: French Public Research Organization, Toulouse, France is responsible for a part of the IAGOS-CORE data and for managing the IAGOS-DC:
    - Ozone and carbon monoxide sensors (part of Package 1)
    - Elaborated products

The creation of a Data Management Group (DMG) has been decided and the DMG will be added in the organization of IAGOS-AISBL, in order to:

- Manage the development of the IAGOS-DC strategy together with the IAGOS Executive Board. It will plan new developments with the IAGOS-DC.
- Establish and monitor the implementation of the IAGOS-DC work programme
- Undertake evaluation and monitor the operations
- Serve as the link to and interact with the data providers
- Ensure the necessary interaction on common technical topics and issues (standards, interoperability, user feedback...)

- Propose changes and further development of IAGOS Data Management Plan (DMP), or discussions and approval by IAGOS executive board before implementation

1.3 Organization of the IAGOS Data Centre

The IAGOS-DC is hosted at the Observatoire Midi-Pyrénées (OMP) in Toulouse, France and is composed of four persons that are part of the OMP Data Service (SEDOO):

- The IAGOS-DC manager
- One developer in charge of IAGOS data workflow
- One developer in charge of the Graphical User Interface (GUI) developments .
- One developer in charge of production of IAGOS elaborated data products

SEDOO is one of the Data and Services Centre of the French Cluster for Atmosphere (AERIS). It includes ten more developers that can be involved in the IAGOS-DC activities as some developments are mutualized.

The IAGOS-DC numerical infrastructure is hosted by the OMP. The IT service of the OMP is in charge of maintaining the numeric infrastructure.

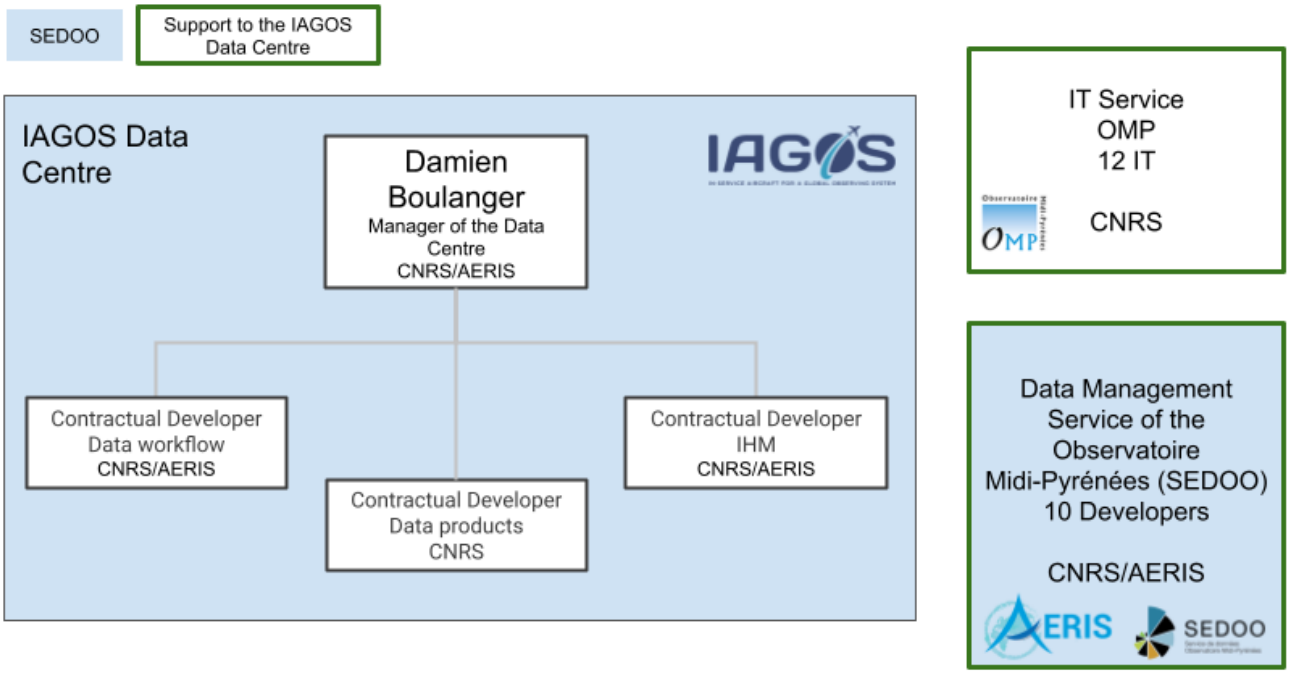


Figure 2 : Human resources of the IAGOS Data Centre

1.4 Related projects

IAGOS is part of the European H2020 project ENVRI-FAIR aiming to implement interoperability within the European environmental research domain. The application of

all the FAIR principles in the IAGOS-DC is currently done in the frame of this project. By the end of the project in June 2023, the IAGOS user's community could be extended to other environmental domains such as Marine, Solid Earth, and Biodiversity.

IAGOS is also part of the European H2020 project ATMO-ACCESS aiming to provide virtual access to the data and services of the Atmospheric Research Infrastructures. New IAGOS services and data products will be implemented in the frame of this project.

1.5 Overall goal of IAGOS Data Management Plan

The IAGOS DMP documents the IAGOS data management life cycle, and the plans for the data collected, processed, generated and published. The goal of the DMP is to describe the present situation and the operational IAGOS-DC.

Furthermore, the DMP also describes the technical solutions agreed, that are currently under implementation, and outline the strategy and development needed towards making IAGOS data FAIR.

The DMP is a living document that will be updated regularly. The goal is to make the DMP accessible for all stakeholders (repository operators, funders, researchers, publishers, infrastructure providers etc.).

This document uses the Horizon 2020 Template.

2 IAGOS data

2.1 Definitions

In the frame of IAGOS, a **dataset** is a collection of data records published or curated by a single source, and available for access or download in one or more formats.

A **data product** is a dataset that is the outcome of a service or process applied to an input dataset. Therefore, the data product is considered more elaborated than the input dataset and may be assigned a new data processing level. Added-value or elaborated data products propose features to enhance the user experience. They are composite datasets that are results of the derivation of observational datasets combined with other datasets (from IAGOS or external).

2.2 Processing levels

IAGOS observational data are submitted to several automatic and manual processes during its life cycle. Each step corresponds to a different processing level.

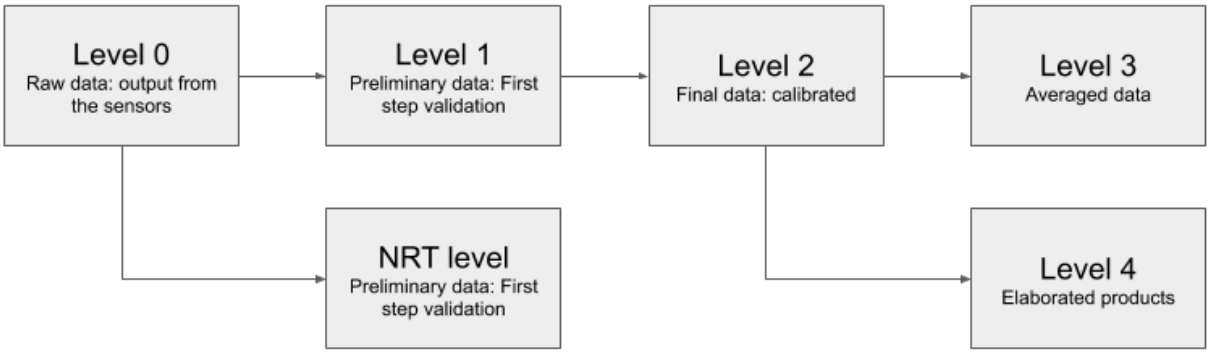


Figure 3: Description of the data processing levels workflow

Level	Description	Type of product
Level 0A	Raw sensor output in sensor or physical units	In situ Observations
Level 0B	Raw sensor output in sensor or physical units with minimum level of quality control	In situ Observations
NRT	Quality assured data with minimum level of automatic quality control provided within three days after the acquisition	In situ Observations

Level 1	Quality assured data with minimum level of quality control by the instrument Principal Investigator or automated	In situ Observations
Level 2	Approved, calibrated and fully quality controlled data product or geophysical variable	In situ Observations
Level 3	Averaged global gridded products	Gridded climatologies (long-term averages) of O3, CO and H2O
Level 4	Elaborated data products derived by post-processing of Level 2 data and data from other sources like ECMWF (re-)analyses, FLEXPART lagrangian model	• Footprints • CO contributions • ECMWF variables • Representation error • PBL profiles

Table1: Definitions of the IAGOS data processing levels

Level 4 products are output from research activities. Partners and/or users can define and prototype some of the level 3 and level 4 data products. The IAGOS-DC takes over the publication of the products and sometimes the production. But the production can stay under the responsibility of the partners.

2.3 Variables

The list of all the variables is available in Appendix C.

3 Data summary at the IAGOS Data Centre

3.1 Purpose of the data production

The primary goal of IAGOS is to produce high quality integrated datasets in the area of atmospheric sciences and provide services on those data. The purpose of the data collection and generation of data products in IAGOS is to provide open access to reactive and greenhouse gases, cloud and aerosols in measurements of high quality, benefiting a large community of scientists involved in atmospheric science and related areas as well as policy makers, private sector, educators and the general public.

The scientific quality of the data is assured by following SOPs (Standard Operating Procedures) and an implemented QA/QC protocol. The SOPs documentation is available on each instrument description on the IAGOS website (https://www.iagos.org/iagos-core-instruments/).

3.2 Relation to the objectives of the project as stated in the statutes of IAGOS-AISBL

As stated in the statutes of IAGOS-AISBL, the main objectives of IAGOS related to the IAGOS-DC activities are :

- to promote science and research by supporting and coordinating research in the field of air pollution and climate change through the establishment and operation of an infrastructure for long-term observations of atmospheric composition and atmospheric properties on the global scale which uses a fleet of in-service aircraft of internationally operating airlines
- to facilitate coordination of activities, support the development of relevant technologies and disseminate knowledge on atmospheric composition to the international community
- the deployment of high-tech instruments for regular airborne observations of atmospheric parameters for atmospheric and climate research and associated fields
- maintaining of and offering publicly accessible databases and other information generated within the AISBL
- dissemination and exchange of information, where applicable via communication/information platforms, including data provision in near-real-time
- joint analysis of the collected data, including joint publications; according to the rules of good scientific practice
- to provide information on the 4D composition and variability of the physical, optical and chemical properties of short-lived atmospheric constituents, from the surface throughout the troposphere to the stratosphere, with the required level of precision, coherence and integration
- to provide efficient open access to IAGOS data and services and the means to effectively use the IAGOS products, according to the FAIR principles

- to ensure and raise the quality of data and use of up-to-date technology used in the RI and the quality of services offered to the community of users

High quality observational data needs to be supported by:

- Documentation of archiving procedures and access to level 0 to level 4 data
- Documented and traceable processing chain of level 0 data
- Documented, traceable processing and long-term archiving and preservation of all IAGOS level 1 to level 4 data and data products
- Access to IAGOS data, data products, and digital tools through a single point of entry
- Documentation of data, data flow, citation service, and data attribution, including version control, data traceability, and interoperability,

3.3 Main users of IAGOS data

IAGOS produces data essential to a wide range of communities:

- IAGOS scientists that are in charge of the instrumentation and responsible for the acquisition and qualification of the data. The IAGOS-DC develops and maintains services dedicated to these data providers.
- Atmospheric science research communities world-wide
- Climate and air-quality, observational/ experimental/ modelling/ satellite communities, national and international research programmes and organisations
- Environmental science research communities and communities from other neighbouring fields: hydro-marine, bio-ecosystem, geosciences, space physics, energy, health, and food domain, to study interactions and processes across different disciplines
- Operational services, National weather services, climate services for model validation, weather and climate analysis and forecasting
- Airlines
- Space agencies for validation and the development of new satellite missions
- National and regional air quality monitoring networks and environmental protection agencies for air quality assessments and validation of air pollution models
- Policy makers and local/ regional/ national authorities for climate, air-quality, health and atmospheric hazards related information for decision making and policy development
- Similar programs asking for repository services

IAGOS has currently more than 730 individual registered users.

IAGOS provides Near Real Time data (NRT, within 3 days) to the Copernicus Atmosphere Monitoring Service (CAMS) for model validation. Real Time data (RRT)

provision is planned in cooperation with Copernicus. That data will be used for real time monitoring purposes (air quality, etc.).

3.4 IAGOS Data Workflow

3.4.1 Data processing

IAGOS-CORE data is automatically transmitted by the instruments to the IAGOS-DC as soon as the aircraft has landed. All the data providers download the original data on the reception server at the IAGOS-DC. Once the data is processed, they upload it on the same reception server.

Observational data sets are provided in different formats according to the data providers. Formats can be NetCDF or ASCII CSV. Automatic and manual procedures, executed by the partners and the IAGOS-DC, check the quality of the data, extend the metadata (provenance, etc.) and format the data sets in a common data format (NetCDF).

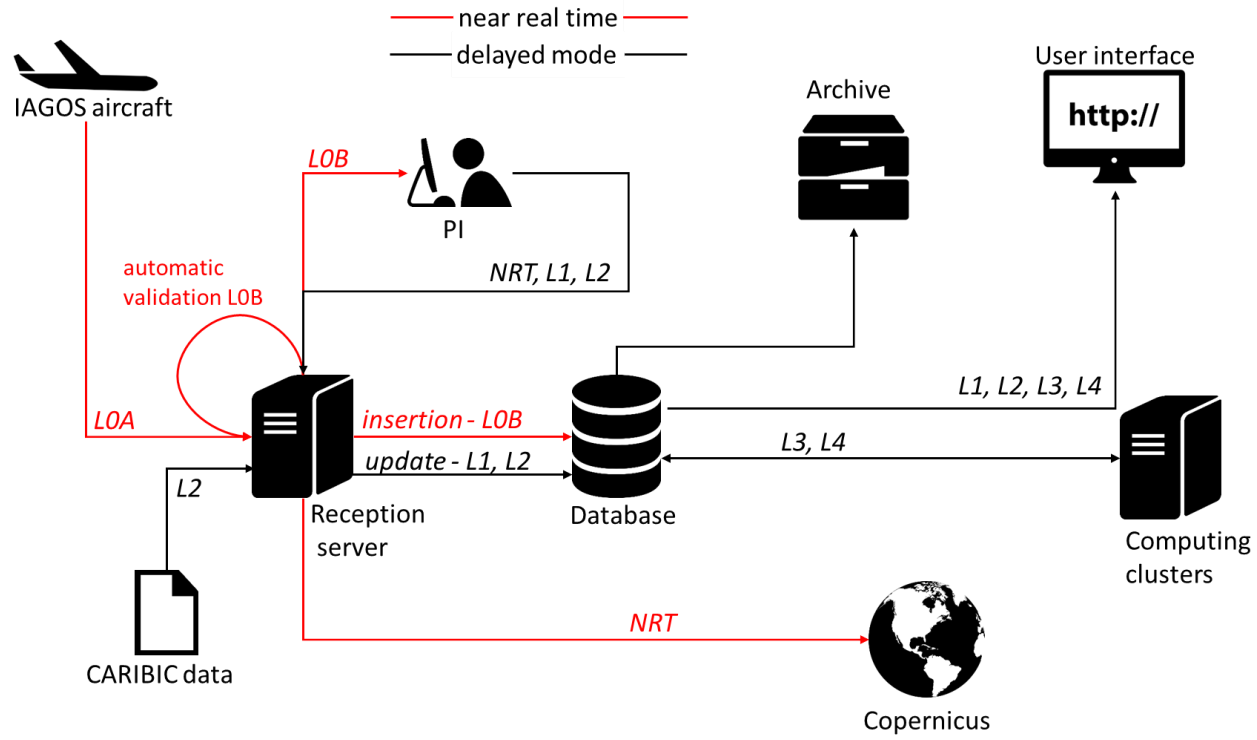


Figure 4: IAGOS data workflow

All versions of the data sets are stored (originals, upgraded after QA/QC procedures, new versions) and documented. Some intermediate quality controlled data sets are produced (Level 1). Only fully qualified data sets are accessible by the scientific community (Level 2 to level 4). Others are only available to the IAGOS partners and Copernicus services.

All the observations acquired during a flight are merged in common files in order to provide homogeneous data sets to the users. The figure below presents the workflow of the homogenisation. At each data processing level a merged file is generated as soon as a data provider provides its data processed for this level. The file is updated as soon as new processed data is uploaded.

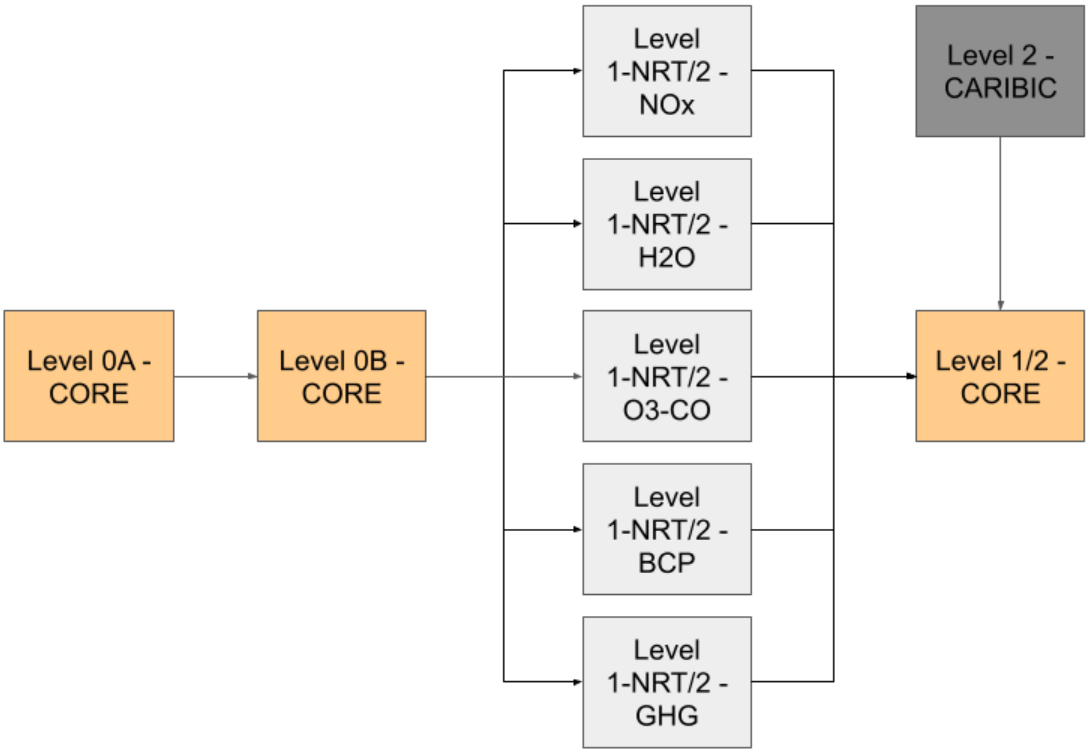


Figure 5: Data files qualification and homogenisation workflow

The IAGOS-DC also produces elaborated data products (Level 3 and Level 4) to improve the use of the datasets and offer a better experience to the users. They are managed in the same way as observational data sets.

3.4.2 Data publication

The IAGOS Data Portal provides data discovery and download as well as added-value services. Data and metadata are also accessible through standard machine-to-machine services. See section 4.2.2 for more information.

3.5 CNRS

3.5.1 Responsibilities

CNRS is in charge of the acquisition and qualification of Ozone and Carbon monoxide measurements and of the acquisition and qualification of the parameters measured by

the aircraft (position: latitude, longitude, altitude; meteorological fields such as wind speed, etc.).

CNRS is also the producer of the following elaborated data products:

- Climatologies (Level 3) : under the responsibility of the IAGOS-DC
- PBL profiles (Level 4)
- ECMWF variables (Level 4): integrated in the IAGOS-DC workflow
- CO contributions (Level 4): integrated in the IAGOS-DC workflow

3.5.2 Types and formats of data

Data files of levels 0 to 2 are provided to the Data Centre in a homemade CSV file format including a set of required metadata.

Data files of levels 3 and 4 are provided in NetCDF v3 format and will be provided in NetCDF v4 when the new Data Portal will be delivered.

3.5.3 Re-use of existing data

ECMWF reanalysis are used to generate Level 4 products (ECMWF variables and CO contributions).

3.5.4 The origin of the data

Observational data is derived from raw data acquired by the instruments.

Level 3 data is derived from Level 2 data.

Level 4 data is derived from Level 2 data and/or model outputs (ECMWF, FLEXPART).

3.5.5 The expected size of the data

The expected size of the data is proportional to the number of flights operated by the IAGOS fleet. The number of flights is also dependent on the number of aircraft in operation. This number can fluctuate according to some factors (negotiations with the airlines, aircraft end of life, etc.).

In the whole IAGOS observation period (including the former projects MOZAIC), there have been around 62000 flights from 1994 to 2020, with an average number of 462 flights per year and per aircraft.

3.5.5.1 Observational data

For Package 1 and Ozone and CO observational data files, the estimated size of the data for a flight is 2.1 Mo.

	aircraft	flights	Total flights	Volume Go
2020	7	3235	62193	133.20
2021	7	3235	65428	140.13
2022	8	3697	69125	148.04

2023	9	4159	73284	156.95
2024	10	4621	77905	166.85
2025	11	5083	82988	177.74
2026	12	5545	88534	189.61
2027	13	6007	94541	202.48
2028	14	6470	101011	216.33
2029	15	6932	107942	231.18
2030	16	7394	115336	247.01

Table 2: Expected evolution of the archive size - P1 + O3 and CO

3.5.5.2 Homogeneous final products

The IAGOS-DC merges all the variables into one single file per flight for all the processing data levels. The estimated size of the data generated for a flight is 3 Mo.

	aircraft	flights	Total flights	Volume Go
2020	7	3235	62193	185.15
2021	7	3235	65428	194.78
2022	8	3697	69125	205.79
2023	9	4159	73284	218.17
2024	10	4621	77905	231.93
2025	11	5083	82988	247.06
2026	12	5545	88534	263.57
2027	13	6007	94541	281.46
2028	14	6470	101011	300.72
2029	15	6932	107942	321.35
2030	16	7394	115336	343.36

Table 3: Expected evolution of the archive size - Final observational dataset

3.5.5.3 Elaborated products

For L4 products (ECMWF variables and CO contributions), the estimated size of the data for a flight is 2.8 Mo.

	aircraft	flights	Total flights	Volume Go
2020	7	3235	62193	172.48
2021	7	3235	65428	181.45
2022	8	3697	69125	191.70
2023	9	4159	73284	203.24
2024	10	4621	77905	216.05
2025	11	5083	82988	230.15
2026	12	5545	88534	245.53
2027	13	6007	94541	262.19

2028	14	6470	101011	280.13
2029	15	6932	107942	299.35
2030	16	7394	115336	319.86

Table 4: Expected evolution of the archive size - L4

3.5.6 Data utility

3.5.6.1 Ozone and Carbon monoxides observations

Observations can be used for models and satellite data validation.

3.5.6.2 L3 and L4 products

Level 3 and Level 4 are of particular interest for the community of scientists working on assessment reports (e.g. TOAR) and global model evaluation (e.g. CAMS, CCMI) in providing essential and unique diagnostics. Such IAGOS products reveal the statistical robustness of the data set and allow testing the ability of the model to reproduce for example the good ozone for the good reasons (as long as they reproduce the good sources of precursors).

3.5.7 Outline of data life cycle (workflow)

Details of the data life cycle and workflow diagrams for data production managed by CNRS can be found in Appendix B.1.

3.6 FZJ

3.6.1 Responsibilities

FZJ is in charge of the acquisition and qualification of the measurements for:

- Water vapor (P1)
- NOx (P2)
- Aerosols (P2)

3.6.2 Types and formats of data

The files formats for each type of data are:

- Water vapor: NetCDF
- NOx: Nasa Ames
- Aerosols: Nasa Ames

3.6.3 The origin of the data

The origin of the data is derived from instrument raw data.

3.6.4 The expected size of the data

The expected size of the data is proportional to the number of flights operated by the IAGOS fleet. The estimated size of the water vapor data for a flight is 2.3 Mo. So far the size of NOx data is negligible and aerosols data not yet available.

	aircraft	flights	Total flights	Volume Go
2020	7	3235	62193	146.58
2021	7	3235	65428	154.21
2022	8	3697	69125	162.92
2023	9	4159	73284	172.72
2024	10	4621	77905	183.61
2025	11	5083	82988	195.59
2026	12	5545	88534	208.66
2027	13	6007	94541	222.82
2028	14	6470	101011	238.07
2029	15	6932	107942	254.41
2030	16	7394	115336	271.84

Table 5: Expected evolution of the archive size - Water vapor

3.6.5 Data utility

Observations can be used for models and satellite data validation.

3.6.6 Outline of data life cycle (workflow)

Details of the data life cycle and workflow diagrams for data production managed by FZJ can be found in Appendix B.2 and Appendix B.3.

3.7 MPI-BGC

3.7.1 Responsibilities

MPI-BGC is in charge of the acquisition and qualification of the measurements for Greenhouse Gases observations (carbon monoxide, carbon dioxide, methane and water vapor).

3.7.2 Types and formats of data

The format of the files is ASCII CSV.

3.7.3 The origin of the data

The origin of the data is derived from instrument raw data.

3.7.4 The expected size of the data

The expected size of the data is proportional to the number of flights operated by the IAGOS fleet. So far the size of Greenhouse gases data is negligible.

3.7.5 Data utility

The data will be used by CAMS and C3S for model validation; and for satellite validation.

3.8 University of Manchester

3.8.1 Responsibilities

The University of Manchester is in charge of the acquisition and qualification of the measurements for cloud particles measured by the BCP.

3.8.2 Types and formats of data

BCP data files are provided in Nasa Ames format.

3.8.3 The origin of the data

The origin of the data is derived from instrument raw data.

3.8.4 The expected size of the data

The expected size of the data is proportional to the number of flights operated by the IAGOS fleet. The estimated size of the BCP data for a flight is 0.2 Mo

	aircraft	flights	Total flights	Volume Go
2020	7	3235	15543	3.30
2021	7	3235	18778	3.99
2022	8	3697	22013	4.77
2023	9	4159	25247	5.65
2024	10	4621	28482	6.64
2025	11	5083	33103	7.72
2026	12	5545	37725	8.89
2027	13	6007	42346	10.17
2028	14	6470	46967	11.54
2029	15	6932	51588	13.01
2030	16	7394	56209	14.58

Table 6: Expected evolution of the archive size - BCP

3.8.5 Data utility

Observations can be used for models and satellite data validation.

3.9 KIT (CARIBIC)

3.9.1 Responsibilities

KIT is in charge of the acquisition and qualification of the measurements for the IAGOS-CARIBIC project.

3.9.2 Types and formats of data

The data is provided as a global file containing all the observations for a flight in NASA Ames format.

3.9.3 The origin of the data

The origin of the data is derived from instrument raw data.

3.9.4 The expected size of the data

The expected size of the data is proportional to the number of flights operated by the IAGOS-CARIBIC aircraft. The estimated size of the CARIBIC data for a flight is 1 Mo

	aircraft	flights	Total flights	Volume Go
2020	1	48	541	0.57
2021	1	48	589	0.62
2022	1	48	637	0.67
2023	1	48	685	0.72
2024	1	48	733	0.77
2025	1	48	781	0.82
2026	1	48	829	0.87
2027	1	48	877	0.92
2028	1	48	925	0.97
2029	1	48	973	1.02
2030	1	48	1021	1.07

Table 7: Expected evolution of the archive size - CARIBIC

3.9.5 Data utility

Observations can be used for models and satellite data validation.

4 Data management at the IAGOS Data Centre

4.1 Data findability

4.1.1 Discoverability of data

The datasets managed by the Data Centre are of two sorts:

- final citable datasets: Level 2 to Level 4
- internal datasets: Level 0 to Level 1

All the datasets are documented by rich metadata including the following information:

- spatial and temporal extent
- contacts information
- keywords (name of the variables, etc.)
- instruments and platforms identification
- provenance information (original dataset, history of processing, etc.)

Only the final data sets are assigned with a DOI. So far the following DOI have been assigned:

- Observational Time series : https://doi.org/10.25326/06
- Observational Vertical Profiles : https://doi.org/10.25326/07
- CO contributions : https://doi.org/10.25326/3
- PBL-referenced profiles of O3 and CO : https://doi.org/10.25326/4

A DOI is also assigned to the IAGOS Data Portal: https://doi.org/10.25326/20.

For the internal datasets non-persistent homemade identifiers are generated. A new system based on ePIC handles is currently under implementation.

4.1.2 Naming conventions

Two naming conventions are used in the metadata profiles and the data files, to represent the variables provided in the datasets, the instruments and the platforms:

- GCMD (Global Change Master Directory)
- CF convention (Climate and Forecast)

In the frame of ENVRI-FAIR, a vocabulary is under implementation for the Atmosphere domain. Once this vocabulary is ready, IAGOS will align its controlled vocabularies to it.

4.1.3 Search keyword

To optimize the possibilities for re-use the following keywords are provided in the metadata for the search:

- name of the variables (GCMD and CF standard names)
- name of the instruments (GCMD)
- name of the platform (GCMD)

- scientific domain keywords (GEMET)

4.1.4 Versioning

So far, versioning information of the data sets and the software used to process them is stored in the database but not yet provided to the users.

The IAGOS-DC distinguishes between two forms of alteration involving the generation of a new version:

- New version and therefore a new dataset: when there is a change to data
- Minor change: when there is a change to metadata, descriptive documents or supplementary files.

A new version is deposited as a new dataset and will therefore receive its own persistent identifier. The new and the previous dataset are cross-referenced in their respective descriptive metadata. Alternatively, when there is a minor change, this change is documented in the administrative metadata; no new persistent identifier is minted.

A new system is currently under implementation. It will provide all the versioning and provenance information to the users in the data files and on the datasets landing page. The PROV-O standard will be used to represent this metadata. A dedicated triple store will be implemented to store those metadata.

4.1.5 Metadata standard

Metadata are stored in a MongoDB database in a homemade pivot format and are mapped to multiple metadata profiles:

- ISO 19115 / Inspire
- Datacite

The conversion to WMO metadata profiles (WIGOS and WIS) and DCAT is under implementation.

ISO 19115 metadata profiles are stored in a dedicated Geonetwork server allowing their harvesting through the standard protocols CSW and OAI-PMH. Metadata is also available in JSON through a REST API.

The metadata endpoints are:

- CSW: http://catalogue2.sedoo.fr/geonetwork/srv/eng/csw-iagos?service=CSW&version=2.0.2&request=GetCapabilities
- OAI-PMH: http://catalogue2.sedoo.fr/geonetwork/srv/eng/oaipmh?verb=ListRecords&metadataPrefix=oai_dc

- REST: https://services.iagos-data.fr/prod/swagger-ui.html

Metadata are also available in the data files following the NetCDF conventions CF and ACDD.

4.1.6 Indexation

IAGOS metadata is indexed on the following portals:

- the IAGOS Data Portal as main entry point: https://doi.org/10.25326/20
- the AERIS catalogue: https://www.aeris-data.fr/catalogue
- re3data: https://www.re3data.org/repository/r3d100013365
- FAIRSHARING: https://fairsharing.org/FAIRsharing.caOJPM
- CatRIs: https://www.portal.catris.eu/service/IAGOS.iagos_data_portal
- Datacite

The registration to the following portals is under implementation:

- GAWSIS (WIGOS/GAW) from WMO
- COPERNICUS (in the frame of ENVRI-FAIR)
- EOSC (in the frame of ENVRI-FAIR)
- GEOSS (in the frame of ENVRI-FAIR)

4.2 Data accessibility

The purpose of the data collection and generation of data products in IAGOS is to provide open access to all of them. A guiding principle is that all IAGOS data should be readable for both humans and machines using protocols that offer no limitations to access

4.2.1 Openness

Only the datasets from level 2 to level 4 are made openly available to the public. Level 0 and level 1 data sets will only be available for the members of the project as their quality is limited. NRT datasets are only provided to Copernicus on the basis of a contract.

4.2.2 Data access

IAGOS data is only stored by the IAGOS-DC. It is accessible in a human way through the IAGOS Data Portal via HTTPS protocol. As the metadata is or will be registered in many portals or catalogues (see above), users will be redirected to the IAGOS Data Portal.

Data is also provided in a machine-readable way through REST API endpoints. The implementation of the OpenDAP and WxS protocols is in progress with the installation of a THREDDS server.

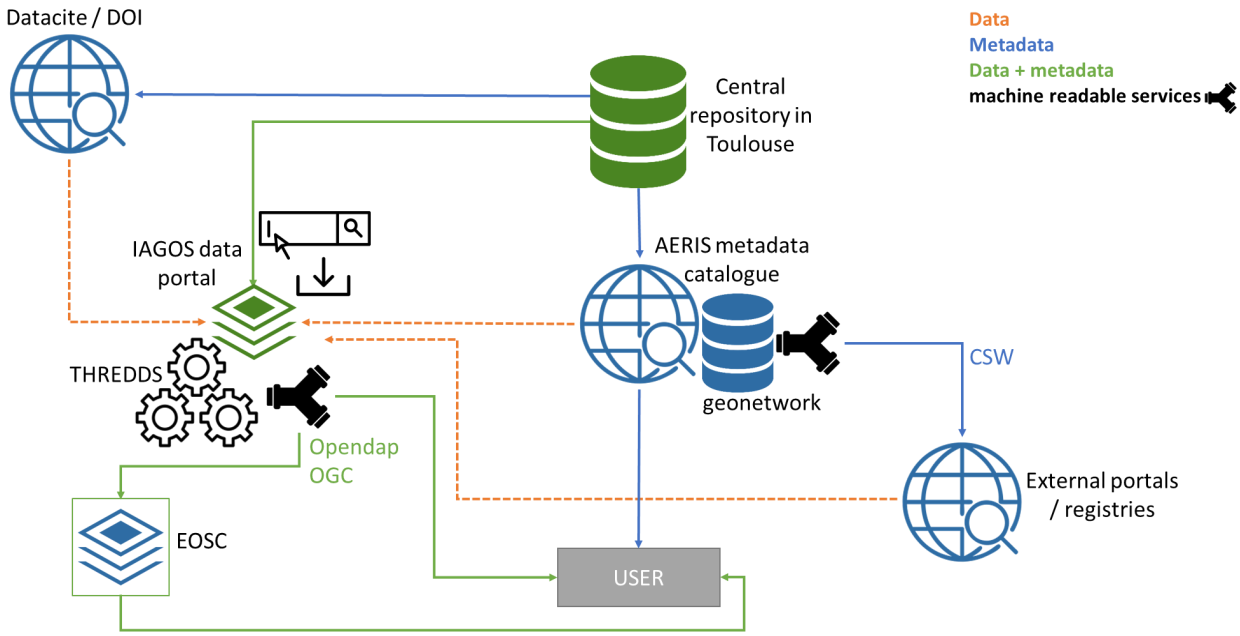


Figure 6: Metadata and data access

4.2.3 Documentation

The data is available in NetCDF of NASA Ames formats that are widely used by the user's community. The files follow the corresponding conventions and their conformity can be checked using the following tools:

- NetCDF (CF and ACDD convention): https://podaac-tools.jpl.nasa.gov/mcc/
- NASA Ames (1001 convention): https://archive.ceda.ac.uk/nachecker

The formats are documented on the Data Portal: http://www.iagos-data.fr/#DataFormatPlace:

All the code developed at the Data Centre is saved and versioned on a private gitlab repository: https://git.icare.univ-lille1.fr/. All the code developed by the partners in order to produce the datasets will be available through the landing pages of the software (PID provided in the provenance information).

4.2.4 Restrictions

The data is open but the users need to be registered on the IAGOS Data Portal in order to access the datasets. This is due to the need to provide rich statistical information (data usage, etc.). It also allows to provide added-value services to the users such as subscription to data sets, queries history, etc.

The authentication and authorization system is based on the Single Sign-On solution hosted and maintained by AERIS. It allows the users to authenticate using their ORCID or EDUGAIN accounts.

4.2.5 Sustainability

The availability of the metadata is sustainable as a system is implemented to provide metadata even if the dataset isn't available anymore. All the landing pages will be maintained and checked periodically to detect any interruption of service.

4.3 Interoperability

As a guiding principle, IAGOS should make sure that metadata and data use a formal, accessible, shared and broadly applicable language for knowledge representation in order to facilitate interoperability.

Metadata and data are provided in standards formats that are commonly used in the scientific communities. The use of the vocabularies, mentioned before, will also ensure the interoperability of the data. Those vocabularies will be mapped with the vocabularies used in the environmental domain in the frame of the ENVRI-FAIR project.

A THREDDS Data Server is currently under deployment in order to serve data and metadata in an automated way as netCDF files through the OPeNDAP protocol. In addition, a REST API is proposed for machine-to-machine interaction. The API serves metadata (info, provenance, versions, quality controls, etc.) in JSON format and data (specific files or datasets previously generated) in NetCDF or NASA Ames format.

The interoperability with the ENVRI-Hub that will serve the metadata and data of the ENVRI cluster to EOSC is under implementation in the frame of ENVRI-FAIR.

4.4 Reusability

The guiding principle is free and open access to IAGOS data and data products, and the IAGOS Data Centre will facilitate data re-use by providing free and open access to IAGOS data following the IAGOS Data Policy and the open research data initiative of the European Commission.

4.4.1 Licence

All public datasets are under the IAGOS license that follows the Resolution 40 of WMO (World Meteorological Organisation) policy. Datasets are provided with free and unrestricted access for scientific (non-commercial) use. Reuse of the data is then strongly encouraged in the scientific domain.

The data is available as soon as qualified. The qualification is dependent on the access instruments facilities and then the access to the aircraft that can be limited due to the commercial aspect. The data is commonly available within two to six months after the observations.

A new license in accordance with the FAIR principles is under discussion in the frame of the ENVRI-FAIR project. This new license will be machine actionable.

4.4.2 Software

All the code developed by the Data Centre and the partners will be available through the landing pages of the software (PID provided in the provenance information).

4.4.3 Citation

The citation strategy of the IAGOS-DC is to assign a DOI to the data collections (e.g. Observational Time series). As recommended by the RDA, a fragment added to the DOI will allow to cite a subset of the collection, for instance:

- data of a single flight
- data matching a user's query
- a version of the collection

The solution is currently under implementation. The fragments will be generated and the matching query will be stored in a query store. The landing pages of the datasets/collections will be adapted if a fragment is provided (list of contacts, extents, etc.).

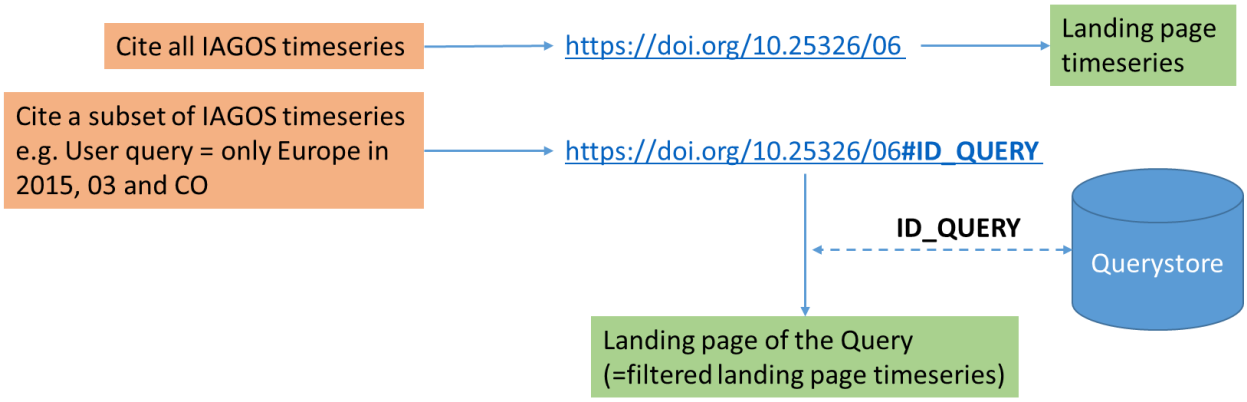


Figure 7: IAGOS data citation strategy

4.4.4 Provenance

In order to increase their reusability, data are completed with rich metadata which are in open access from the website.

A system is currently implemented to provide the full provenance metadata for each dataset. This information will be available on the Data Portal in a human-readable way and in the data files in a machine-readable way. The system is based on the implementation W3C standard PROV-O and will allow to provide full history for all the datasets including: link to the previous versions of the datasets, identifier of the input datasets used to create this dataset, identification of the software, data providers and instruments, etc. A SPARQL endpoint will provide access to the provenance metadata.

5 Allocation of resources

CNRS (French Public Research Organization) has been designated responsible for the Data Centre by the IAGOS-AISBL and delegated the responsibility to AERIS (French Data and Services Cluster for Atmosphere) and the Observatoire Midi-Pyrénées (OMP) in Toulouse, France. The OMP is composed of several Research Units and one Services Unit which hosts one of the four Data and Services Centres of AERIS. The Data Centre is hosted and managed by the Services Unit. OMP and AERIS are partly funded by CNRS, University Paul Sabatier and Météo-France which are all members of the IAGOS-AISBL.

The Data Centre has operated since 2005 and has always had continuous and substantial support from CNRS. Thanks to the recognition at national and european level, IAGOS gets substantial support through project grants.

The full implementation of the FAIR principles in the Data Centre is done in the frame of the European projects ENVRI-FAIR, ATMO-ACCESS and RI-Urbans. It will be finished by the end of the project in June 2023. Combined funding from CNRS and the European Commission will cover the costs of making IAGOS data FAIR.

Operation	Cost	Funding	Type of budget
IT (Data Centre Manager)	90k€ / year	CNRS	Permanent
Hardware	10k€ / year	CNRS	Permanent
IT Data processing	50k€ / year	ENVRI-FAIR	4 years (2019-2023)
IT Web development	50k€ / year	ENVRI-FAIR	4 years (2019-2023)
IT Data products and services	55k€ / year	ENVRI-FAIR ATMO-ACCESS	6 years (2019-2024)
IT Data products and services	55k€ / year	RI URBANS	1 year (2021-2024)

Table 8: Table of costs of the IAGOS Data Centre operations

6 Data security

IAGOS data is stored in a MongoDB database that includes all metadata and data relative to the projects. The data is also stored on disk in a files archive that is used to generate the database. A full backup of the file archive is automatically made several times a day. All versions of the datasets are saved.

The preservation plan is continuous preservation as long as the Data Centre continues to exist and that its expertise is maintained in the long term. IAGOS infrastructure is maintained by AERIS with long-term commitment for archiving and preservation. To avoid lost data, a sustainable archiving is provided by a dedicated service at OMP. The data is duplicated on different geographical sites in Toulouse, France and also at another site in Tarbes, France.

In case the OMP could not maintain the Data Centre anymore, the responsibility could be transferred to another Data Centre within AERIS. If any AERIS data centres would not be able to maintain the Data Centre, the responsibility could be transferred to another IAGOS partner. However, the disappearance of AERIS would be highly improbable as it's strongly supported at national level. The data and services management policy implemented within AERIS, such as containerization of the applications, would facilitate the migration to another partner and would allow to easily transfer the services and maintain the availability and accessibility of the data. The French environmental community is currently organizing under the Environmental Data and Services Cluster called Data Terra. In its frame, a project named GAIA Data is dedicated to the implementation of a common numeric infrastructure for all the environmental domains. The project just started in 2021 and will last eight years. The IAGOS-DC will benefit from the services that will be provided by Data Terra and then improve its sustainability.

In the frame of the FAIRsFAIR project the IAGOS-DC is currently working to achieve the CoreTrustSeal certification by the end of 2021.

7 Ethical aspects

The Data Centre does not manage any data with disclosure risk. The Data Centre stores personal data about the users as they need to register. Personal data management follows GDPR.

The data providers are well aware of their responsibility for the correctness of data and metadata.

In general, everyone in IAGOS should work in a socially ethical way keeping the integrity and FAIRness, and maintaining a high level of trust and respect among the people working in IAGOS and with the users and other stakeholders. One should always take into account that the mission of IAGOS is to provide effective access for a wide user community to its resources and services, in order to facilitate high-quality Earth system research, to increase the excellence in Earth system research, and to provide information and knowledge on developing sustainable solutions to societal challenges.

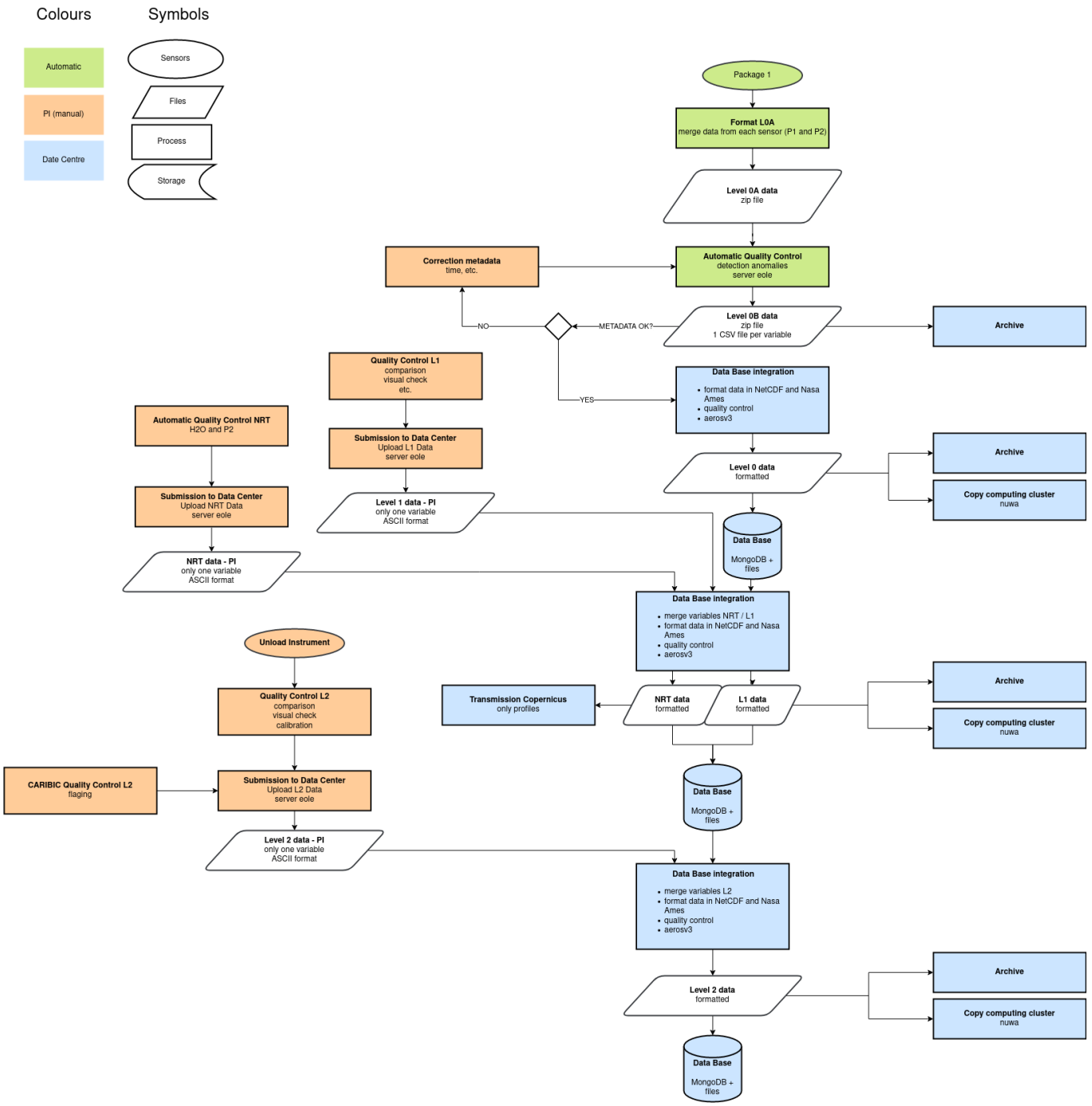
A. Appendix: Definition of Terms

- ACDD: Attribute Convention for Data Discovery, https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3
- AERIS: French Data and Services Cluster for the Atmosphere, https://en.aeris-data.fr/
- ATMO-ACCESS H2020 project: Sustainable Access to Atmospheric Research Facilities, https://www.atmo-access.eu/
- C3S: Copernicus Climate Change Service
- CAMS: Copernicus Atmosphere Monitoring Service, https://atmosphere.copernicus.eu/
- CARIBIC: Civil Aircraft for tde Regular Investigation of the atmosphere Based on an Instrument Container, https://www.caribic-atmospheric.com/
- CF: Climate and Forecast convention, https://cfconventions.org/
- CNRS: French National Centre for Scientific Research, http://www.cnrs.fr/en
- CSV: Comma-separated values
- Copernicus: European Union's Earth observation programme, https://www.copernicus.eu
- Data Terra: French Integrated Earth Observation, https://www.data-terra.org/
- DLR: Deutsches Zentrum für Luft- und Raumfahrt e.V.
- DMG: Data Management Group
- DMP: Data Management Plan
- EDUGAIN: EDUcation Global Authentication INfrastructure
- ENVRI-FAIR H2020 project: ENVironmental Research Infrastructures building Fair services Accessible for society, Innovation and Research, https://envri.eu/home-envri-fair/
- EOSC: European Open Science Cloud, https://eosc-portal.eu/
- ERI: European Research Infrastructure
- FAIR Principles: Findable, Accessible, Interoperable, Resuable
- FZJ: Forschungszentrum Jülich GmbH
- ECMWF: European Centre for Medium-Range Weather Forecasts, https://www.ecmwf.int/
- FLEXPART: FLEXible PARTicle dispersion model, https://www.flexpart.eu/
- GAW: Global Atmosphere Watch,
- GAWSIS: GAW Station Information System, https://gawsis.meteoswiss.ch/GAWSIS
- GEOSS: Global Earth Observation System of Systems, https://www.earthobservations.org/geoss.php
- IAGOS: In-service Aircraft for a Global Observing System, https://www.iagos.org/
- IAGOS-AISBL: IAGOS International not for profit Association, https://www.iagos.org/organisation/
- IAGOS-CORE: only IAGOS core instrumentation and data, https://www.iagos.org/iagos-core-instruments/
- IAGOS-MOZAIC: instrumentation and data from former project MOZAIC
- IAGOS-CARIBIC: instrumentation and data from project CARIBIC
- IAGOS-DC: IAGOS Data Centre, http://www.iagos-data.fr/
- KIT: Karlsruhe Institute of Technology

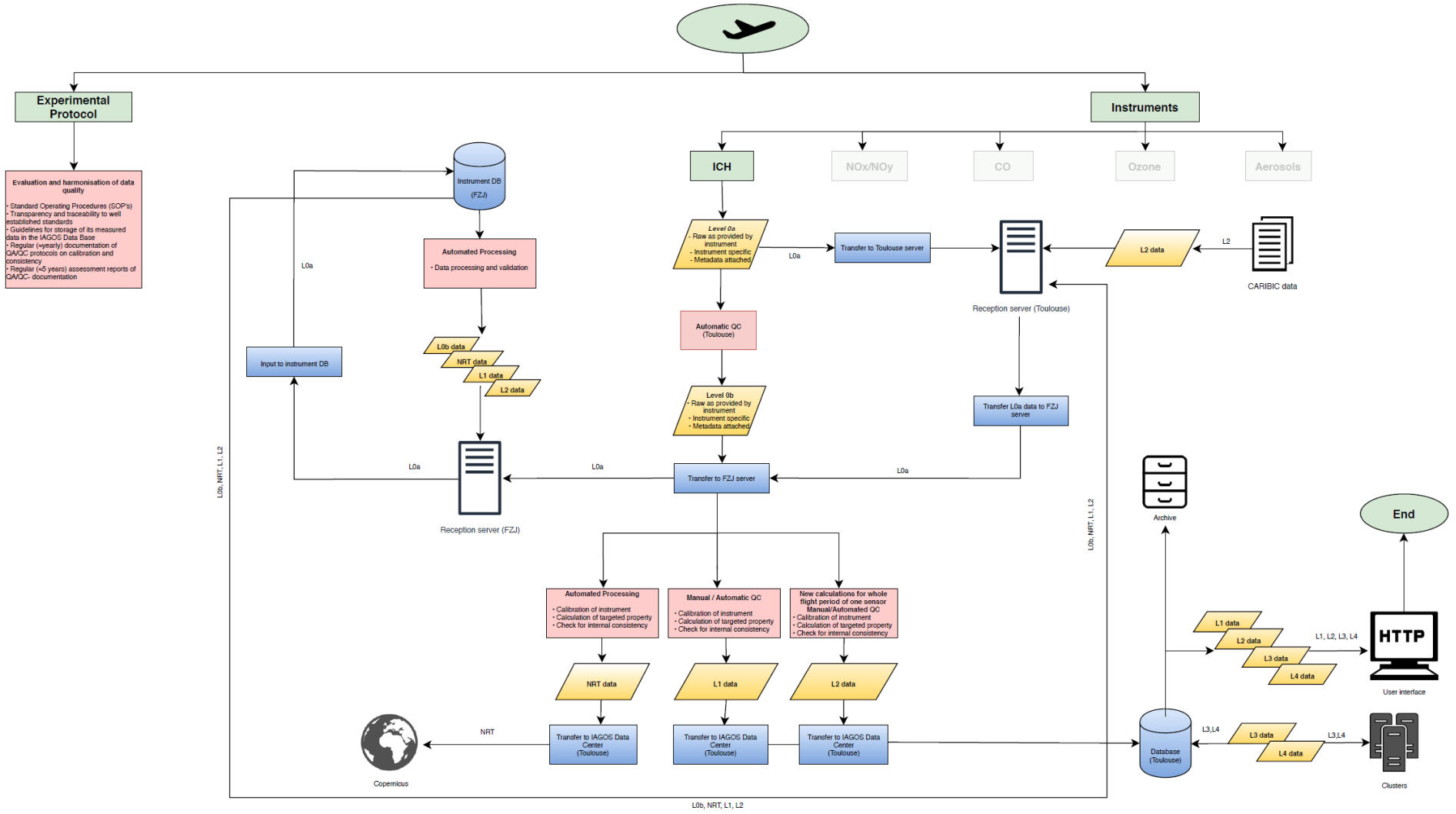
- MOZAIC: Measurements of OZone and water vapour by in-service AIrbus airCraft
- MPI-BGC: Max Planck Institute for Biogeochemistry
- NASA Ames: National Aeronautics and Space Administration Ames Format for Data Exchange
- NetCDF: Network Common Data Form
- NOXAR: Nitrogen Oxides and Ozone along Air Routes, http://www.megdb.ethz.ch/cmp_noxar_data.php
- NRT : Near Real Time
- PI: Principal Investigator
- PID: Persistent IDentifier
- RI-Urbans: Reinforcing Air Quality Monitoring Capacities in European Urban & Industrial AreaS
- RRT: Real Real Time
- OAIS: Open Archival Information System
- OMP : Observatoire Midi-Pyrénées, https://www.omp.eu/
- ORCID: Open Researcher and Contributor ID
- QA/QC: Quality assurance (QA) and quality control (QC)
- SEDOO: OMP Data Service: https://www.sedoo.fr/
- SOP: Standard Operating Procedure
- TROPOS: Leibniz-Institut für Troposphärenforschung e.V.
- WMO: World Meteorological Organization, https://public.wmo.int/en

B. Appendix: Workflows

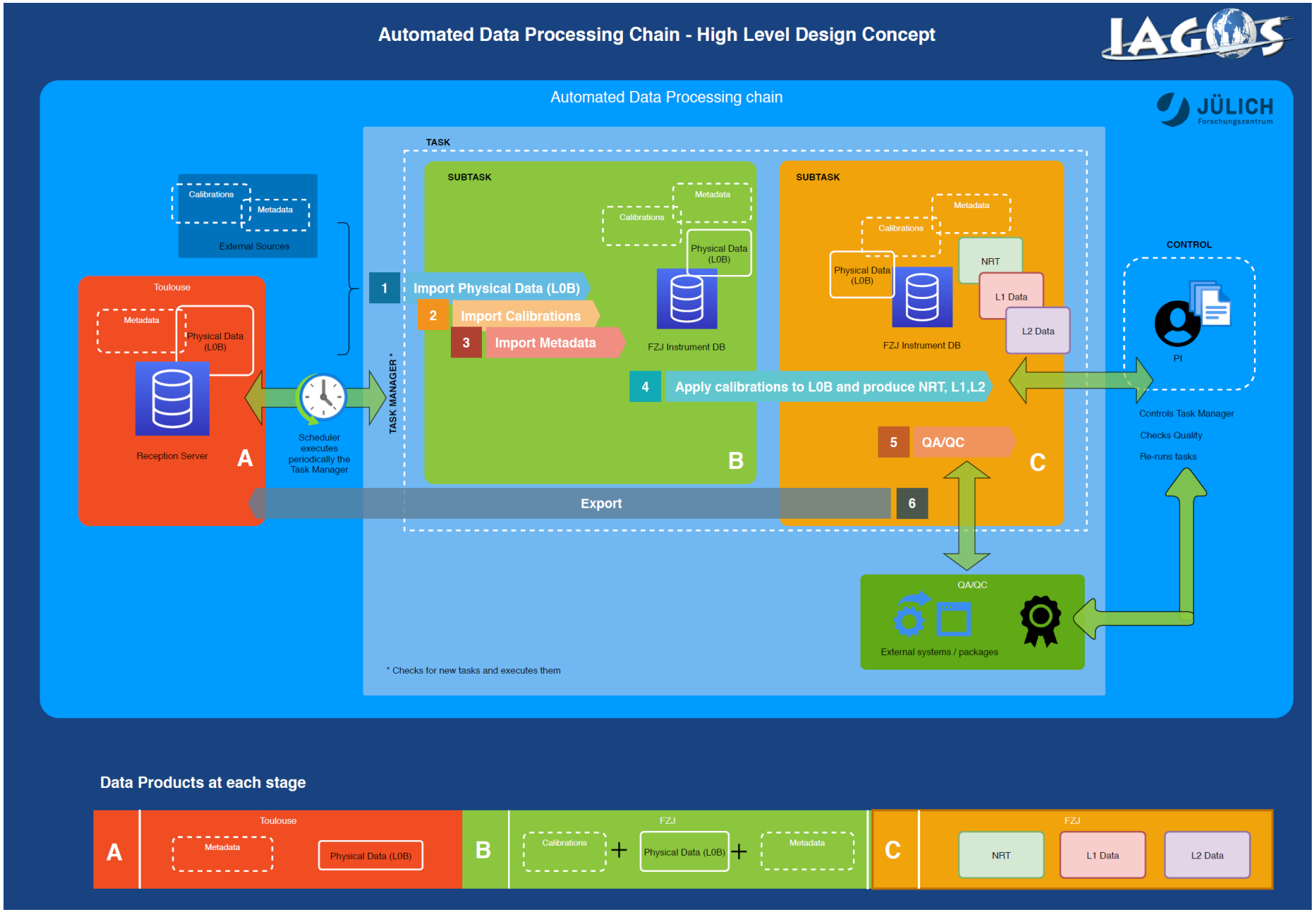
B.1 - IAGOS Data Centre data and metadata workflow



B.2 - FZJ Data and metadata workflow



B.3 - FZJ concept of the Data Processing Chain



C. Appendix - IAGOS variables

- P1: Package 1
- PM: Package MOZAIC
- PC: Package CARIBIC
- P2: Package 2

IAGOS variable	Long name	CF standard name	Unit	Processing levels	Availability	Source	Data provider	Access
air_press	Air pressure	air_pressure	Pa	L0, L1, L2	since August 1994	aircraft	CNRS	public
air_press_left	Left air pressure	air_pressure	mBar	L0, L1, L2	since November 2017	aircraft	CNRS	iagos team
air_press_right	Right air pressure	air_pressure	mBar	L0, L1, L2	since November 2017	aircraft	CNRS	iagos team
air_press_total	Total air pressure	air_pressure	mBar	L0, L1, L2	since May 2005	aircraft	CNRS	iagos team
air_speed	Aircraft air speed	platform_speed_wrt_air	m s-1	L0, L1, L2	since August 1994	aircraft	CNRS	public
air_stag_temp	Stagnation air temperature	stagnation_air_temperature	K	L0, L1, L2	since August 1994	aircraft	CNRS	public
air_temp	Air temperature	air_temperature	K	L0, L1, L2	since August 1994	aircraft	CNRS	public
altitude	Altitude rate	platform_altitude_rate	m s-1	L0, L1, L2	since May 2005	aircraft	CNRS	public
attack_angle	Angle of attack	platform_attack_angle	degree	L0, L1, L2	since May 2005	aircraft	CNRS	iagos team

baro_alt	Barometric altitude	barometric_altitude	m	L0, L1, L2	since August 1994	aircraft	CNRS	public
drift	Drift	platform_drift	degree	L0, L1, L2	since November 2017	aircraft	CNRS	iagos team
gps_alt	GPS altitude	gps_altitude	m	L0, L1, L2	since August 1994	aircraft	CNRS	iagos team
ground_speed	Aircraft ground speed	platform_speed_wrt_ground	m s-1	L0, L1, L2	since August 1994	aircraft	CNRS	public
lat	Latitude	latitude	degree_north	L0, L1, L2	since August 1994	aircraft	CNRS	public
lon	Longitude	longitude	degree_east	L0, L1, L2	since August 1994	aircraft	CNRS	public
mer_wind	Meridional wind	northward_wind	m s-1	L0, L1, L2	since August 1994	aircraft	CNRS	public
orientation	Orientation	platform_orientation	degree	L0, L1, L2	since May 2005	aircraft	CNRS	iagos team
pitch_angle	Pitch angle	platform_pitch	degree	L0, L1, L2	since May 2005	aircraft	CNRS	iagos team
radio_alt	Radio altitude	altitude_wrt_radar		L0, L1, L2	since August 1994	aircraft	CNRS	public
roll_angle	Roll angle	platform_roll	degree	L0, L1, L2	since May 2005	aircraft	CNRS	iagos team
vertical_speed	Vertical speed	platform_vertical_speed	m s-1	L0, L1, L2	since May 2005	aircraft	CNRS	iagos team
wind_dir	Wind direction	wind_from_direction	degree	L0, L1, L2	since August 1994	aircraft	CNRS	public

wind_speed	Wind speed	wind_speed	m s-1	L0, L1, L2	since August 1994	aircraft	CNRS	public
zon_wind	Zonal wind	eastward_wind	m s-1	L0, L1, L2	since August 1994	aircraft	CNRS	public
CO	Carbon monoxide mixing ratio	mole_fraction_of_carbon_monoxide_in_air	ppb	L0, L1, L2, NRT	since December 2001	P1, PM, PC	CNRS	public
H2O_gas	Water vapor volume mixing ratio	mole_fraction_of_water_vapor_in_air	ppm	L0, L1, L2, NRT	since August 1994	P1, PM	FJZ	public
NO	Nitrogen monoxide mixing ratio	mole_fraction_of_nitrogen_monoxide_in_air	ppb	L0, L1, L2	since August 1994	P1	FJZ	public
NOx	NOx expressed as nitrogen mixing ratio	mole_fraction_of_nox_expressed_as_nitrogen_in_air	ppb	L0, L1, L2	since August 1994	P1	FJZ	public
NOy	Total reactive nitrogen mixing ratio	mole_fraction_of_total_reactive_nitrogen_in_air	ppb	L0, L1, L2	since August 1994	P1	FJZ	public
O3	Ozone mixing ratio	mole_fraction_of_ozone_in_air	ppb	L0, L1, L2, NRT	since August 1994	P1, PM, PC	CNRS	public
RHL	Relative humidity (Liquid Water)	relative_humidity	1	L0, L1, L2, NRT	since August 1994	P1, PM	FJZ	public
air_rec_temp	Recovery temperature at RH-sensor	recovery temperature_at_RH-sensor	K	L0, L1, L2, NRT	since August 1994	P1, PM	FJZ	public
air_stag_temp	Stagnation air temperature	stagnation_air_temperature	K	L0, L1, L2, NRT	since August 1994	P1, PM	FJZ	public
air_temp	Air temperature	air_temperature	K	L0, L1, L2, NRT	since August 1994	P1, PM	FJZ	public
cloud	Cloud particles total concentration	number_concentration_of_cloud_liquid_water_particles_in_air	no cm-3	L0, L1, L2	since July 2011	P1	U. Manchester	public

cloud_presence	Cloud presence	number_concentration_of_cloud_liquid_water_particles_in_air status_flag	1	L0, L1, L2	since July 2011	P1	U. Manchester	public
CH4	Methane mixing ratio	mole_fraction_of_methane_in_air	ppb	L0, L1, L2		P2, PC	MPI-BGC	public
CO2	Carbon dioxide mixing ratio	mole_fraction_of_carbon_dioxide_in_air	ppm	L0, L1, L2		P2, PC	MPI-BGC	public
cloud_water	Cloud liquid water mass concentration	mass_concentration_of_cloud_liquid_water_in_air	kg m-3	L2		PC	CARIBIC	public