Preservation Policy

IAGOS Data Centre

PUBLIC

16 June 2021

Version: 1.0

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

Contents

Purpose

This document defines the IAGOS Data Centre (IAGOS-DC) policy on preservation, including for the designated community of users. This policy generally conforms to the OAIS Reference Model, with additions and alterations that are specific to the data managed by the IAGOS-DC.

The IAGOS-DC mission is to ensure long-term access to the data produced by the IAGOS Research Infrastructure (https://www.iagos.org), supporting and promoting their use, value and impact. To achieve this for current and future data providers and other beneficiaries, the IAGOS-DC offers long-term preservation services for the digital objects in its collection.

To ensure the continued use of its data products the IAGOS-DC follows a policy of preservation with the aim of ensuring the authenticity, reliability and logical integrity of all resources entrusted to it for long-term digital preservation while providing formats suitable for research, teaching or learning.

This policy codifies long-standing good archival practice at the IAGOS-DC which has been the unique repository for the IAGOS data since 2005. The responsibilities for long-term preservation of assets may change during the lifetime of this service. The IAGOS-DC also aims continually to improve all aspects of the preservation-related workflow by embedding an awareness of quality in all processes.

The basic principles of the preservation plan are outlined in Part I. This is the Preservation Strategy, being the general framework in which the IAGOS-DC operates.

The preservation plan also contains the Preservation Policy as referred to in CoreTrustSeal requirement R9 "Documented storage procedures", to be found in Part II. The Preservation Policy is the implementation of the principles outlined in the Preservation Strategy. It details the commitment to support the long-term management of data and also outlines the roles and responsibilities of all those involved in the collection and management of data. The goal of this OAIS function is to ensure that the data in the data archive remains accessible, understandable, and sufficiently usable over the long term.

Part I - Preservation Plan

1. Scope of the plan

The scope of this plan is limited to the IAGOS-DC. It deals with all aspects of preservation and applies to all the data products managed by the IAGOS-DC. This plan does not consider preservation of other materials such as web pages and internal documents.

The preservation plan follows a variety of external guidelines and standards for digital preservation such as OAIS, the FAIR principles and CoreTrustSeal.

2. Objectives of the plan

The IAGOS-DC mission is to ensure long-term access to the data produced by the IAGOS Research Infrastructure. The IAGOS-DC assumes responsibility for the long-term preservation and accessibility of digital objects. Most research data deserve to be kept available in the long term in order to be accessible for new research. In addition to possible reuse, data should also be kept available for replication and verification purposes.

The IAGOS-DC is also responsible for ensuring the authenticity and integrity of the data. Any strategy for the long-term preservation of any digital information must address the issue of software dependence. For most digital information and in particular research data it is generally possible to eliminate software dependency. Thus the primary goal of the IAGOS-DC's preservation plan is to ensure the long-term accessibility of electronic information while ensuring the highest level of authenticity possible.

The specific aims of the preservation plan are to:

- provide authentic and reliable instances of datasets to researchers
- maintain the integrity and quality of the datasets
- ensure that digital resources are managed throughout their lifecycle in the medium that is most appropriate for the task they perform
- ensure that the relevant level of information security is applied to each dataset
- be a "trustworthy digital repository"

3. Mission of the IAGOS Data Centre, content and designated community

3.1 Mission

The IAGOS-DC’s mission is to provide long-term access to the data produced by the IAGOS Research Infrastructure. To ensure the continued use of these resources the IAGOS-DC follows a policy of active preservation with the aim of ensuring the authenticity, reliability and logical

integrity of all resources entrusted to its care while providing formats suitable for research for the long term. The designated community of the IAGOS-DC mainly consists of scholars in Atmosphere and Environmental Sciences.

3.2 Characterisation of the content

The content of the IAGOS-DC consists of digital research data from the Atmosphere sciences. The number of published datasets is more than 200000 sorted in four data collections. The data collections are homogeneous in terms of data types, file format, size and usage. The data are generated for different purposes and through different processes. Not only the original data must be kept, but also the context that permits the data to be interpreted. The IAGOS-DC holdings consist of observational data (composition of atmospheric chemistry) and elaborated products that can be used to improve the comprehension of the observations.

3.3 Community watch and dedicated services

The designated community of the IAGOS-DC consists predominantly of scholars in the Atmospheric and Environmental sciences including:

- Scientific partners which are part of IAGOS and are in charge of the instrumentation, the acquisition and qualification of the data. The Data Centre develops and maintains services dedicated to these data providers.
- The international scientific community that is mainly focused on the study of Atmospheric composition and Climate.

The IAGOS-DC monitors the community through substantial contacts, for instance during data acquisition and ingestion, membership of the IAGOS European Research Infrastructure and by offering discipline specific services. The current section describes a selection of activities and services for specific communities, whereas section 10.3 will provide an overview of generic processes for monitoring and improving the quality of the IAGOS-DC.

IAGOS is part of the European H2020 projects ENVRI-FAIR and ATMO-ACCESS aiming to implement interoperability within the European environmental domain. By the end of ENVRI-FAIR in June 2023, the IAGOS users community should be extended to other environmental domains such as Marine, Solid Earth, and Biodiversity. These projects, with their link with the other environmental Research Infrastructures, enable IAGOS to stay in close contact with their communities, in particular on the data and metadata formats they prescribe.

4. Requirements

4.1 The IAGOS-DC's requirements

The IAGOS-DC has developed a series of requirements which it strives to ensure as closely as possible:

- the data that the IAGOS-DC acquires are accompanied by adequate documentation to enable their use and re-use for analytical and research purposes

- the datasets are checked and validated according to strict data ingest procedures (see section 9.2)
- the data are professionally catalogued according to appropriate metadata standards
- the data, documentation, metadata and other representation information are preserved for the long-term
- the authenticity, integrity and reliability of datasets preserved for future use are retained

4.2 Legal and regulatory framework

The legal and regulatory frameworks for the management of the data acquired by the IAGOS-DC are as follows.

IAGOS is led by the IAGOS-AISBL (international non-profit association under Belgian law) and its members are public institutional partners from European countries (Germany, France and UK). CNRS (French Public Research Organization), member of the IAGOS-AISBL, 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 relationship between the data provider of a dataset and the IAGOS-DC is based on a legally-binding deposit agreement and licence which:

- confirms the rights and obligations of both parties
- states that the IAGOS-DC shall ensure, to the best of its ability and resources, that the deposited dataset is archived in a sustainable manner and remains legible and accessible
- states that the IAGOS-DC has the right to modify the format and/or functionality of the dataset if this is necessary in order to facilitate the digital sustainability, distribution or re-use of the dataset
- states the conditions under which access may be given to third parties, as specified by the depositor (the IAGOS-DC applies the principle 'Open if possible, protected if necessary')
- states that the data provider has cleared all necessary permissions

Data providers are asked to clear, possibly unresolved, rights of third parties on (parts of) the datasets beforehand. The IAGOS-DC is not liable for the contents of the datasets made available through the IAGOS Data Portal, nor for the documentation associated with those datasets. The IAGOS-DC is also not liable for content errors or incorrect inferences from the datasets and the data contained therein.
The relationship between the user of a dataset and the IAGOS-DC is based on legally-binding General Conditions of Use which concern:

- the use of the data

- the special restrictions that apply to datasets with personal data according to the General Data Protection Regulation
- the required bibliographic reference to the dataset
- no content liability by the IAGOS-DC

5. Roles and responsibilities

All IAGOS-DC staff assist in implementing the policy contained in this preservation plan as appropriate to their roles and responsibilities. The manager of the IAGOS-DC is responsible for maintaining this policy. All IAGOS-DC staff, including temporary staff and trainees are accountable for keeping up confidentiality when processing data, in particular personal data, in any way whatsoever.

6. Content coverage

Formats of the files accepted by the Data Centre are agreed separately with each data provider. NetCDF and ASCII formats are accepted and they have to follow common recommendations (names of the variables, units, quality flags, mandatory metadata). Files which are not following the recommendations and the expected format are refused by the Data Centre and need to be submitted again after correction.

All data and metadata are converted in pivot formats and then stored in the database. Metadata is converted from the pivot standard to standard format (ISO-19115, Data Cite, etc.). Data files are provided to the users in NetCDF and ascii format NASA Ames.

All formats of digital files stand the risk of becoming obsolete in the future. NetCDF format has been selected because it has a high chance of remaining usable in the far future due its spread use in the community.

The data files that the IAGOS-DC intends to acquire, preserve and make available are of a static nature, i.e. they are not work "in progress" anymore. When data are changed or extended, the resulting updates are considered as new datasets.
Until now the IAGOS-DC does not acquire or preserve all the software programs that have been used to generate the data, although the data providers are encouraged to deposit documentation of the applied software (such as version and configuration parameters) in conjunction with the data and if possible a link to the software code on a public repository through a DOI.

7. Integrity and security

7.1 Integrity

The complete chain of the IAGOS-DC's custody of all datasets is documented through metadata. All actions are explicit, complete, correct and current. However, only the “original"

version can be said to be an integral copy of the version deposited at the IAGOS-DC. The preservation and dissemination versions are considered to be authentic and there is provenance information of all alterations in the preservation and dissemination versions that relates back to the original deposited version.
Provenance information is currently stored in the database but not provided in a standard way to the users. It's a work in progress done in the framework of the project ENVRI-FAIR.

7.2 Security

The preservation of the IAGOS-DC's collections relies on an IT infrastructure that is fit for purpose and is continually monitored and periodically reviewed to ensure timely upgrades in both hardware and software.
Adequate storage capacity for all holdings is maintained. The IAGOS-DC provides necessary secure networking and communications equipment, providing adequate connectivity.
All servers in the IAGOS-DC are protected by power surge protection systems. Disaster recovery procedures are in place. See section 9.3.1 for more details.

The IAGOS-DC is committed to taking all necessary precautions to ensure the physical safety and security of all data collections that it preserves:

- fire prevention and protection system
- physical intruder prevention and detection systems
- environmental control systems

The server rooms are equipped with key entries. All machine room computer systems are locked to prevent unauthorised access in the case of a security breach of the room.

8. Sustainability plans and funding

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

AERIS is committed to supporting continued funding for all the operations relating to preservation management. Resource management for preservation of digital resources includes:

- technical infrastructure, including equipment purchases, maintenance and upgrades, software and hardware obsolescence monitoring, network connectivity, etc.
- financial plan, including strategy and methods for financing the digital preservation programmes and commitment to long-term funding
- staffing infrastructure, including recruitment, induction, and ongoing staff training. The IAGOS-DC has established a rolling planning scheme for lifetimes of computer equipment and storage media to help forward planning for the necessary upgrades

The preservation of data and documentation to ensure they remain usable over time is a core activity of the IAGOS-DC. The IAGOS-DC, therefore, makes every effort to remain up to date with any relevant technological advances to ensure continued access to its data.

The sustainability plan is continuous preservation as long as the IAGOS-DC continues to exist and that its expertise is maintained in the long term. The Research Infrastructure IAGOS has a development plan, including data management, for the next five years.

In the case that the data would be lost, 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 IAGOS-DC 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 disparition 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.

Part II - Preservation Policy

9. Implementing the preservation strategy

The following chapters are structured around the main functional concepts of the Open Archival Information System (OAIS) reference model for digital preservation environments as well as the FAIR principles. OAIS reference model is an international standard which proposes common terms and concepts and a framework for entities and relationships between entities in digital preservation environments. The IAGOS-DC recognises the benefits of OAIS as a conceptual framework for defining and aligning archival activities and not a concrete implementation plan. The IAGOS-DC follows the broad guidance given in the OAIS reference model.

Preservation decisions at the IAGOS-DC are made within the context of the IAGOS-DC's mission and strategy, balancing the constraints of costs, scholarly value and user accessibility. The IAGOS-DC's processes are organised according to this model.

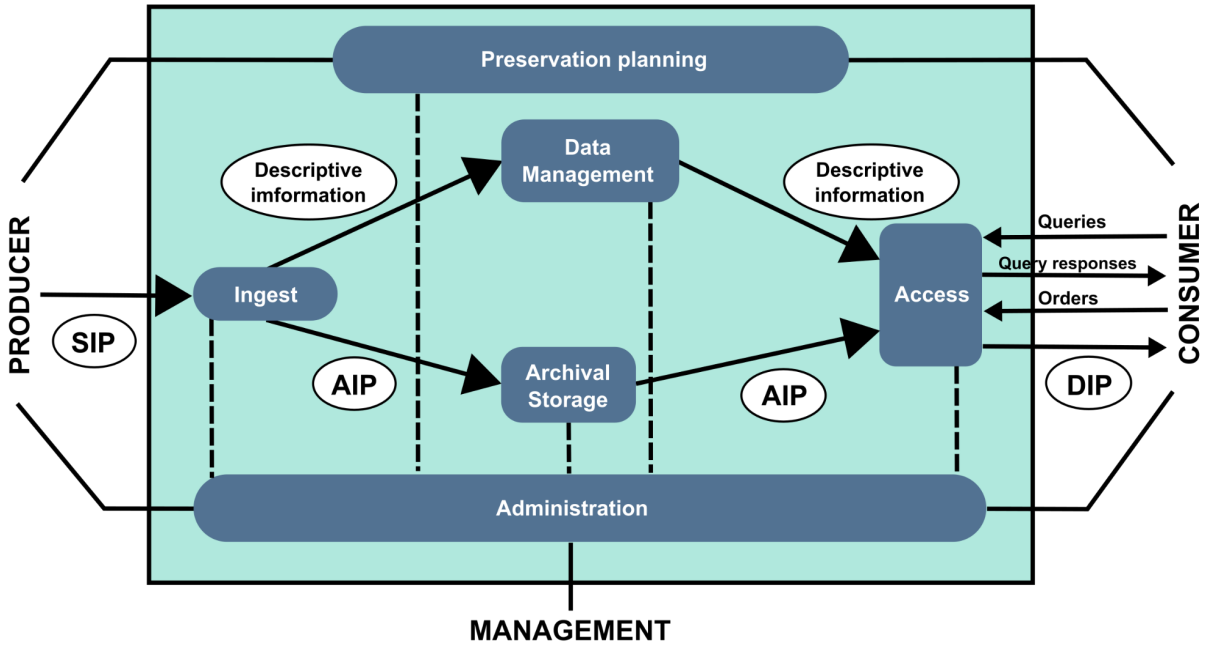

Figure 1: OAIS model

9.1 Pre-ingest function

Officially, the pre-ingest function is not part of the OAIS model. However, pre-ingest services help to ensure the usability and accessibility of datasets through the improved quality of metadata and documentation. This way, they also reduce costs within the ingest phase.
In particular, the IAGOS-DC provides guidelines documents and support to the IAGOS data providers.

9.2 Ingest function

Ingest is the first functional component of the OAIS reference model. It includes the receipt of information from a data provider and the validation that the information supplied is complete. This process also identifies the specific properties of the information which is to be preserved; it authenticates that the information is what it claims to be. The data provider is responsible for successfully documenting and depositing the data. However automated procedures implemented at the IAGOS-DC check if the data submitted by the data providers respect some requirements (data format, mandatory metadata, etc.). Files which are not following the recommendations and the expected format are refused by the Data Centre and need to be submitted again after correction.

The IAGOS-DC staff performs quality control by following a data processing protocol to safeguard that the supplied data will be findable, accessible and comprehensible for the long term. The version supplied by the data provider is known within the IAGOS-DC as the "original" version and this is retained for preservation in its original format. This supplied version has a close correspondence to the Submission Information Package (SIP) in OAIS terms.
Only the approved/curated data and metadata will be published. If the data and metadata required transformations, then the original data are kept, but not published. For some data collections, data files are merged in order to provide a single file to the final users. The data providers are aware of these changes. The IAGOS-DC may apply minor changes to the metadata and add additional metadata. Larger issues will be consulted with the data provider.
As a member of IAGOS, the data provider is informed that the material has been transferred to the IAGOS-DC's custody. When the IAGOS-DC staff has subsequently processed the dataset and it has been accepted, then the dataset is published and assigned with a unique persistent identifier minted by the IAGOS-DC's system.
The version resulting from the ingest process is an Archival Information Package (AIP). All actions related to the preservation of the data are documented in the IAGOS-DC provenance database (triplestore) following the PROV-O standard.
All datasets are ingested in a fully automated procedure by SFTP. The datasets can also be manually uploaded on the IAGOS Data Portal through dedicated tools for the data providers.

9.3 Archival storage function

In essence, the purpose of archival storage is to ensure that what is passed to it from the ingest process remains identical and accessible. In the IAGOS-DC this function receives AIPs from the ingest function and adds them to the permanent storage facility, oversees the management of this storage, including media refreshment and monitoring. This function is also responsible for ensuring that AIPs can be retrieved.

9.3.1 Physical data preservation and storage

In order to best safeguard long-term preservation, the IAGOS-DC follows a policy of multiple copy resilience. Three versions of the complete preservation system are held in Toulouse, France. A main near-line copy and a shadow copy on two separate preservation servers. There

is also a near-site online copy on a server located in another location at the University of Tarbes, France.

9.3.2 Media monitoring and refreshing strategy

The IAGOS-DC operates a media monitoring procedure. This allows it to check for potential future problems of wear and tear on media and act before the problems become severe. Bad blocks on discs are automatically identified, administrators are notified and the disc replaced.
If any media have either recoverable or non-recoverable errors then they are regenerated from the on-site mirror preservation server.

9.3.3 Compression

In order to reduce the risk of damage to data, the IAGOS-DC uses industry standard lossless compression tools. Sufficient redundancy is used for long-term storage to warrant use of these tools. Compression software technology is monitored continuously to ensure timely updates to enable successful migration.

9.4 Data management function

Data Management is the third major function of the OAIS reference model. It maintains databases of descriptive metadata; supports external finding aids; and manages administrative metadata which support internal operations, including change control.

9.4.1 Version control/change procedures

Ensuring that any alteration to the preserved version of any part of a dataset is accurately documented is integral to the authenticity of any dataset. The IAGOS-DC distinguishes between two forms of alteration post ingest:

- 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.

In the case of data conversion to another file format for preservation or access purposes, the IAGOS-DC maintains the original files. The conversion aims to preserve the content of the data, because this is seen as the most significant property of the data. Preservation of other aspects, such as the layout of the input format (the "look and feel") is considered to be of lesser importance for most research data.

9.4.2 Data collection withdrawal

Deleting data would be an extreme case of data change once published. In principle however, in principle the IAGOS-DC does not delete published data, unless sufficiently weighty grounds exist for removal.

9.5 Access function

This OAIS function contains the services and functions that make the data collections and related services visible to consumers. End users interact with the IAGOS-DC through the IAGOS Data Portal (http://www.iagos-data.fr) to find, request and receive datasets. By default these processes are web-based, but with support by the IAGOS-DC staff. Machine-actionable endpoints are under development and will soon be available to the users.
Apart from the processes that support these three activities (i.e. find, request and receive datasets), the access function also implements the security that is related to access.
To enhance findability, metadata is checked and completed where necessary. The map feature enables users to locate datasets on a map through their coordinates. Other search criteria are also maintained such as temporal extent and names of the variables.

9.6 Administration function

In the OAIS model the administration function manages the day-to-day operations of the IAGOS-DC. Processes covered here e.g. relate to the negotiation of the license agreement and the general conditions of use and to system engineering functions to monitor the IAGOS-DC's system operations (see section 9.3).

10. Preservation planning

10.1 Preservation strategy overview

The IAGOS-DC has chosen to implement a preservation strategy based upon open and available file formats, data migration and media refreshment. Different ingest processes may be required for material with different levels of quality and significance. Specifications for ingest standards including validation and integrity checks are documented.
The IAGOS-DC's preservation strategy is predicated on two basic principles: first, that digital storage media are inherently untrustworthy unless stored appropriately; second, that all file formats and physical storage media will ultimately become obsolete.

Therefore, the environmental parameters which control the storage media are tightly controlled to reduce the vulnerability of these media. Additionally, the strategy to reduce the risk of obsolescence is based on storing multiple copies on different storage media.
A similar strategy is employed to deal with the obsolescence of file formats. Appropriate information-rich preservation formats have been identified and are used in conjunction with formal documentation procedures. These formats are chosen with specific reference to the 'data

types' under consideration. The IAGOS-DC follows international best practice in its choice of preservation formats and data migration procedures.

When new formats are created from data files either through migration into new file formats or through creating new file formats for dissemination, the old files are retained alongside.
The preservation strategies of the IAGOS-DC aim to maintain a flexible preservation system that evolves to meet the demands of changing technology and new and increasing user expectations.

Every data collection within the preservation system follows a consistent directory structure for storage, and this is enforced by automated checks. This has many benefits, such as the ability to find set types of information and also to allow automated tasks (e.g. migration of file formats) to be run without the need for complicated locator scripts. In addition to this structure, file label details are kept in an in-house system to provide extra information about a file in addition to its filename. Further, file extensions are always standardised, with a single extension allowable for each type of file.

10.2 Preservation principles: FAIR

The IAGOS-DC aims to operate according to the FAIR Guiding Principles in making its data Findable, Accessible, Interoperable and Reusable. Below is described, per principle, how the data comply. The full implementation of the FAIR principles is currently in progress in the framework of ENVRI-FAIR.

To be Findable

- Each dataset deposited in the IAGOS-DC is automatically assigned a Persistent Identifier (PID) to facilitate discoverability and sustainable reference. Digital Object Identifiers (DOI) are assigned for citable data sets and ePIC PIDs (Persistent Identifiers for eResearch) for internal datasets.
- Metadata, compliant to ISO 19115-2 / INSPIRE, are quite rich with regard to machine-readability. They can be exported through the user interface in XML, CSV and JSON formats so as to be discoverable by both humans and computers.
- All the IAGOS metadata records can be harvested through its OAI-PMH or CSW service, which also makes datasets more visible to providers of search and discovery services.

To be Accessible

- Metadata are always openly accessible. No authentication or authorisation is ever needed either through the user interface or by using the open and universal OAI-PMH/CSW protocol.
- To access datasets through the IAGOS Data Portal, new users are requested to create a user account through registration (authentication). The activation of the user account is moderated by IAGOS principal investigators (authorisation).

To be Interoperable

- Metadata are mapped to ISO 19115-2 standard, a formal, accessible, shared and broadly applicable language for knowledge representation. Also, to comply with the needs of other partners, the IAGOS-DC maps and exposes the metadata to other metadata schemas such as DataCite and is currently implementing mapping to WIGOS, WIS and DCAT schemas.
- To facilitate interoperability controlled vocabularies are used for the name of the variables, the sensors and observation facilities:
    - Climate and Forecast (CF) standard names : https://cfconventions.org/standard-names.html
    - Global Change Master Directory (GCMD) Keywords: https://earthdata.nasa.gov/earth-observation-data/find-data/idn/gcmd-keywords

To be Reusable

- Data providers are encouraged to describe their data with as much detail as possible. Metadata must contain mandatory elements: Title, Creator, Description, Date (created). The required elements richly describe the data's provenance information; what are the data about, by whom and when they were created as well as how the data were created and for which purposes.
- To enable proper reuse data are always released with clear conditions of use.
- Having recognised the need to lower the barriers and facilitate data reuse by different communities, the IAGOS-DC implements some measure of interoperability among different metadata standards which are exposed through its CSW and OAI-PMH service. The IAGOS-DC is committed to use both domain-specific and domain-neutral metadata standards when necessary.
- Once a dataset is published only the IAGOS-DC can make changes, keeping control of the authenticity.
- Fixity checking is in place, verifying that data have not been altered or corrupted.

10.3 Monitoring, review and feedback

So far no procedure exists for monitoring and reviewing the preservation policy.

A. Appendix: Definition of Terms

- 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/
- CNRS: French National Centre for Scientific Research, http://www.cnrs.fr/en
- ENVRI-FAIR H2020 project: ENVironmental Research Infrastructures building Fair services Accessible for society, Innovation and Research, https://envri.eu/home-envri-fair/
- IAGOS: In-service Aircraft for a Global Observing System, https://www.iagos.org/
- IAGOS-DC: IAGOS Data Centre, http://www.iagos-data.fr/
- OAIS: Open Archival Information System
- OMP : Observatoire Midi-Pyrénées, https://www.omp.eu/

B. Appendix: References

- IAGOS Data Management Plan: https://services.iagos-data.fr/prod/v2.0/documents/60df1fa584045e26a7a0778a
- CSW metadata access endpoint: http://catalogue2.sedoo.fr/geonetwork/srv/eng/csw-iagos?service=CSW&version=2.0.2&request=GetCapabilities
- OAI-PMH metadata access endpoint: http://catalogue2.sedoo.fr/geonetwork/srv/eng/oaipmh?verb=ListRecords&metadataPrefix=oai_dc
- REST data access endpoint: https://services.iagos-data.fr/prod/swagger-ui.html
- CoreTrustSeal: https://www.coretrustseal.org
- Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18