



Project Title Prototype of HPC/Data Infrastructure for On-demand Services

Project Acronym PHIDIAS

Grant Agreement No. INEA/CEF/ICT/A2018/1810854

Start Date of Project 01.09.2019

Duration of Project 36 Months

Project Website www.phidias-hpc.eu

Deliverable 6.1.1 – Specifications for long-term data archiving procedures in respects of RDA recommendations

Work Package	WP 6 : Ocean use case
Lead Author (Org)	Gilbert Maudire (Ifremer)
Contributing Author(s) (Org)	Olivier Rouchon (CINES)
Due Date	29.02.2020
Date	18.03.2020
Version	V1.6

Dissemination Level

- PU: Public
- PP: Restricted to other programme participants (including the Commission)
- RE: Restricted to a group specified by the consortium (including the Commission)
- CO: Confidential, only for members of the consortium (including the Commission)



The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Versioning and contribution history

Version	Date	Author	Notes
0.0	04.03.2020	Gilbert Maudire (Ifremer)	Initial draft version
1.5	11.03.2020	Corrections after review by Olivier Rouchon (Cines) and Joel Sudre (CNRS)	Deliverable version (with revision marks)
1.6	13.03.2020	Charles Troupin (ULiege)	Minor corrections Final version

Disclaimer

This document contains information which is proprietary to the PHIDIAS Consortium. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to a third party, in whole or parts, except with the prior consent of the PHIDIAS Consortium.

Table of Contents

Executive Summary.....	5
1 Objectives of the WP6 – Ocean use cases.....	6
2 Managed data.....	9
2.1 In-situ Data.....	9
2.2 Remote Sensing Data.....	10
2.3 Output Data.....	11
3 Existing publication/Archiving services.....	12
3.1 Permanent identifiers.....	12
3.2 Metadata.....	13
3.3 Published/Archived data.....	14
3.4 Preservation of evolving datasets.....	15
3.5 Preparation for long term preservation.....	15
4 New requirements.....	17
1.1 Service scalability.....	17
1.2 Back office exchanges.....	18
1.3 Securing long-time archive.....	18

TERMINOLOGY

Terminology/Acronym	Description
CSW	Catalogue Service for the Web (ISO/OGC Protocol)
DIAS	Data Information and Access Service
DIVA	Data Interpolating Variational Analysis
DIVAnd	DIVA in n dimensions (new version of DIVA)
EMODnet	European Marine Observation and Data Network
EOSC	European Open Science Cloud
EUDAT	Collaborative Data Infrastructure in Europe
HPC	High Performance Computing
HPDA	High Performance Data Analytics
OGC	Open Geospatial Consortium
OPeNDAP or OpenDAP	Open-source Project for a Network Data Access Protocol
SeaDataNet	European Research Infrastructure for Marine Data Management
WCS	Web Coverage Service (OGC Protocol)
CSV	Coma Separated Values, Flat Text File Format
CF-convention	Climate and Forecast convention
NetCDF	Network Common Data Form
BGC	Bio-Geo-Chemical variables
DOI	Data Object Identifiers
IFCB	Imaging FlowCytoBot
API	Application Programming Interface
ORCID	Open Researcher and Contributor Identifier
iRODS	Open Source Data Management Software (irods.org)

Executive Summary

This document provides the specifications and new requirements for long term archiving procedures within the scope of the “Ocean use case” defined and developed in WP6 of the PHIDIAS project. This version has been reviewed by Olivier Rouchon (CINES), Joël Sudre (CNRS) and Charles Troupin (ULiege). It is the final deliverable version.

It presents the use case context before detailing the technical requirements in order to present the main components to readers.



1 Objectives of the WP6 – Ocean use cases

The general objective of the WP6 - Ocean use case is to improve the use of cloud services for marine data management, data service to users in a FAIR perspective, data processing on demand, taking into account the European Open Science Cloud (EOSC) challenge and the Copernicus Data and Information Access Services (DIAS).

Since the marine environment is evolving continuously, and because marine observation is still expensive, observation data are unique and must be well preserved and easy to be retrieved.

In practice, several functions (cf. Fig. 1) have been implemented either in France, by the Ocean – Odatis cluster of the French Earth Observation Research Infrastructure DataTerra and, in Europe, by the SeaDataNet Research Infrastructure, the European Marine Observation and Data Network for in-situ data and the Copernicus Marine Environment Monitoring Service for Operational Oceanography and Satellite Data:

- Distributed Data and Services Centres manage data that are acquired in routine modes. Those data centres have established a direct interface with observation systems: satellite missions, in-situ observatories, research fleets in relationship with National Oceanographic Data Centres.
- Data that are acquired more occasionally (“long tail data”) or data that require manual work at laboratories to be elaborated are often not directly interfaced with data centres. In order to facilitate the ingestion of these data in the marine data management infrastructure, **publication/archiving services** have been provided to scientific teams.

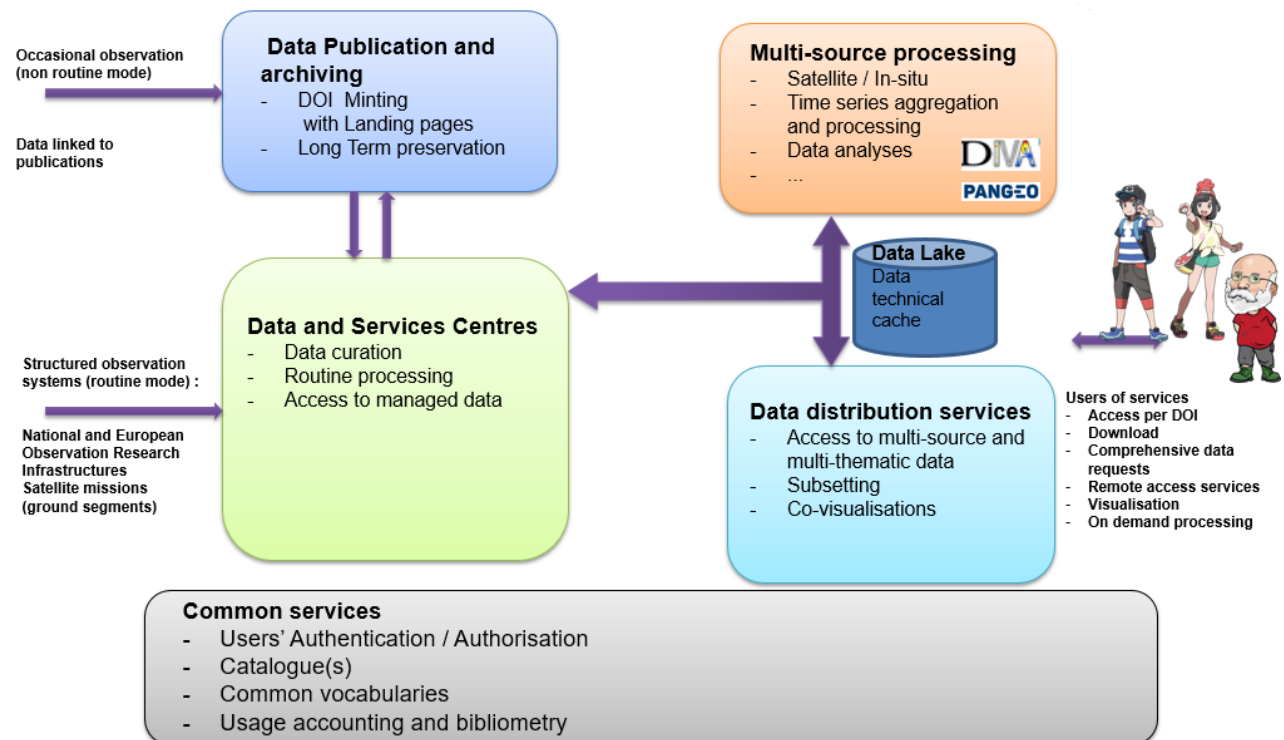


Figure 1: General architecture of services provided by French and European Marine Data Infrastructures

Most of these services are already implemented by the French Ocean Data Management Cluster (Odatis) and by European Infrastructures such as SeaDataNet and EMODnet. However, upgrades are necessary to facilitate and to speed-up response time of data access and data browsing, to enlarge data management capacities, and to improve the long-term stewardship of marine data, especially for the long tail in-situ observations.

In this context, the technical objectives (cf. Fig. 2) of the WP6 - Ocean Use Case are, by making an adapted use of HPC (high-performance computing) and HPDA (high-performance data analytics) capacities:

- Task 6.1: Improvement of long-term stewardship of marine in-situ data**
- Task 6.2: Improvement of data storage for services to users**
 - For two contexts :
 - 1- Fast and interoperable access for visualization and subsetting purposes (web portal)
 - 2- Parallel processing within dedicated high-performance computing
- Task 6.3: Marine data processing workflows for on-demand processing**

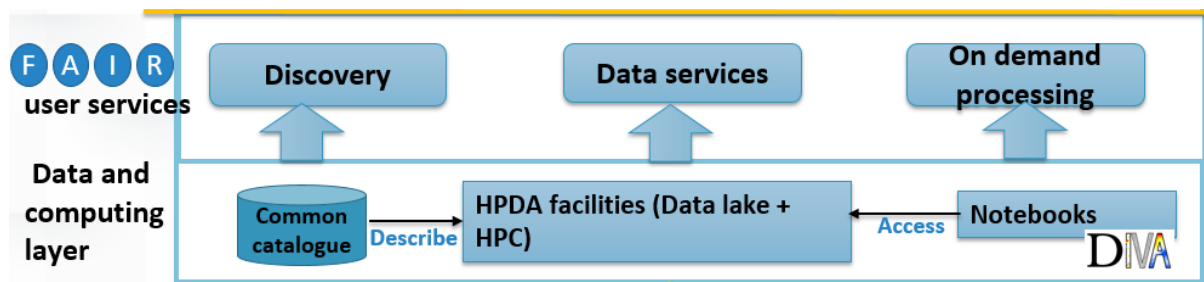


Figure 2: General sketch of the WP6 - Ocean use-case

In respect of those objectives, two main requirements have been identified about data storage:

1. Reinforce the long-term preservation capacities of observational data;
2. Manage a work copy of data with adapted structure in order to speed up and facilitate data retrieval and data processing. This working copy can be considered as a “**technical cache**” (called “**Data Lake**” in this document) that accelerates and harmonizes data requests and data processing.

The specifications of the requirement 2 – “Data Lake” are described in the deliverable 6.1.2 “Specifications of data storage”.

This document will focus on the requirement 1 – “Long-term preservation capacities”.

2 Managed data

Marine observations are made up of several data types, with different characteristics:

- Satellite data (large datasets, covering only the sea surface);
- In-situ observations (many small datasets, often with measurements below the sea surface).

And from distributed data repositories, such as:

- Operational oceanography observations that are, for example, stored in the DIAS Wekeo;
- Scientific observations that are, for example, stored using EUDAT services (SeaDataNet European infrastructure) or by the Data Centre themselves;
- Long tail data sources that are partly managed by online publication/archiving services.

Access conditions may differ from one data source to another one, for instance for some repositories, access is only granted to registered users (MarineID for SeaDataNet, Mercator Ocean International for Copernicus ...).

Most of those data are recorded using one of these formats: NetCDF (following the CF-Conventions), Ocean Data View spreadsheet (CSV compatible + semantic header). Other formats must also be considered such as Geographical Shape Files (SHP) and imagery formats (TIFF, Geo-TIFF, JPEG and MPEG for the animations).

This use case will target only two specific areas: the North Atlantic (North-East for the Chlorophyll-a) and the Baltic Sea. These 2 regions represent 10 millions of observations in the North Atlantic and the Baltic Sea, accounting for a total of approx. 250 GBytes. However, the progresses achieved during the Phidias project will be extended to other data sources and data types as far as possible.

Important notice: It is considered that the long term-preservation of satellite data is out of the scope of this use case because they are under the responsibility of spatial agencies. Consequently, the requirements of long-term preservation described in this document will target only in-situ data and products derived from both in-situ data and satellite data.

The following data sets will be considered at a first stage by the use case:

2.1 In-situ Data

- SeaDataNet and EMODnet-Chemistry marine in-situ data collections, managed at EUDAT partner CSC (Finland), especially:
 - o Temperature & Salinity



The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

- Chlorophyll-a concentration
- Access conditions at <https://www.seadatanet.org/Data-Access>
- Copernicus in-situ data collections, managed by DIAS-WEKEO, Ifremer and FMI, especially:
 - Temperature & Salinity
 - Chlorophyll (BGC Argo & FerryBox)
 - Access conditions at <http://www.marineinsitu.eu/access-data/>
- Euro-Argo data managed by Ifremer, especially:
 - Temperature & Salinity
 - Chlorophyll (BGC Argo)
 - Access services:
 - ftp servers
<ftp://ftp.ifremer.fr/ifremer/argo>
 - DOI (Data Object Identifiers)
<http://www.argodatamgt.org/Access-to-data/Argo-DOI-Digital-Object-Identifier>
 - synchronization service (rsync)
<http://www.argodatamgt.org/Access-to-data/Argo-GDAC-synchronization-service>
 - Thredds server API
<http://tds0.ifremer.fr/thredds/catalog/CORIOLIS-ARGO-GDAC-OBS/catalog.html>
- Imaging FlowCytobot (IFCB): in-situ automated submersible imaging flow cytometer at Utö field station
- Long-tail observations managed by EMODnet Ingestion and SeaNoe online publication/archiving service

2.2 Remote Sensing Data

- SMOS Sea Surface Salinity products, managed at Ifremer
 - De-biased 10-day average & monthly salinity field products from SMOS satellite (mixed orbits)
 - Access services :
 - ftp server: <ftp://ext-catds-cpdc:catds2010@ftp.ifremer.fr/>
 - DOI: 10.12770/0f02fc28-cb86-4c44-89f3-ee7df6177e7b
 - 20 MBytes per day
- Sentinel-3 imagery, managed at ESA-DIAS (variables to be defined)
 - Sentinel 3 (OLCI) service
 - Two datasets: 1) Full resolution with 300 m spatial resolution and 2) Reduced Resolution is approximately 1.2 km on ground. Resolution 20m
 - access conditions at <https://sentinel.esa.int/web/sentinel/user-guides/sentinel-3-olci>

2.3 Output Data

Products will be 4-D gridded fields (latitude, longitude, depth, time) or 3-D (fixed depth of fixed time) recording using netCDF according to the CF conventions (Climate Forecast, <http://cfconventions.org/>).

Examples may be found at: <https://www.seadatanet.org/Products#/search?from=1&to=30>.

See also figure 3 and 4 below.

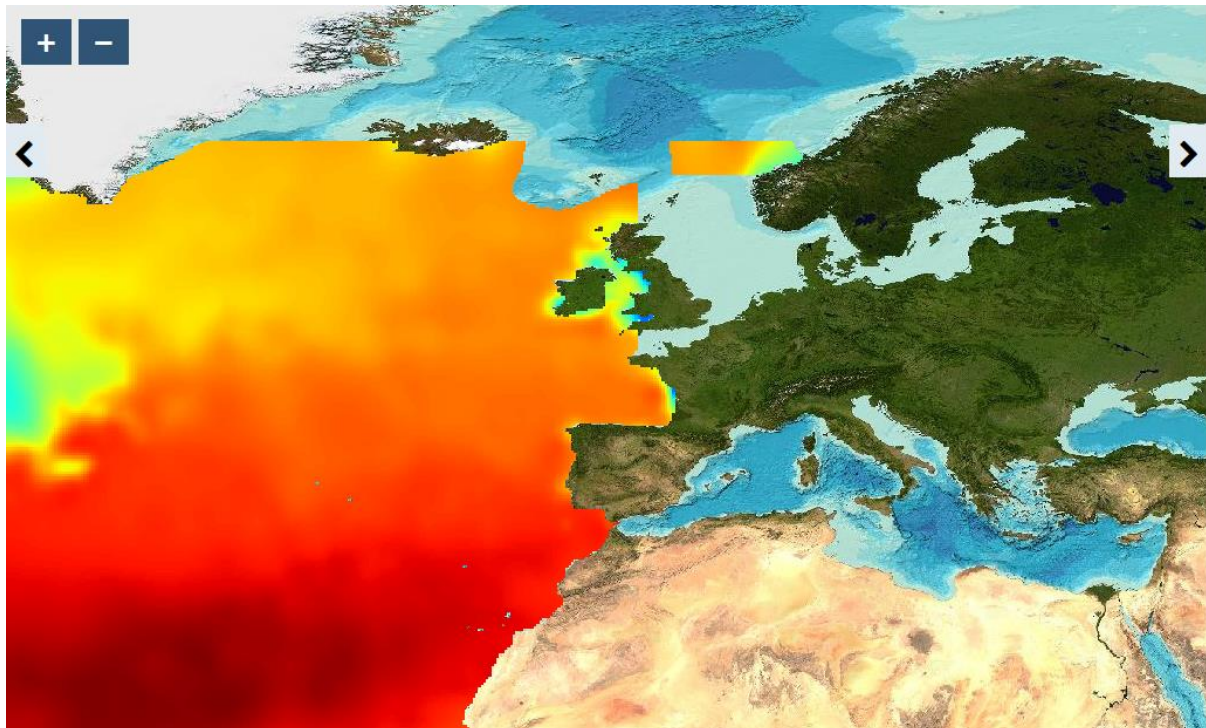


Figure 3: North-East Atlantic Mean Salinity at surface in January (Ocean Browser – ULiege)

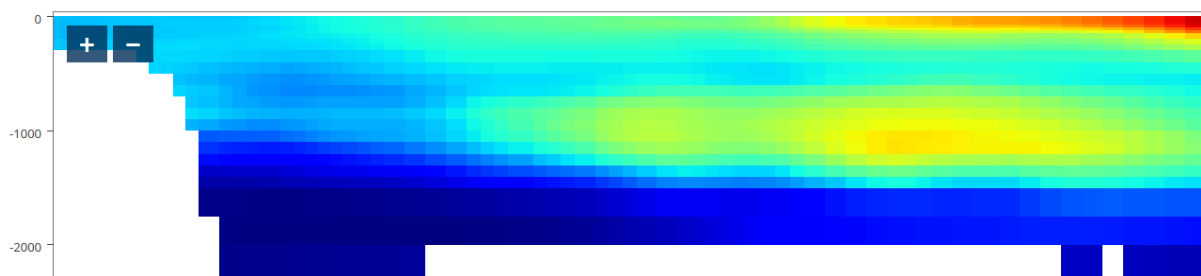


Figure 4: Vertical section of North-East Atlantic Mean Salinity at longitude 20°W, with Mediterranean water bodies (Ocean Browser – ULiege)

3 Existing publication/Archiving services

In France, the SeaNoe publication service (<https://www.seanoe.org>) has been set up for marine data in the framework of the Odatis cluster. In Europe, the EMODnet Ingestion service is also now available. The EMODnet ingestion service relies also on SeaNoe for marine research data.

Some other similar services also exist such as EUDAT services, Pangaea in Bremen-Germany (<https://www.pangaea.de/>), Zenodo (<https://zenodo.org/>), or the Dataverse software ([Dataverse.org](https://dataverse.org)). However, these services are more generic and less dedicated to marine and earth observation sciences. For instance, Zenodo allows users not only to upload dataset, but also software codes, journal articles or presentations.

These publication/archiving services are in line with the usual recommendations such as:

- DataCite Metadata Working Group. (2016). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. doi:10.5438/0012.
- DataTerra, Groupe interpôle. (2016) : Bonnes pratiques autour de la citation.
- Research Data Alliance working Group: "Addressing the Gaps: Recommendations for Supporting the Long Tail of Research Data",

These services address the following requirements:

- Minting of permanent identifiers (DOI – Digital Object Identifiers),
- Management of metadata
- Generation of landing pages
- Preparation of archiving

More information about SeaNoe services may be found below and at <https://www.seanoe.org/html/publish-your-data.htm>.

These services are now certified as part of a certified repository by the Core Trust Seal (RDA & ICSU-WDS).

3.1 Permanent identifiers

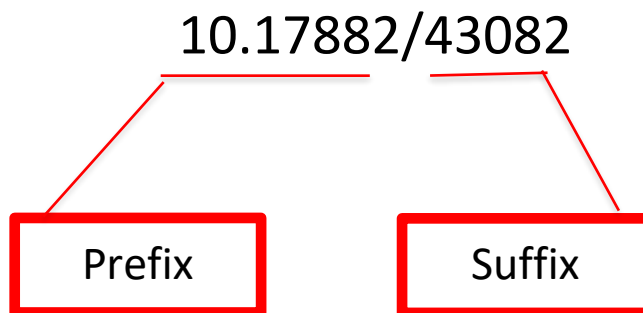
Permanent identifiers are minted in order to:

- Rationalize Data citation;
- Provide traceability on data usage.
- Simplify data access, especially for data sets that are linked to scientific publications.

The permanent identifiers in use are DOI – Digital Object Identifier, attributed by DataCite.

Minted DOI's are not meaningful for human readers (only digits), in order to make them as persistent as possible. For example, they do not include any organization name as the organization name may change.

The used DOI syntax is:



The prefix is fixed and attributed to an organization (e.g. seano.org). The suffix is the identifier of the dataset within the organization.

In some cases, especially for managing version of constantly evolving datasets (cf. § “Preservation of evolving datasets” below) such as time series, fragments may be used (# symbol): [10.17882/42182#42350](https://doi.org/10.17882/42182#42350).

3.2 Metadata

Managed metadata are compliant with the Dublin Core standards. If relevant, for example for geo-referenced data, metadata are made compatible with ISO 19115 standard (e.g. by the addition of geographical extend...). Main managed information are:

- General metadata (Dublin Core)
 - Title
 - Author(s) and affiliations (link with ORC ID)
 - Publication date
 - Abstract
 - References
 - Use Conditions (Possible limitations...)
 - Reference to data user’s manual (if any)
- Access conditions
 - Data Licence (Creative Commons license, ...)
 - Provided data citation according to a format suggested by DataCite: "Creator (Publication Year): Title. Publisher. Identifier"
 - Access service(s)
 - Data format and size
- Keywords (CodeList provided):
 - Variables (link with the Essential Ocean Variables Code List),
 - Method(s),
 - Instrument(s),
 - Project(s)
- Geographical extends

- Min and Max latitudes and longitudes
- Location map
- Temporal extends
 - Min and Max time
- Data preview(s)
 - Figures, maps, visualisation services
- List of associated datasets
- List of citing publication
- ...

These metadata can be imported and/or harvested using various protocols such as OGC-Catalogue Service for the Web and OAI-PMH or downloaded in the RIS format. This format allows automatic import in bibliographic management tools (e.g. Endnote).

Generating these metadata can be automated using tools. For example, the DIVA data interpolation tool is able to generate compatible metadata when generating interpolated fields (data products).

3.3 Published/Archived data

Using SeaNoe, data may be directly uploaded by data providers. Data centres perform checks in order to ensure that data match what has been described in metadata.

Checks include:

- Completeness of metadata in respect of data types;
- Data files are readable as described in metadata;
- Used vocabularies are the recommended vocabularies.

In case of detected non-conformance, additional information are requested from the data provider.

These checks may include format conformance tests. In the Phidias WP6-Ocean use case, only two file formats will be considered: netCDF using CF-convention, Ocean Data View Spreadsheet (CSV compatible, including mandatory metadata headers). Specific reader software tools have been developed for those data formats.

Checksums are computed (SHA256 hash function) in order to ensure that data will not be modified

3.4 Preservation of evolving datasets



The PHIDIAS project has received funding from the European Union's Connecting Europe Facility under grant agreement n° INEA/CEF/ICT/A2018/1810854.

Since data may not be modified after publication to ensure reproducibility of results, the preservation of evolving datasets such as data produced by running observation systems have been organized as follow:

- Snapshots of the datasets are periodically extracted from the evolving database (e.g. one snapshot per month);
- A DOI and the associated landing page is minted to the evolving dataset. Associated metadata describe the full datasets.
- DOI fragments are minted to the periodic snapshots and describe the temporal interval covered by the snapshot. The DOI fragment refers to the global DOI.

3.5 Preparation for long term preservation

The SeaNoe service relies on a catalogue hosted by a SQL database. However, to facilitate the long-term preservation of both metadata and data, the following file system structure (cf. Fig. 5) has been adopted:

- One directory per published data set which includes:
 - The dataset file(s) itself (themselves),
 - The associated metadata formatted in JSON,
 - Any document(s) referenced in the associated metadata (illustrations using image formats, texts using PDF format).

This directory is self-sufficient to describe and retrieve data and can be copied in a persistent data library.

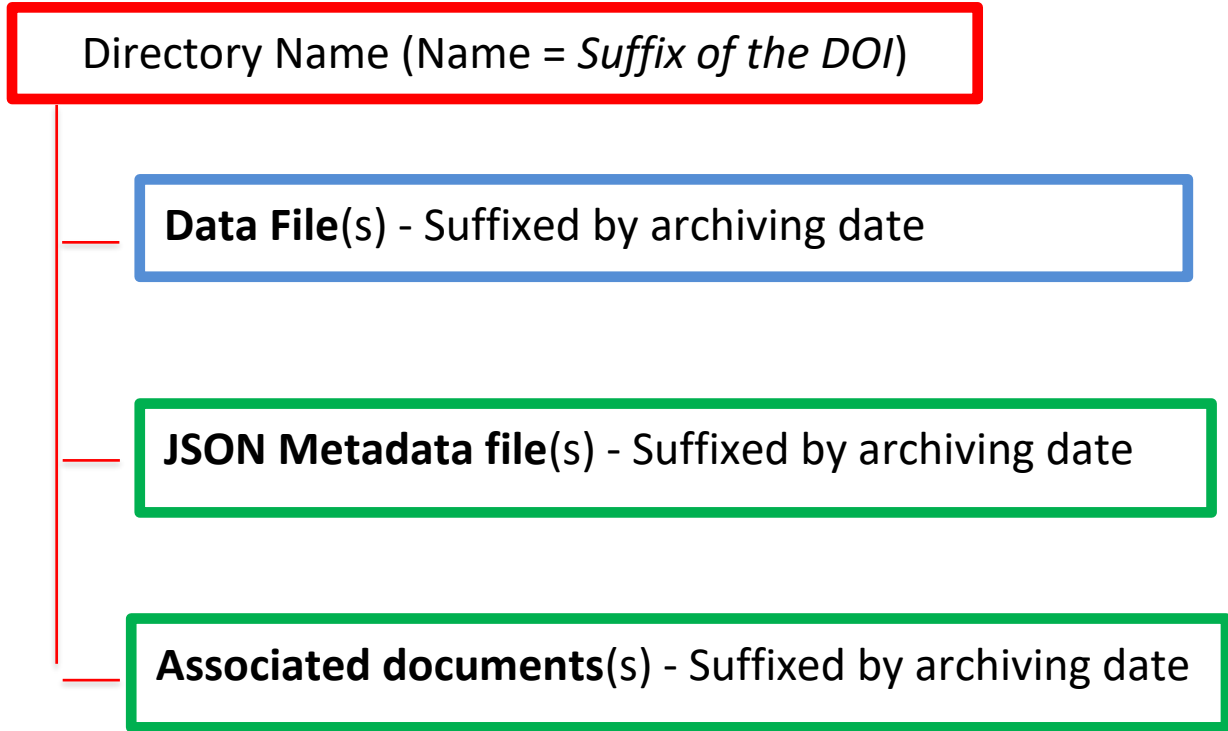
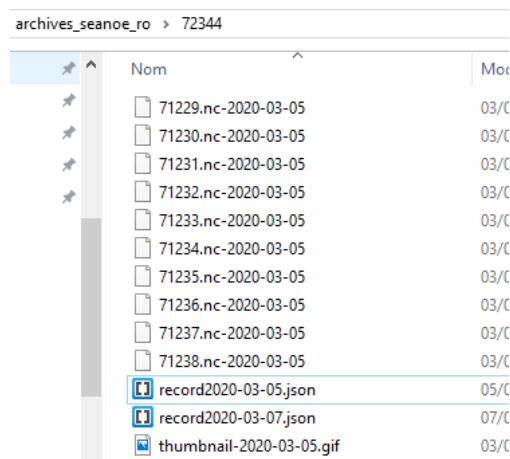


Figure 5: Structure of data files for data preservation.

Example:



Nom	Mod
71229.nc-2020-03-05	03/0
71230.nc-2020-03-05	03/0
71231.nc-2020-03-05	03/0
71232.nc-2020-03-05	03/0
71233.nc-2020-03-05	03/0
71234.nc-2020-03-05	03/0
71235.nc-2020-03-05	03/0
71236.nc-2020-03-05	03/0
71237.nc-2020-03-05	03/0
71238.nc-2020-03-05	03/0
record2020-03-05.json	05/0
record2020-03-07.json	07/0
thumbnail-2020-03-05.gif	03/0

4 New requirements

The SeaNoe publication/archiving service is now well established and fits quite well with the need of marine data producers. Data management procedures have been set up in back office to check published datasets, and to notify data centres that are in charge of the related data type.

At present, these data management procedures are performed manually by EMODnet helpdesk (Odatis operated by Ifremer/Sismer in France) after each data publication:

- Check the thematic type of data and the Data Provider country, if permitted by the data licence granted by the data provider;
- Warn the thematic Data Centre in charge of this data type in France or in Europe;
- Provide access to the data files for the thematic Data Centre;
- Notify the data provider that submitted data are also transmitted to another data centre for integration in a collection, according the granted data licences

The formal identification of all data flows between data centres is now studied in the framework of EMODnet Ingestion service (“Pathways” task).

Consequently, new requirements that will be implemented as part of the use case are more oriented towards:

- Improving the scalability of the service;
- facilitating back-office exchanges between data centres in charge of related data types;
- securing long-time archive by curating several copies in several locations.

1.1 Service scalability

Due to the allocated storage resources, and due to the limitations of http protocol that is used for uploading datasets, the maximum allowed size of data sets is presently 0.2 TBytes.

To supersede this limitation, use of other protocols, that can be asynchronous, might be considered.

In addition, sharing the allocation necessary storage resources from different infrastructures may improve the total capacity of the systems.

It is foreseen to study how far the use of Virtual File Systems is relevant for those purposes.

1.2 Back office exchanges

One of the objectives of the SeaNoe service is to collect as much as possible long-tail data. As explained before, in the marine domain, long-tail data are very valuable because it is merely impossible to reproduce observations (environmental changes, cost of observation at sea).

As a consequence, it is of great interest to assemble these data in large “data collections”. This task has to be performed in routine mode by the Data Centres that are part of the French Ocean Cluster Odatis or by all data centres that are partners of the pan-European SeaDataNet marine data management infrastructure.

This task requires many data exchanges between the involved Data Centres. At present, this task is performed mainly manually.

Improving data exchange facilities between data centres and automatize them will be of great interest, by for example implementing iRODS data flows.

1.3 Securing long-time archive

At present, most of data safeguarding is performed by the Data Centres themselves, using they own technical infrastructures. These infrastructures are not always well adapted to this purpose, and dedicated teams for organizing and performing long term archives are not available everywhere.

Relying on professional long-term repositories (i.e. certified repositories by Core Trust Seal, ISO 14721 ...) seems to be extremely relevant.

Distributing, by e.g. using iRODS data flows, datasets that have to be archive in several geographically distributed repositories is also one of the measure that are recommended by long-term archive standards.