



D4.7.1 IMPLEMENTATION PLAN OF THE INTEGRATED POLAR DATA REPOSITORY



Deliverable number:	D4.7.1
Work package:	WP4 – Atmospheric Domain
Intermediate Objective:	IO4.3
Deliverable type:	<input checked="" type="checkbox"/> Document, report
	<input type="checkbox"/> Websites, patent filings, videos, etc.
	<input type="checkbox"/> Other: please specify
Dissemination level:	<input checked="" type="checkbox"/> Public
	<input type="checkbox"/> Restricted
Estimated delivery (bimester):	B6
Actual delivery date:	15/01/2024 (version 4)
Author(s) (Partner-OU):	Vito Vitale, Giulio Verazzo, Alice Cavaliere, Mauro Mazzola (CNR-ISP-BO)
Reviewed by:	Lucia Mona (CNR-IMAA), Gianluca Di Fiore (CNR-IMAA), Ilaria Rosati (CNR-IRET), Andrea Tarallo (CNR-IRET)
Note:	

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System - CUP B53C22002150006 (D.D. n. 130/2022)
 Funded by EU - Next Generation EU
 Mission 4 “Education and Research” - Component 2: “From research to business” -
 Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Table of contents

1.	<i>INTRODUCTION</i>	4
2.	<i>THE BASIC CONCEPTS</i>	4
3.	<i>POLAR DATA REPOSITORY</i>	6
3.1	Implementing the System of Systems approach	7
4.	<i>METADATA FLOW AND MODEL</i>	9
4.1	Metadata Model	10
4.2	Help users to populate metadata catalogue	13
5.	<i>SOFTWARE APPLICATIONS AND FUNCTIONALITIES</i>	13
5.1	Common node and First Level Nodes: GEONETWORK	14
5.2	Second Level Nodes: ERDDAP	14
6.	<i>THE DATA FLOW</i>	15
7.	<i>DATA INTEGRATION AND ADDED VALUE</i>	19
7.1	The analysis environment and tools for data integration	21
7.2	Added value and products from data integration	21
7.3	Visualization as first powerful tool for data integration	22
8.	<i>CONCLUSIONS</i>	24
	<i>REFERENCES</i>	24
	<i>ANNEX DETAILED DATA MANAGEMENT PLAN</i>	25

Index of tables

Table 1 - A schematic description of sections of Metadata Model we will adopt.	12
---	----

Index of figures

Figure 1: Polar Data System overall Architecture	6
Figure 2: Implementation of the System-of-Systems approach in NADC and IADC	7
Figure 3: Data and Metadata Flow	10
Figure 4: Hierarchical structure of metadata.....	11
Figure 5: The lifecycle model of the Digital Curation Centre (SOURCE: https://www.dcc.ac.uk/guidance/curationlifecycle-model).....	15
Figure 6: The Journey of our data from the measurement site (or laboratory) to the correct insertion in a data center or data repository (DEPOSIT/INGEST, PRESERVE) includes a fair number of steps and the use of different devices/tools/software.	16
Figure 7: The typical data life cycle for a researcher.....	16
Figure 8: How duties and main responsibilities will be distributed in the polar Data Management system (model SOURCE: Strasser et al. Promoting Data Stewardship Through Best Practices, DataONE, 2011).....	17
Figure 9: The Data journey from observations to the Data Repository	18
Figure 10: SIOS Core Variables CNR should provide to SIOS through Data Polar Repository.....	19
Figure 11: Plan for development of added value services	20

Figure 12: Visualization tools in EmodNET.....22
 Figure 13: Visualization tools in ARICE Data Management.....23
 Figure 14: Interactive Sea Ice Graph at NSIDC.....23

1. INTRODUCTION

Since 1985, Italy has been present in the Antarctic territory and carries out research activities through the National Antarctic Research Program (PNRA) whose ownership and financing is entrusted to the MUR (Ministry of Scientific Research), while the scientific and logistical management is entrusted currently at CNR and ENEA, respectively. The Program is by its nature a multidisciplinary program that embraces all areas and disciplines that have an interest in conducting research activities in Antarctica: from marine to atmospheric and astrophysics sciences, from biological sciences to earth sciences such as geology, glaciology, geophysics.

Among the obligations (Article III 1 c of the Antarctic Treaty) that each Country that adheres to the Antarctic Treaty and participates in the SCAR is that of contributing to the development of the Antarctic Data Management System (ADMS) according to a very specific data policy (SCAR, 2011) and following the Data and Information Management Strategy (DIMS) developed by the Standing Committee on Antarctic Data Management (SCADM).

Considering the technological developments that have been made since the early 2000s based on approaches that favored distributed networks, interoperability, brokering services, the creation of a National Antarctic Data Center (NADC - <https://www.pnra.aq/it/dati>) was carried out starting from 2018-2019 based precisely on these new technical principles and capable of managing the large quantity of data and information acquired from the PNRA research projects. The ideal objective is to be able to ensure as much as possible at least the data collected since 2010. As first step focus was on metadata, using functionalities of the open-source software adopted to provide link to data (Chiarelli et. al., 2020).

At the same time in the Arctic, Italian participation in the large research infrastructure SIOS (Svalbard Earth Observation Integrated System) led to the start of the development of an Italian Arctic Data Center (IADC – www.programmaricercaartico.it/iadc) with the aim of collecting the mass of data collected in particular at the Arctic station Dirigibile Italia and make the same available to the SIOS Data Management System (SDMS), also based on the principles of distributed network and interoperability.

The implementation plan illustrated below starts from these experiences and what has been done so far in the development of both the NADC and IADC.

The general objective is to use the resources made available by ITINERIS to primarily strengthen Data Management of the observations collected in the Arctic, with a special focus on long-term activities carried out in all domains (atmosphere, marine, cryosphere, ecosystems), and the Italian Arctic Data Center (IADC). But, at the same time, we also seek as much as possible the obvious synergies with NADC and in this way in a cost-effective way optimize and strengthen the entire Italian polar research system from the point of view of management and display of data collected in the polar areas.

2. THE BASIC CONCEPTS

The overarching target of the implementation plan is to create an infrastructure that has the following characteristics:

- Be an integrated system that provides a single point of access for the end user to metadata and data stored in different repositories;
- Ensure quality, traceability and accessibility of data collected within the PNRA projects, promoting research and scientific dissemination;
- Facilitate the dissemination of data produced by the PNRA to the national and international scientific community;
- Adopt and promote Open Science (European Commission, 2016) according to the Findable, Accessible, Interoperable, and Re-usable data (FAIR) guiding principles.

Following also what reported in several guidelines (European Commission 2015, GEO 2015), two concepts/approaches are very important in order to accomplish this overarching objective:

A distributed system

The panorama of Italian research in the polar regions offers an extremely varied scenario, with a plurality of subjects and systems both if we look at the current situation and even more so if we look back at the history of Italian research since 1985. The variety and complexity with which the NADC and IADC systems must interface, brings with it the need to build flexible data infrastructures to the point of guaranteeing identity and management autonomy of the different parts, while integrating and connecting them in a common structure with its own well precise identity.

The approach based on a distributed network of system components for the management and sharing of data and information allows all of this, especially if the implemented architecture is based on the "System-of-Systems - SoS" concept.

System-of-Systems approach/architectures

With this expression we indicate an infrastructure where a set of systems - functional nodes – are interconnected through mediation and adaptation services, implemented by a central software component - Common Node - which allows the sharing of resources within a more complex system (Nativi, 2013; Nativi, 2015).

This approach guarantees synergy and integration between the existing information systems (or in the process of being created) of the Public Research Bodies, Universities and other organizations participating in the PNRA, preserving their technological and strategic autonomy. This distributed architecture allows you to pool the resources and capabilities of existing systems (and those coming in the near future) to create a more complex system that offers functionality and performance that is not the simple sum of existing systems. The "bottom-up" architectural approach guarantees the managerial autonomy and technological diversity of each individual information system, and at the same time the full interoperability of its resources through the use of standard interfaces.

The interoperability of the system is guaranteed by:

- The adoption of common standards for metadata and international standard protocols
- The use of controlled vocabularies that will guarantee the sharing and semantic harmonization of metadata and data attributes.
- The definition of a data sharing and governance policy dictated by the PNRA
- the service level (Quality-of-Service: QoS) and indication (via appropriate flags) of the qualification of the data reported in the IADC and NADC Polar Data Repositories

The characteristics of a similar infrastructure if realized following common standards allow to communicate and share resources both inside the ITINERIS overall system and with the main European and international systems, programs and initiatives in the polar research sector and beyond.

3. POLAR DATA REPOSITORY

The Italian Polar Data Repository is a scientific and technological infrastructure designed to gather, handle, publish and provide access to scientific data and metadata regarding Polar regions.

The research activities in the Polar Area are promoted and supported by two government funded research programs: the Italian Antarctic National Research Program (PNRA) and the Italian Arctic Research Program (PRA).

Figure 1 presents the overall conceptual model of Polar Data Repository

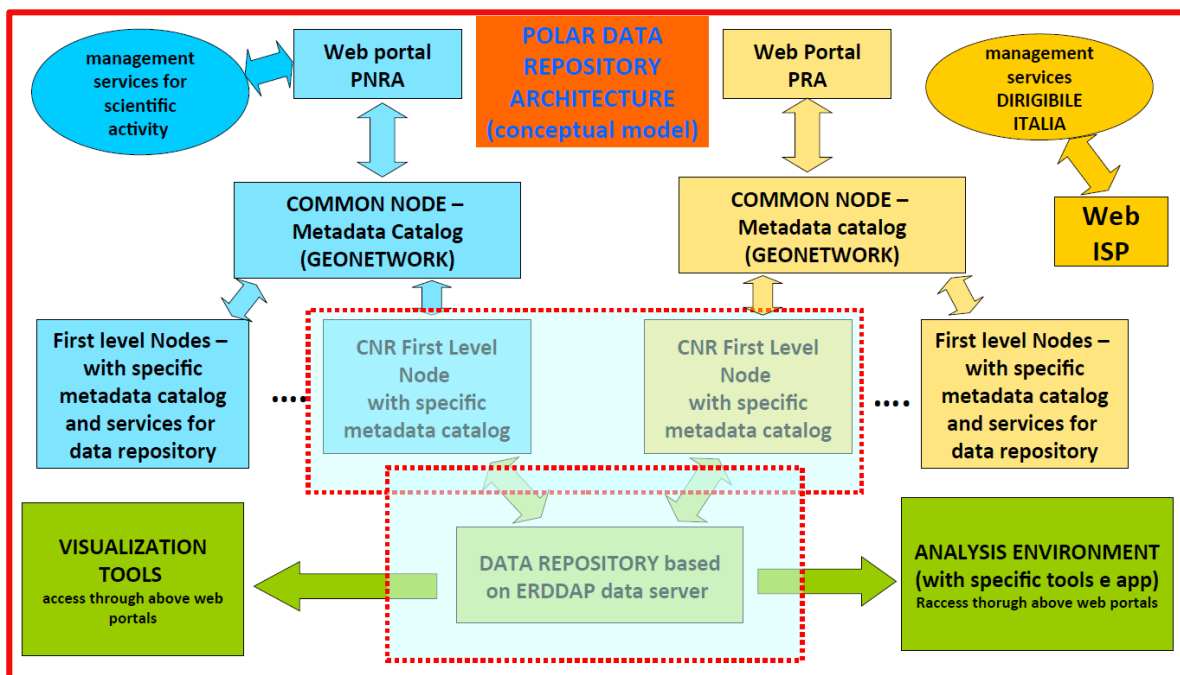


Figure 1: Polar Data System overall Architecture

The presence of two separated programmes makes necessary to keep separate the front-end through which NADC and IADC data infrastructure and repositories can be reached. Figure 1 clearly shows how data services will be generally accessed through the web portals that both the PNRA and the Arctic (PRA) have set up.

In any case, Figure 1 also shows how, the backend structures, in particular those with data repository function, will be developed in common. This will optimize hardware and software resources and will facilitate the future integration of Arctic Antarctic data repositories. Currently metadata catalogues are at disposal for NADC and IADC (for NADC also at the level of First Nodes). Instruments to populate metadata catalogue is also available (cfr. www.pnra.aq/dati). Instead, data repository functionalities are currently implemented only for IADC.

3.1 Implementing the System of Systems approach

Following the basic concept in section 2, the architecture of the IT infrastructure is based on the concept of "System-of-Systems - SoS": a set of systems interconnected with each other through mediation and adaptation services, which are commonly called "brokering services" (Nativi, 2013), implemented by a central software component - hereinafter called the "Common Node" (Nativi, 2015).

This distributed architecture allows to pool the resources and capabilities of existing systems (and of those that will come in the near future) in order to create a more complex system that offers functionalities and performance that are not the simple sum of the existing systems.

The proposed architectural approach ("bottom-up") guarantees the management autonomy and the technological diversity of each single information system, and at the same time the full interoperability of its resources through the use of standard interfaces.

In fact, the "brokering services" provided by the "common node" mediate between the different standard communication interfaces implemented by the existing information systems, which therefore must not modify either their technology or the communication standards adopted. This methodology allows to interconnect heterogeneous digital systems and infrastructures in a flexible, expandable and sustainable way (Nativi 2012). These systems can thus be part of an integrated system that not only allows diversity (of technology and scientific disciplines) but also eases of use, allowing scientists to focus more on their research, leaving interoperability technologies to expert components and services managed by the "Common Node". The digital infrastructure to be created provides for the activation of various distributed components (hereinafter called functional system nodes) for the management and sharing of data and information. Each node represents an existing (or under construction) data system and/or information.

The functional nodes are interconnected and federated with each other (see Figure 1) by means of international standard interfaces and metadata models widely supported by scientific communities. Referring to Figure 2, three types of nodes are identified:

- Common Node
- First Level Nodes
- Second Level Nodes

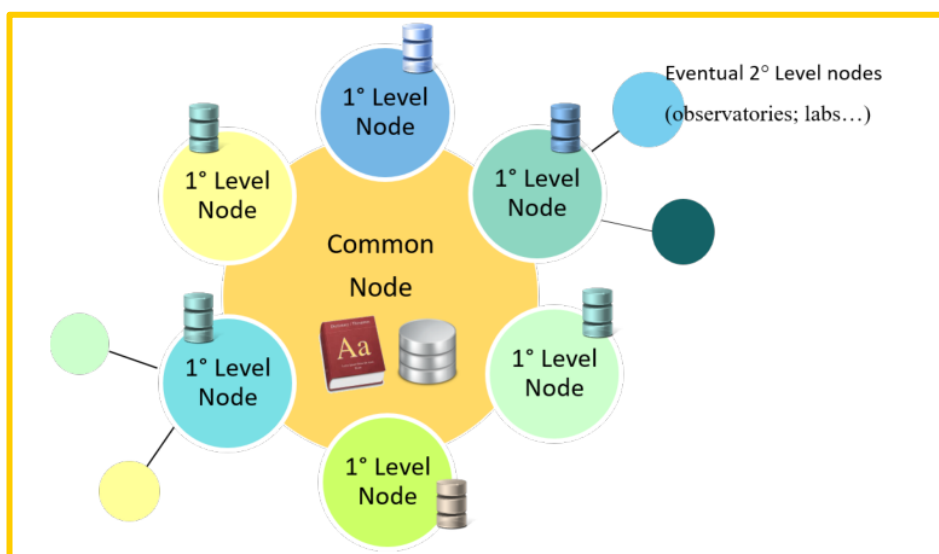


Figure 2: Implementation of the System-of-Systems approach in NADC and IADC

COMMON NODE

There is only one Common Central Node and it is the only node of the distributed infrastructure that must be managed in a shared way as it has the following functions:

1. Interconnect all First Level Nodes by implementing the adapters necessary to interface with them.
2. Perform a regular "harvest" of the metadata published by the First Level Nodes, harmonizing them according to a common scheme and implementing the mediators necessary to map the metadata schemes published by the First Level Nodes. Among the metadata conveyed by the Common Node there must also be useful information for accessing the data residing in the First Level Nodes.
3. Allow remote access to data shared by First Level Nodes.
4. Maintain a common vocabulary for shared semantics to allow harmonization of metadata published by First Level Nodes and facilitate the discovery of resources.
5. Implement the Data Policy defined by the PNRA/PRA/ITINERIS.
6. Support and publish the infrastructure web portal where it will be possible the search, access and use of resources shared by First Level Nodes.
7. Create the interface to the outside networks.
8. Ensure interoperability towards the main European and international programs and initiatives in the polar research sector.
9. Guarantee the level of service (Quality-of-Service: QoS) defined by the PNRA/PRA/ITINERIS.
10. Ensure shared governance with First Level Nodes.

The Common Node manages a harmonized copy of all the metadata shared by the First Level Nodes. However, the "master" copy of the metadata remains in the First Level Nodes, which have the task of maintaining and updating it. The common node has the task of periodic harvesting to ensure synchronization and update. The Common Node does not store or manage any data. These are and remain resident and managed only in First Level Nodes (and where required in Second Level Nodes).

FIRST LEVEL NODES

The First Level Nodes are heterogeneous systems that represent different Bodies and/or Organizations involved in PNRA/PRA/Polar research.

These contribute to PNRA/PRA/Polar Research by making data, information and knowledge available and by guaranteeing their care and preservation. Precisely in order to guarantee their sustainability and evolution over time, their autonomy and their technological diversity must be guaranteed by the overall infrastructure.

First Level Nodes must implement the following functions:

1. Publish on the Internet one or more shared resource discovery services (data, information, services, tools, etc.).

2. Publish on the Internet one or more access services to shared resources (data, information, services, tools, etc.).
3. Share with PNRA/PRA/ITINERIS a common Data Policy.
4. Guarantee the required level of quality (i.e. completeness of metadata, standard data format, etc.) to resources shared.
5. Guarantee the updating, maintenance and preservation of the services and resources shared.
6. Participate in the governance of the Common Node.
7. Allow Second Level Nodes (who request it) to manage and publish their resources: data, information, services, etc.
8. Request and control a minimum level of quality from resources shared by Second Level Nodes, in order to be shared in the general infrastructure.

Each First Level Node guarantees a level of management, "data curacy" and "preservation" in line with European standards; implements the recommendations of the European Commission - H2020 for Open Access to it; plays a role of aggregator of the data and information provided by the Second Level Nodes that is connected/referred to this specific First Level Node.

SECOND LEVEL NODES

Second Level Nodes are typically data systems (or databases) created for project and/or experimental needs. They interface directly with a First Level Node that has the task of ensuring their sharing in the distributed infrastructure. The functions required for a Second Level Node are:

1. Publish their resources on a First Level Node in order to guarantee sharing and preservation.
2. Guarantee the required level of quality (i.e. completeness of metadata, standard data format, etc.) of the shared resources.

Each Second Level Node uses a First Level Node to ensure the preservation of data and metadata generated or used in research financed by the PNRA/PRA/Others. Each Second Level Node guarantees the level of "data quality" defined by the First Level Node to which it belongs.

4. METADATA FLOW AND MODEL

The flow of data and metadata in the infrastructure is described in Figure 3.

The Common Node manages a harmonized copy of all the dataset metadata shared by the First Level Nodes. However, the "master" copy of the metadata remains in the First Level Nodes, which have the task of maintaining and updating it. The Common Node executes a periodical harvesting from all First Level Nodes in order to ensure synchronization and update.

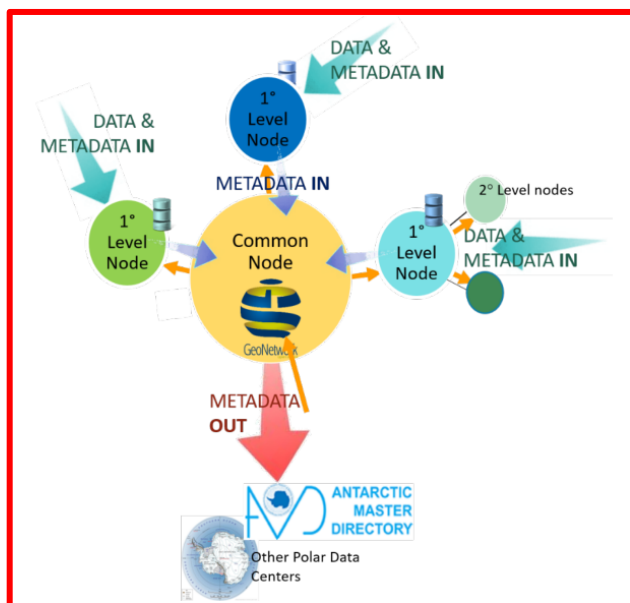


Figure 3: Data and Metadata Flow

Each Second Level Node uses a First Level Node to ensure the preservation of the data and metadata generated. In all the cases, it is recommended that metadata sent to Common Node contains the link to the original data and metadata location.

Finally, the Common Node ensures the interoperability towards ITINERIS Hub, the main European and international programs, and initiatives in the polar research sector.

This data and metadata flow allows high-level discovery through the Common Node and then moves down towards a more detailed browsing directly at the First Level Node's level. We will come back later on data flow in order to consider and describe it not only for the part referring to the IT infrastructure and repositories, but also in relation to all the steps necessary to prepare them for data repositories, with the aim to give evidence and discuss the cooperation work between researchers and data experts/managers. A cooperation that is necessary to activate/pursuit to assure a robust and efficient flow of data in the system.

4.1 Metadata Model

Choosing the metadata format for Common Node faced several challenges:

- It had to be shared by the entire Italian Polar research community;
- it had to consider that some of the top-level Nodes already had different metadata formats depending on their data domain;
- it had to be easily shareable with national and international scientific community in general, and with ITINERIS community in particular, through standard interfaces.

The model chosen for the metadata is the ISO 19115 standard, which is widely used within the scientific community for the description of georeferenced data. Compared to the DIF format, the ISO 19115 model integrates descriptive metadata, i.e. metadata that contains information that identifies and describes the resource to which they refer with a more general character. The ISO 19115 standard scheme in fact provides information such as: title, abstract and objective of the resource, program body/coordinators, project contacts, temporal and geographical references of the data. Metadata may

contain links to external resources such as dataset repositories, project pages, and other information. The resources described through metadata can be both datasets and projects. Finally, the possibility of creating links between the different metadata gives rise to a hierarchical structure that allows a detailed description of different datasets referring to the same project context.

Hierarchical structure of metadata

Dynamically sharing metadata at the dataset level is becoming increasingly important. In any case, considering how research activity is developed/identified in the frame of Italian polar programmes PNRA and PRA, is also very important to share project-level metadata. This means to include high-level information like the title, funding agency, location of the project, etc. for project monitoring, program coordination and other purposes.

As optimal solution we will implement a hierarchical format which not only allows the level to be flatly described of the metadata of the georeferenced dataset, but also to describe its design context in an articulated way.

As described in Figure 4, considering that a specific activity/project could lead to many different datasets, we will define the project itself through the ISO19115-3 format (as already done for the British Geological Survey), in order to describe it in detail and connect it then to all possible datasets, in turn described through ISO19115-3.

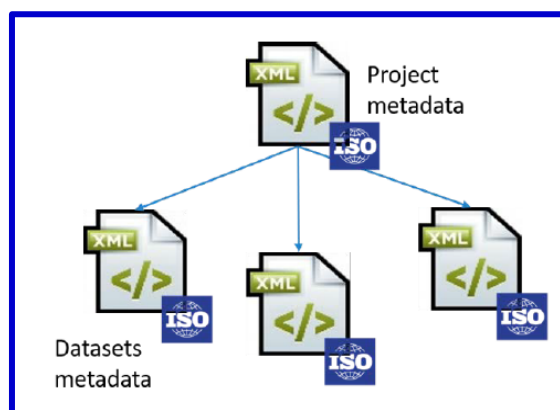


Figure 4: Hierarchical structure of metadata

An advantage of the ISO 19115-3 metadata standard is that it allows interconnected layers of information to be effectively combined. Metadata about projects can be nested within metadata about scientific datasets. This hierarchical approach makes connection across the project-data lifecycle possible.

This approach is also flexible, scalable and efficient, as nested metadata can be dynamically embedded with "XLinks" - external URLs embedded in XML tags - so that projects and datasets are brought together with information from different and appropriate sources (in a distributed system of interoperable nodes) (similar use has been made for the Arctic Observing Viewer scientific data and project management system).

The fields of the metadata model

The fields requested during the compilation of the template will be chosen with the aim of guaranteeing the validation of the metadata for the ISO 19115 standard. The scheme, designed for the collection of information concerning a resource characterized by a temporal and geographical

reference, offers a description both specific to the resource as well as generic to the context (e.g. project) in which it is inserted. By resource we typically mean a dataset which can be a single file, a compressed archive containing multiple files, files present on a data repository platform (e.g. Zenodo, Hyrax, ERDDAP, its own FTP server, etc) or other types of formats of data.

The organization of the fields to describe the metadata will be structured into six main sections:

- Identification info
- Reference System Information
- Distribution Information
- Data quality info
- Metadata
- Associated resources.

The table below (Table 1) provides a schematic description of these sections.

Table 1 - A schematic description of sections of Metadata Model we will adopt.

Identification Info Citation Identifier Point of contact Maintenance Keywords and adopted vocabularies Resource constraints Aggregation information Spatial resolution, language and topic category Extent	Provides a series of generic information structured into several subsections describing the resource to which it refers	Mandatory
Reference System Information	specifies the reference system used for georeferencing the resource	Mandatory
Distribution information	specifies information regarding the distribution of the metadata	Mandatory
Data quality info	specify information regarding the quality of the dataset	Mandatory Some items will be pre-populated for INSPIRE validation

Metadata	specifies the characteristics of the metadata and who is responsible for it	Mandatory
Associated resources	resources associated with the metadata: link to a descriptive page of the dataset or project, link to a service that distributes the dataset or the dataset itself.	Optional inserting it into the form allows you to delve deeper into the information related to the resource

4.2 Help users to populate metadata catalogue.

Various tools will be defined and created to make researchers' activity in populating metadata catalogue as independent and autonomous as possible. This is from the already mentioned perspective that Data Management activities must provide for the widest possible synergy between the community of researchers who are the suppliers, and the community of IT experts/data managers. This material will be added to a template that will also be created and which aims to guide the researcher and simplify the insertion work. Webinars and a build guide will be among the main tools produced and implemented.

The purpose of the webinar is to provide researchers and PIs with the skills to independently insert and manage metadata relating to their datasets/scientific projects on the IT platform we selected for this scope (cfr. Section 5 below). The webinar will be organized in a theoretical part concerning the data infrastructure and the characteristics of the system, and in a practical part which consists of the description and explanation of the fields required by the metadata scheme through an example of compiling the record.

The guide for compiling and managing metadata represents in some way an off-line analogue of the template. Through this guide, the researcher will be able to understand the quantity and type of information he must have available to compile the metadata, thus facilitating the subsequent approach to the IT platform and the template. It also remains as a consultation tool while the metadata is compiled and inserted.

All this material will be made available and consultable through the portals of both the PNRA and the PRA.

5. SOFTWARE APPLICATIONS AND FUNCTIONALITIES

A fundamental choice obviously concerns the technologies and software on which developing the entire infrastructure and all its different functions. From this point of view, the choice made is to use Open Source technologies, software and standards. This allows for easy reuse and no licensing costs, and allows you to allocate all economic resources on the development phases. Advantages that also connect to this choice, if carefully made, are (i) being able to count on products always developed at the state of the art by the consortia in charge of the management and upgrade of the selected software, and (ii) being able to share a vast community of users who in turn are also developers of specific application solutions which can be drawn upon where necessary/useful. Based on these considerations we have therefore defined two guiding criteria for choosing the technical solutions to adopt:

1 - the presence of strong and consolidated development teams and a large community of users

2 - to consistently adhere to the principles of Open Science, try as far as possible to target products and consortia that are public. This will guarantee the highest possible level of adherence to all the principles connected to data management: open methodology, Open source, Open data, Open access.

Using these criteria, after a careful comparative analysis of the software available for managing georeferenced metadata, it was decided to adopt GeoNetwork Open Source, a mature software, currently used in numerous spatial data infrastructure initiatives around the world, including FAO, WHO, WFP, etc. While for data management we focused on the ERDDAP software considering that it is supported by NOAA, also considering that it is the software used by some of the most important environmental data integrators at European level, such as EMODnet.

5.1 Common node and First Level Nodes: GEONETWORK

GeoNetwork allows to manage the insertion, maintenance and faceted search of metadata in various formats including ISO19115-3 and permits validation of records against ISO and INSPIRE rules (INSPIRE, 2013). It provides many harvesting services that adapt to different standard protocols (such as OAI-PMH;WFS; REST...) and to different metadata formats. It has a well-developed interface that can easily be configured and changed, it has an interactive map viewer and it is already multilingual. Moreover, it has built-in schema.org Dataset and Catalog annotations in order to make datasets searchable through Google Datasets. All these characteristics perfectly fitted the role of Common Node in the architecture. As from the initial goals of the project the evolution and technical choices of the First Level Nodes had to be preserved and independent from the Common Node technical choices. Therefore, some First Level Nodes (especially the ones that didn't have an Information System already in place) decided to install a local version of GeoNetwork in order to manage their metadata through it, some others preferred to maintain their own information system and make their metadata available through standard protocols and formats.

5.2 Second Level Nodes: ERDDAP

While Common Node and First Level Nodes manages metadata, data are in charge of Second Level Nodes. Partners of PNRA/PRA that does not have a custom solution for data management are opting to use ERDDAP, an open-source software made by NOAA.

ERDDAP data server is open-source software written in Java that builds upon the open-source ideals of the OPeNDAP, WCS, SOS and OBIS standards. ERDDAP supports both human interaction (e.g. OPeNDAP requests) and machine-to-machine interoperability. ERDDAP data server supports several common data file formats (html table, netcdf, csv, txt, mat, json, etc.) and output files are created on-the-fly in any of these formats. ERDDAP implements FGDC Web Accessible Folder (WAF) with FGDC-STD-001-1998 and ISO 19115 WAF with ISO 19115-2/19139.

- ERDDAP tries to solve the problem of managing different data formats and sharing protocols.
- ERDDAP offers an easy-to-use, consistent way to request data: via the OPeNDAP standard. Many datasets can also be accessed via ERDDAP's Web Map Service (WMS).
- ERDDAP returns data in the common file format of your choice. ERDDAP offers all data as .html table, ESRI .asc and .csv, Google Earth .kml, OPeNDAP binary, .mat, .nc, ODV .txt, .csv, .tsv, .json, and .xhtml. Eliminating the need to manually reformatting data.
- ERDDAP can also return a .png or .pdf image with a customized graph or map.

- ERDDAP standardizes the dates + times in the results.
- ERDDAP has web pages (for humans with browsers) and RESTful web services (for computer programs).

6. THE DATA FLOW

The data lifecycle or data flow has been largely conceptualized and modeled by the IT community over time, precisely in consideration of the importance of its correct understanding in terms of Data Management. Figure 5 shows the conceptual model produced by the Data Curation Center (DCC) of Edinburgh. On the right, the meaning and objective of the various specific actions it envisages is summarized.

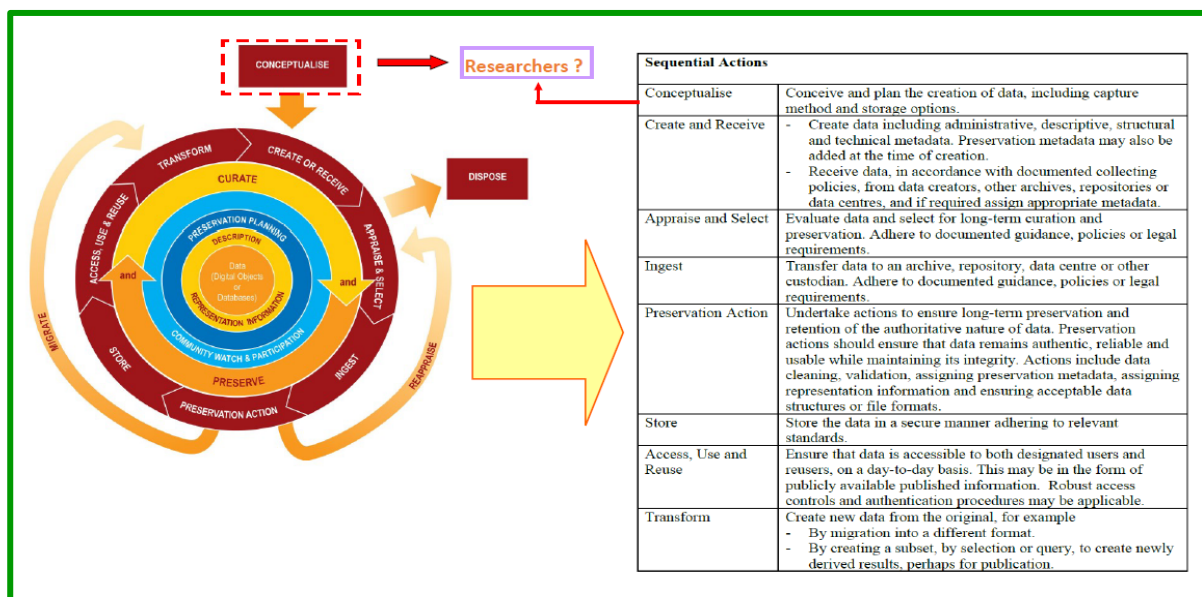


Figure 5: The lifecycle model of the Digital Curation Centre (SOURCE: <https://www.dcc.ac.uk/guidance/curationlifecycle-model>)

The DCC model is one of many that have been developed by many groups and is useful for two reasons: 1 – is graphically clear and compact; 2 – also clearly indicates how the IT community tends to not consider too much the long journey the data need to do before their arrival in the data repository. The consequence being that (i) the role of the research community is unclear/marginal (cfr. Figure 5), and (ii) they are in general very little supported and helped to be connected with the data management system and populate catalogues and repositories with resources. In this model the language and graphic solutions tend to represent a clear separation between the researcher/data producer and the IT expert/data manager. DCC is an example of a issue/problem that is more general. However, the path from the tool to the Data Repository is much longer and more complex than that exemplified by the DCC model. Figure 6 tries to give a vision of all the steps that precede the creation/transformation phase for the data infrastructure (last two boxes with dotted borders in Figure 7). Meanwhile, Figure 7 exemplifies what for the researcher is the typical life cycle and use of the data.

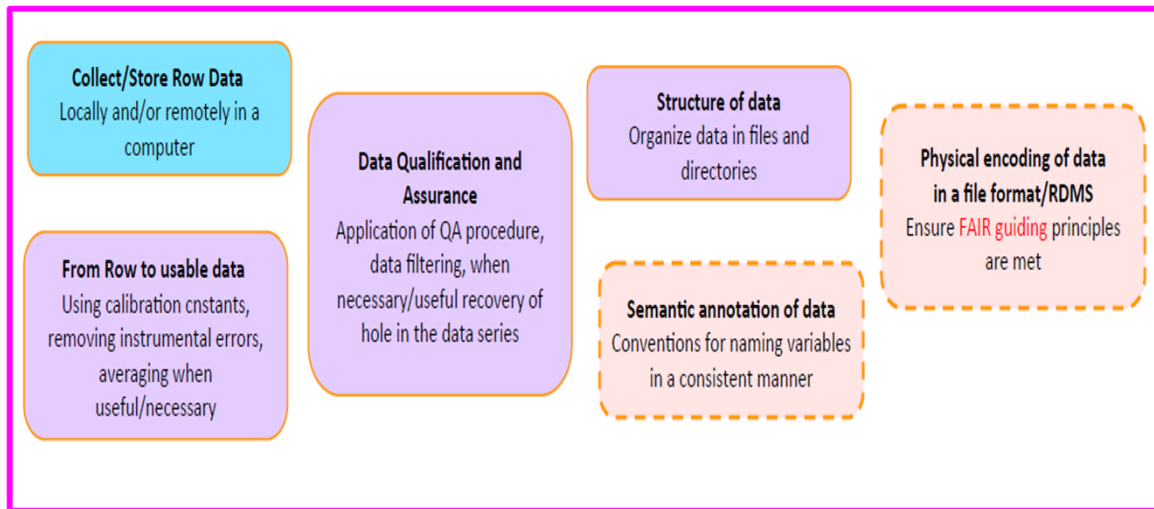


Figure 6: The Journey of our data from the measurement site (or laboratory) to the correct insertion in a data center or data repository (DEPOSIT/INGEST, PRESERVE) includes a fair number of steps and the use of different devices/tools/software.

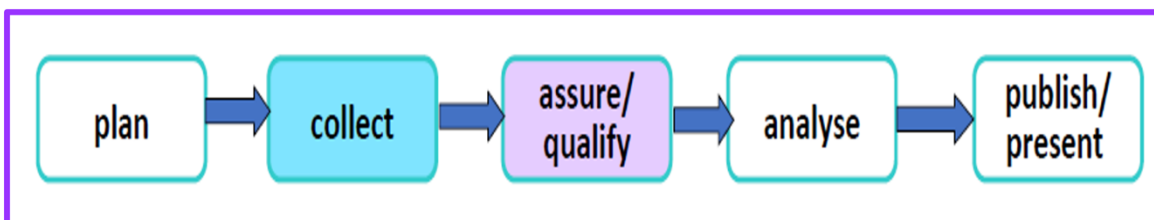


Figure 7: The typical data life cycle for a researcher

Figures 5-7 highlight how data managers tend to underestimate and/or pay little attention to (i) the great work that data producers have to do, (ii) the fact that some actions useful for data management are done for another purpose (publishing), and (iii) the data creation part for the data repository is something completely new for researchers and often still too little valued in the research evaluation system and therefore not very useful for career and CV purposes.

These considerations lead to the obvious conclusion that if one wants to implement a robust and constant flow within a data infrastructure, and create the best conditions for its sustainability and expansion in the medium and long term, the philosophy that must be adopted is that of a close cooperation, and to work as much as possible to create tools and functions that help the researcher, helping to ensure that reduced motivation is not reflected in a non-submission of data. In the long term, the advantages of data sharing will also come in handy, especially if the infrastructure manages to create good services to extract value and new knowledge from the integration of different data (much more important in this sense is the "bio-diversity" of this data rather than their quantity. But in the short term, this policy and approach is certainly the most valid choice and will be the one we will follow. Figure 8, providing another graphical representation of the data life cycle, can be used to highlights how roles and tasks between data providers and data managers will be distributed in the system we aim to implement for polar data. In the Figure, areas where the cooperative approach will be developed to the maximum, are clearly shown. In addition to the area of creation and reception of

data sets for the data repository (following the terminology of the DCC lifecycle model), the area aimed at extracting new knowledge from the analysis of the data collected in the repository will also be important. We'll talk more about this in the next section.

A last important aspect to consider is the fact that the data is not all the same and that there is a big difference both if we look at different domains and if we consider different areas of the same domain. For example, focusing on the atmosphere, it is clear that data obtained continuously from automatic tools are very different from data collected during measurement campaigns, and the measurements/observations of physical parameters are very different in their treatment and management from a Data Management perspective, from chemical data that are usually produced through laboratory analysis.

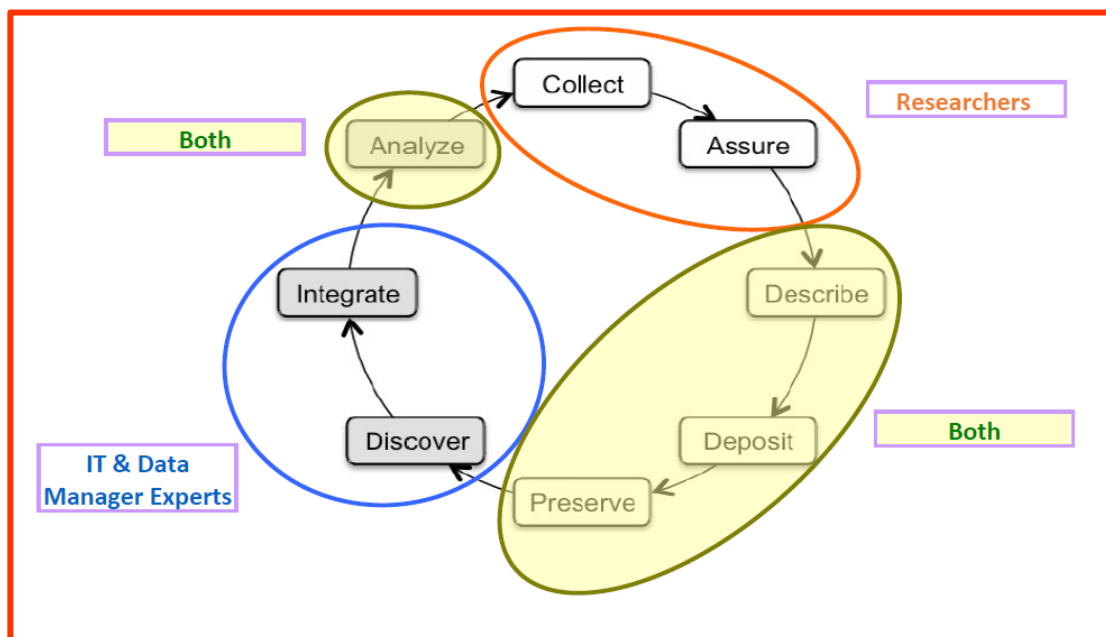


Figure 8: How duties and main responsibilities will be distributed in the polar Data Management system (model SOURCE: Strasser et al. Promoting Data Stewardship Through Best Practices, DataONE, 2011)

Figure 9 provides a schematic view of the system we aim to implement to ensure data flow towards the Italian polar repositories (National Antarctic Data Center - NADC; Italian Arctic Data Center - IADC). The approach arises from all consideration made above. In particular, it is adapted both to the type of data and to the capacities/resources of the different providers.

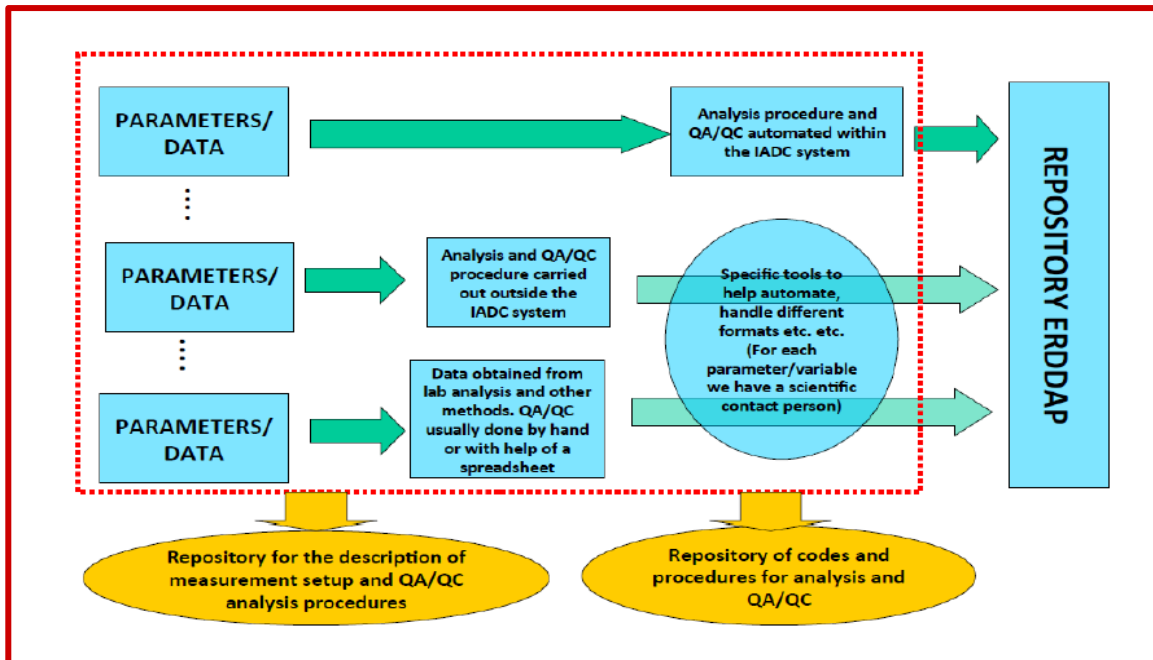


Figure 9: The Data journey from observations to the Data Repository

Figure 9 provides a data flow perspective very focused on the actions that are upstream of what is the usual area of expertise of Data Management plans. While Figure 3 reported in section 3 about metadata allows us to appreciate what the data flow will be within the distributed infrastructure. Note in Figure 9 the importance that it will be for us to have complete documentation not only of the phases of inserting the data into the repository, but also of all the actions that have been carried out on the original data, including information on the experimental setup which is the basis of the acquisition of the data themselves.

Speaking of the data sets and parameters on which we will focus attention in the implementation of the scheme illustrated in Figure 9, our starting point will be the "core data", as defined within SIOS. One of the main products and services of SIOS are data useful to address identified key research questions in Earth System Science (ESS). The core observational programme of SIOS should provide the research community with systematic long-term observations, yet flexible enough to integrate upcoming new methods and research questions.

SIOS Core Data has to fulfil the following criteria:

1. Scientific requirements: the variable is critical to answer the key research questions as defined and addressed by SIOS. Connection with GCOS ECVs and other Essential variable schemes, as for example Essential Ocean Variables (EOVs), and marine Essential Biodiversity Variables (EBVs), provide guidelines and criteria for selection and prioritization.
2. Data availability: SIOS core data should be available through SDMS. Where possible, existing measurement and calibration protocols should be used, e.g. WIGOS, GAW, BSRN, ACTRIS and ICOS, to secure comparability.

3. Members commitment: for SIOS core data, there should be a commitment from the providing institute to maintain the measurement for more than 5 years and make the data available through SDMS.

Based on the first two overarching criteria, a set of 51 core variables were identified by SIOS Working groups and additional scientific experts, 30 of them pertaining to the Atmospheric domain. Considering the whole process adopted, the selected variables are critical for characterising climate system and its changes (connection with ECVs) and are also essential in answering the ESS science questions outlined in the SIOS Infrastructure Optimization Report. All of them has been identified and characterised by the Essential Climate Variables (ECVs) defined by The Global Climate Observing System (GCOS), WMO standards and the Global Change Master Directory (GCMD) Keywords. Information and complete list of the core variables and definitions can be found on SIOS web page (<https://sios-svalbard.org/CoreData>).

CNR, as member of SIOS, has ensured the acquisition and management of a certain number of these parameters, as well as making proposals to broaden the view to parameters of interest to Italian research groups. Figure 10 shows the list of variables that the CNR has ensured for SIOS for the atmospheric domain, and in red the new proposals formulated in 2020. Thanks to the resources of ITINERIS and further resources that the CNR has ensured in recent years, this commitment will be able to be expanded and strengthened.

SIOS Core Data variables	SIOS Core data products (following GCOS terminology)		
SCD 1.1. WIND SPEED			
SCD 1.2. WIND DIRECTION			
SCD 1.3. AIR TEMPERATURE			
SCD 1.4. NET RADIATION			
SCD 1.5. SHORTWAVE RADIATION			
SCD 1.6. LONGWAVE RADIATION			
SCD 1.8. HUMIDITY			
SCD 1.16. OZONE	SCD 1.16.3 Total column ozone		
	SCD 1.16.4 UV radiation		
SCD 1.19. AEROSOL PARTICLE PROPERTIES	SCD 1.19.1 Particle number size distribution -mobility diameter		
	SCD 1.19.3 Particle number concentration		
SCD 1.20. CHEMICAL COMPOSITION	SCD 1.20.3 Mass concentration of particulate organic tracers.		
	SCD 1.20.5 Mass of major ions in PM10 or PM2.5.		
SCD 1.21. CO ₂ , FLUX			
SCD 1.23. AEROSOL PARTICLE ABSORPTION			
SCD 1.24. AEROSOL PARTICLE SCATTERING			
SCD 1.31. RATE OF CHANGE OF TOTAL ELECTRON CONTENT (ROT)			
		SCD 1.32. PHASE SCINTILLATION INDEX	
		SCD 1.33. AMPLITUDE SCINTILLATION INDEX	
		SCD 1.34. TURBULENCE IN THE ABL	SCD 1.34.1 sensible heat flux
			SCD 1.34.2 friction velocity
			SCD 1.34.3 roughness length
		SCD 2.10. SNOW COVER	
		SCD 3.2 Water level	
		SCD 3.3 Water Electrical conductivity	
		SCD 3.4 Water temperature	
		SCD 4.3. OCEAN CURRENTS	
		SCD 4.5. SALINITY	
		SCD 4.8. WATER TEMPERATURE	

In red the proposal for additions to SIOS core data proposed by CNR.

Figure 10: SIOS Core Variables CNR should provide to SIOS through Data Polar Repository

7. DATA INTEGRATION AND ADDED VALUE

Once data have been collected and described, there is the need to extract value from its. The possibility to obtain new information from analysis and integration of large amount of data provided by different domains and/or different measurements sites, gives the meaning and reason to invest in Data Management and in implementing complex and expensive data infrastructures: in approaching the issue of the added value extraction from data repositories we can consider two parallel but different lines: (i) make it possible to extract new knowledge and information by integrating and comparing even very different data, and (ii) increase data use by an ever-increasing number of users from the most disparate categories .

Relatively to this point, the intention is to develop a sufficient number of services and tools by working along three precise lines of action:

- 1 - through the development and implementation of an analysis Environment that allows expert users to access data and process it freely according to their needs.
- 2 - designing a series of tools and instruments for simple and/or medium complexity processing available to less experienced users, tools that help and guide the user where needed, including in the selection of suitable data sets and parameters
- 3 - developing graphic tools that help carrying out data integration, taking advantage of their characteristics as tools that intuitively allow us to understand/glimpse relationships and can lead to subsequent in-depth, more quantitative analyses.

Figure 11 provides a schematic view of the above and allows you also to visualize the connection with the remaining part of the data management and flow.

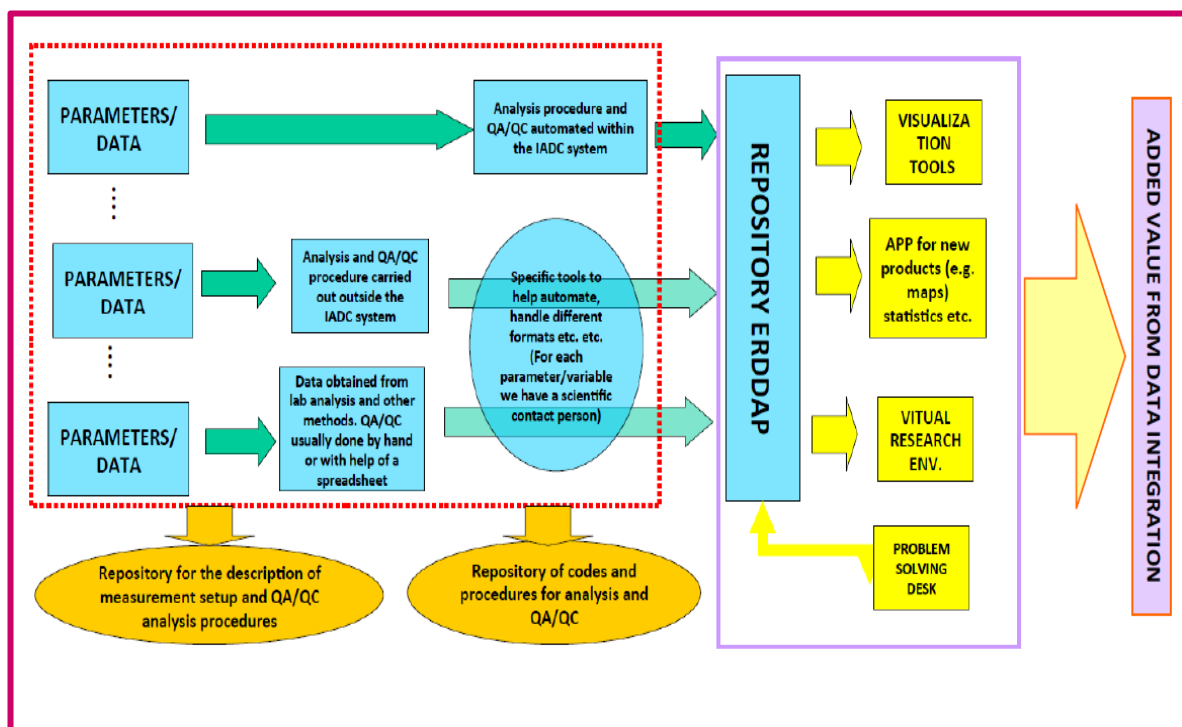


Figure 11: Plan for development of added value services

7.1 The analysis environment and tools for data integration

The Italian Polar Data Repository will integrate an analysis environment based on a Jupyter Hub development platform and a set of ad-hoc software solutions.

Jupyter Hub

JupyterHub is a powerful open-source platform designed to facilitate collaborative and interactive computing environments. It serves as a centralized hub that enables multiple users to access and work with Jupyter notebooks, a popular tool for creating and sharing documents containing live code, equations, visualizations, and narrative text. With JupyterHub, organizations, educational institutions, and research teams can provide a shared computing environment where users can collaborate on data analysis, machine learning experiments, and scientific research. This platform simplifies the deployment and management of Jupyter notebooks, making it easier for users to leverage the benefits of reproducible and data-driven workflows while maintaining security and scalability.

Jupyter Notebook

Jupyter Notebook offers fast, interactive new ways to prototype and explain your code, explore and visualize your data, and share your ideas with others.

Notebooks extend the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results. The Jupyter notebook combines two components:

- A web application: a browser-based editing program for interactive authoring of computational notebooks which provides a fast interactive environment for prototyping and explaining code, exploring, and visualizing data, and sharing ideas with others
- Computational Notebook documents: a shareable document that combines computer code, plain language descriptions, data, rich visualizations like 3D models, charts, mathematics, graphs and figures, and interactive controls

The main idea is to develop Jupyter Notebook documents that works with data from the Italian Polar Data Repository.

7.2 Added value and products from data integration

In many cases, the user needs to perform standard and/or relatively simple analyses on the data he or she is going to recover. In these cases, the added value of the infrastructure is found in being able to make a large number of data available both in terms of quantity and perhaps above all in terms of typology and multidisciplinary. Making tools and services available for these simple calculations (for example simple statistical averages and/or regressions) can make it easier for the user of the Polar Data Repository who does not have to download the data set(s) and then work locally, but can do it directly in the system the analysis and then download the result. Beyond this, there are calculations and products that are usually derived from available data. For example, we can think of the erythemal dose if we talk about UV radiation, of surface albedo if we think of solar radiation data, heat fluxes if we start from eddy-covariance data or/vertical profiles of weather parameters, integral quantities of vapor d water and/or gas if we start from vertical profiles, and for each variable or group of variables we can certainly find examples of products that are routinely then used much more than the original parameter to carry out further analyses and/or describe processes, to know the state of the system. Each researcher knows very well what can be useful and what can be done with calculations that are not too demanding from a computational point of view. They often have already

codes in their computers to make these calculations, given that they are often a normal routine from which to start in order to move forward in the development of research and work for a publication. Other times researcher has the knowledge and skills, has ideas in his/her drawer that for reasons of time and resources he/she was unable to materialize as wanted. This precious mine will be used and exploited to try to equip the Polar Data Repository with as large a number of calculation "apps" as possible which will allow also non-expert users to obtain the products and results they need, in a simple way and with little effort. Jupyter Notebook can be a resource to achieve this result. But it won't be the only one. As mentioned, these codes are often already available. And therefore, from time to time we will try to understand whether executables, or codes in Python or subroutines found in open source libraries, cannot already be used with a few modifications.

7.3 Visualization as first powerful tool for data integration

Plotting the data, and through this graphing making comparisons, and starting to understand if and what type of relationships can exist between the different variables in relation to the phenomenon and process that one is trying to investigate: this is the method that researcher usually uses every time he/she has to face a new topic. In recent years, many steps forward have been made in the direction of making graphical tools available to users of databases and data repositories to effectively visualize the selected dataset. Two examples taken from the SOOS MAP system and the EU ARICE project are shown in Figure 12 and Figure 1, respectively.



Figure 12: Visualization tools in EmodNET



Figure 13: Visualization tools in ARICE Data Management

Usually, however, these tools are limited to represent, for a single chosen variable and a single measurement station/platform, the temporal trend of the variable. Some services in EMSO and/or ICOS make two-dimensional maps of quantities possible. These tools are therefore far from the idea expressed above that researchers use to investigate and try to hypothesize relationships before then trying to determine them accurately with quantitative calculations, which can sometimes be onerous in computational terms. The Polar Data Repository, in addition to implementing tools and functions such as those presented in figures 12 and 13, will try to develop a small graphic library that makes it possible to compare the temporal trend of a parameter in different stations, or the temporal trend of different parameters in the same station/platform, comparing the trend of the same quantity in different years, as is done on the American National Snow and Ice Data Center (NSIDC) website in relation to the extent of sea ice (Figure 14).



Figure 14: Interactive Sea Ice Graph at NSIDC

8. CONCLUSIONS

The plan for the implementation of polar data repositories has been illustrated in the previous sessions in a schematic manner but with sufficient details. We have above all tried to illustrate the principles underlying the architectural choices of the system, the flow of data and metadata within this architecture, the possible connections with the outside at a national (in particular within ITINERIS) and international level, the main technological and software choices.

It should be taken into account that we arrived at ITINERIS with actions already underway and with which it was necessary to interface. From here some almost obligatory choices arise. However, having followed international standards and best practices since 2018, all this does not harm or make it complicated to place and interface in the ITINERIS system. Additionally, considering that at the polar level there are obligations regarding data at the Antarctic level, and that also at the Arctic level there are strong coordination actions that started 10 years ago in the IASC context and then merged into dialogue with the SCAR institutions in which is the Polar Forum, which will hold its fifth conference in Cambridge at the end of October. Even in this case, obligations and interactions arise which cannot be avoided for us if we want to be in the right way in the international context. A context which in the case of the polar is an unavoidable point of reference for Italian research and which can be found above all else.

In order to provide further technical details, we included in the Annex (Detailed Data Management Plan) as an example the data management plan for the polar data repository.

REFERENCES

- Chiarelli C., Longo S., Principato A., Verazzo G. and Vitale V., 2020, NADC – Italian Antarctic Data Centre. 2020, Data Science Journal.
- European Commission DG-R&I. (2015, October 01). “Guidelines on Data Management in H2020”. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- European Commission, 2016, “Open Science (Open Access)”, URL: <https://ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access>
- GEO. (2015, October 01). GEOSS Data Management Principles. Retrieved from http://www.earthobservations.org/documents/dswg/201504_data_management_principles_long_final.pdf
- GEO. (2015, October 01). The GEOSS Data Sharing Principles. Retrieved from https://www.earthobservations.org/geoss_dsp.shtml
- INSPIRE, 2013, “Guidelines for the encoding of spatial data” available at: http://inspire.ec.europa.eu/documents/Data_Specifications/D2.7_v3.3rc3.pdf
- Nativi S., Craglia M., Pearlman J., 2012, The Brokering Approach for Multidisciplinary Interoperability: A Position Paper. International Journal of Spatial Data Infrastructures Research, Vol.7, 1-15. Available at: <http://ijsdir.jrc.ec.europa.eu/index.php/ijsdir/article/view/281/319>
- Nativi,S., Craglia M., Pearlman J., 2013, “Earth Science Infrastructures Interoperability: The Brokering Approach”, Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of , vol.6, no.3, pp.1118,1129, June 2013.

Nativi S., Mazzetti P., Santoro M., Papeschi F., Craglia M., Ochiai O., 2015, Big Data challenges in building the Global Earth Observation System of Systems, Environmental Modelling & Software, Volume 68, June 2015, Pages 1-26.

SCAR - Scientific Committee on Antarctic Research. 2011. SCAR Report 39 - 2011 June - SCAR Data Policy. Cambridge, UK: Scott Polar Research Institute. Accessed 8 Oct. 2018 at <https://www.scar.org/scar-library/reports-and-bulletins/scar-reports/2717-scar-report-39/>

ANNEX DETAILED DATA MANAGEMENT PLAN

The purpose of a Data Management Plan (DMP) is to document how the data generated are handled. Together with general principles, architecture and technological choices of the whole system, all described in the report with great detail, there are also specific technical questions on standards and generation of discovery and use metadata, data sharing and preservation and life cycle management. A large variety of Template have been developed by many to help to introduce in a DMP all these information. The web site at Virginia University <https://library.virginia.edu/data/data-management-plan-templates> provide a good example of this variety and richness. For Horizon 2020 programme, EU Commission also support development of a specific DMP template for the funded project. What reported below is a DMP based on this H2020 FAIR Data Management Plan template designed to be applicable to any H2020 project producing, collecting and/or processing research data. This template can be found and downloaded at the link https://ec.europa.eu/research/participants/data/ref/h2020/other/gm/reporting/h2020-tpl-oa-data-mgt-plan-annotated_en.pdf.

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Data collected will be used to study polar regions area.

What types and formats of data will the project generate/collect?

A high variety of data formats could be generated within the ITINERIS project. Most of the data will be time series data and the most popular formats are expected to be Microsoft Excel (XLS); Comma-separated values (CSV), Text (TXT) and netCDF (NC). Parameters collected pertain to all environmental domains, with a particular focus on atmospheric and marine domains.

Will you re-use any existing data and how?

Data already produced will be reused by leveraging the harvesting feature of the repository if supported by the current platform where they are hosted. If this is not possible, existing data will be added by the default uploading process of the repository.

What is the origin of the data?

Data will be collected from scientific instruments by Italian researchers in the frame of their project activities.

What is the expected size of the data?

The precise of the data cannot be determined at this stage. This information will be provided in the following versions of the DMP.

To whom might it be useful ('data utility')?

Data will be used by researchers, stakeholders and decision makers.

2. FAIR data

2.1. Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

Yes, all the data will be described by discovery and use metadata and identified by a DOI.

What naming conventions do you follow?

The metadata standard ISO 19115 will be used as naming convention.

Will search keywords be provided that optimize possibilities for re-use?

Yes, standard vocabularies such as GCMD Science Keywords, SCAR Gazetteer of Antarctica and others will be used to provide keywords.

Do you provide clear version numbers?

Data version numbers will not be provided. Data and metadata will be tracked by their upload date.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Two types of metadata will be created, discovery and use metadata.

Discovery metadata is used to describe general information about the data and to make data discoverable in data centers in a standard format. The ISO 19115 format will be used.

Use metadata contains precise information about the data described and it used to understand the data. It is included in the data formatted as NetCDF and standards like Climate and Forecasts conventions and OceanSITES will be used.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

All metadata will be made openly available and accessible. Access to the datasets depends on the PI decision. Some data will require access request to the data originators.

How will the data be made accessible (e.g. by deposition in a repository)?

Data will be made accessible through a data server software.

What methods or software tools are needed to access the data?

The data server software used to serve the data will be ERDDAP, developed by NOAA. ERDDAP provides both a human accessible web interface and a computer accessible data API to access the data.

Is documentation about the software needed to access the data included?

Some documentation is needed to access the data and it will be provided through the web interface of ERDDAP.

Is it possible to include the relevant software (e.g. in open source code)?

All software used to develop the Italian Polar Repositories are open source and publicly available to everyone.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Data, metadata, documentation and code will be stored to owners' repositories. The Italian polar repository is meant to be a distributed infrastructure where data from other repositories will only be made accessible but not stored.

Have you explored appropriate arrangements with the identified repository?

Such arrangements will be made in the nearest future.

If there are restrictions on use, how will access be provided?

Datasets under restriction will be accessed submitting a request to the owners.

Is there a need for a data access committee?

The need for a data access committee will be evaluated in the nearest future.

Are there well described conditions for access (i.e. a machine readable license)?

Yes, data and metadata will provide machine readable licenses.

How will the identity of the person accessing the data be ascertained?

Access to restricted data will require an account for data download.

2.3. Making data interoperable

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

Yes, all data available in the Italian polar repositories will be interoperable and reusable. All the software used to implement the repository will be open source and re-combination with different datasets from different origins will be facilitated thanks to usage the ERDDAP data server that acts as a data broker.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Since we will use ERDDAP as a data server, data will be available in a large variety of formats to broaden the compatibility with other systems as much as possible. Particular focus will be placed to produce NetCDF files which is becoming a community standard for sharing scientific data. NetCDF files will be produce following the appropriate standard for each use case. Metadata will be available in the ISO 19115 format through the GeoNetwork metadata catalogue. Common vocabularies such as NASA Science Keywords and others will be used according to interoperability needs.

Will you be using standard vocabularies for all data types present in your data set, to allow interdisciplinary interoperability?

When applicable, standard vocabularies will be used.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Yes, mappings will be provided.

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest re-use possible?

The polar repository promotes free and open data sharing. Each dataset needs a license attached. The recommendation is to use Creative Commons attribution license for data (<https://creativecommons.org/licenses/by/3.0/> for details). Nevertheless, the PIs are able to apply different licensing.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

We will fix through the data policy a maximum value for the embargo period. We can imagine at three moment a 2 year period. However, this will also depend from the general rules ITINERIS will adopt as consortium. Inside this limit, the embargo period for the data will depend on the PIs. The repository will do its best to make the data available as soon as possible.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

The data can definitely be re-used by third parties after the end of the projects. There is no restriction on such usage.

How long is it intended that the data remains re-usable?

As long as they are needed.

Are data quality assurance processes described?

Data quality assurance processes will be described in its metadata or external documentation.

3. Allocation of resources

What are the costs for making data FAIR in your project?

This question arise from the scope of this template. In our context answer is more more articulated than normal. The NADC and IADC polar data repositories are, as mentioned, supported primarily by the respective PNRA and PRA programs which are programs that annually allocate resources for the maintenance of these systems. Occasionally, extraordinary PNRA and PRA resources or from projects such as ITINERIS can allow developments and upgrades of the system. Each funded project then has the obligation to contribute to bringing the data collected into these systems. Since there is therefore no overall budget, it is absolutely not possible to estimate a precise figure for FAIRNESS. However, we can estimate that a percentage of around 5-7% of all allocated resources are necessary to keep the systems efficient and allow a regular flow of data collected towards them. For example, in the PRA on a budget of 1 million euros/year, around 40 million euros are typically allocated to

data management, and this is net of the resources individual funded projects must allocate to make their data available to IADC on a FAIR basis.

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Also in this case the question is formulated specifically for EU projects. But wanting to give an answer to the general request of "how do you support yourself", the answer is already reported in the answer to the question above: with the ordinary resources of PNRA and PRA, with the resources that the individual funded projects must commit to make FAIR i their data, with extraordinary resources from PNRA, PRA other sources when it comes to thinking about strong upgrades and advancements of the system.

Who will be responsible for data management in your project?

The Italian Institute of Polar Sciences (CNR-ISP) will be responsible for the data management.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

In the current situation there is no overview of the costs of long-term preservation of data. This information will be updated in further versions of the DMP.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Encrypted HTTPS connections will be used and private data will be available under request or registration/login mechanism. All data will be backed up and replicated on separate server.

Is the data safely stored in certified repositories for long term preservation and curation?

Yes.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Once again the question is very much aimed at the context of a European project. In any case, the ethical question regarding data is a question that has a general aspect and is very important. Both NADC and IADC take and will take into account EITCI aspects, as required by the international bodies and agreements to which Italy participates/adheres. For atmospheric data in general there are no problems, except when we are talking about parameters that can have effects on human health (e.g. erythemal dose, concentrations of minor gases and contaminants). For the marine part there are also no problems for oceanographic data, while some limitations could affect biological data.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Legal notices including privacy policy for personal data use will be publicly available on the repository website.

6. Other issues

*Do you make use of other national/funder/sectorial/departmental procedures for data management?
If yes, which ones?*

Polar Data Repositories will be a well integrated system, with the only obligation to respond to national and international standards and best practices.