Integrating Activities for Advanced Communities

D4.1- Data Management Plan

Project No.730938– INTERACT

H2020-INFRAIA-2016-2017/H2020-INFRAIA-2016-1

Start date of project: 2016/10/01 Duration: 48 months
Due date of deliverable: 2017/03/31 Actual Submission date: 2017/07/07

Lead partner for deliverable: METNO

1. Author: Øystein Godøy, Boris Radosavljevic, Boris Biskaborn

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the Consortium (including the Commission Services)	
CO	Confidential, only for members of the Consortium (including the Commission Services)	

Table of Contents

Publishable Executive Summary

The main objective of the International Network for Terrestrial Research and Monitoring in the Arctic (INTERACT) is building capacity for identifying, understanding, predicting and responding to diverse environmental changes throughout the high latitude and altitude regions in the northern hemisphere. For this purpose, INTERACT stations collect various types of heterogeneous data that are largely not available to the public. The purpose of the **Data Management Plan** is to describe the data that will be created and how it will be shared and preserved. The goal is a unified view[1] of INTERACT data that will improve the impact of INTERACT and individual stations.

The basic principles of INTERACT data management is that INTERACT is following a **metadata driven approach.** This implies that a number of physically distributed data centres are connected through interoperability interfaces that provide machine access to metadata and data. For this to work, INTERACT datasets are described using standardised discovery metadata. This shall ensure the data are archived and re-usable for future generations and relevant to technologically driven data analysis developments. INTERACT shall rely on discipline specific efforts to establish interoperability at the data level. INTERACT **promotes free and open access** to data in line with the European Open Research Data Pilot (OpenAIRE). Furthermore, this plan is a living document that will be updated during the project and is based upon the template provided by the Digital Curation Centre.

According to this data management plan, INTERACT:

- shall facilitate interoperability by following international best practises for data management
- shall document datasets using ISO19115 where text elements are populated using controlled vocabularies available in machine readable form (GCMD Vocabularies are default)
- metadata and data products shall be free and open (Creative Commons attribution license), although some data may have temporal restrictions
- shall use self explaining file formats/data encoding
- shall use the NetCDF format following the Climate and Forecast Convention where possible
- shall make data available in a timely fashion
- data shall be archived in repositories with a long term mandate
- promotes and encourages the implementation of globally resolvable Persistent Identifiers (g.e. Digital Object Identifiers) at each contributing data centre
- responsible data centres are storing and maintaining data and metadata will be stored and maintained by responsible data centres, with the metadata harvested in the central node
- contributing data centres must support discovery metadata through OAI-PMH and serve ISO19115 metadata (with controlled vocabularies)
- data access through OPeNDAP is preferred for combination and segmentation of datasets
- shall establish a central node providing unified data discovery at http://interact.met.no
- legacy data shall be identified and future handling planned

To accomplish these goals, a Data Forum will be established. This forum will include representatives from data providers (e.g. station managers, INTERACT GIS and INTERAccess) and data repositories. It will meet annually to improve awareness of INTERACT data among repositories, create awareness of international repositories among station managers, and to agree on appropriate archives for project metadata and "environmental data". In addition, representatives of international networks such as ITEX, CALM, ILTER etc. will be invited to coordinate standardisation efforts and liaison with data providers and repositories. Through connection with established and future networks, and by following international documentation and exchange standards, INTERACT stations will ensure that the data generated are relevant, useful and reachable for both Arctic and non-Arctic users.

[1] All available INTERACT data are made searcable in the same search interface.

1. Introduction

1.1. Background and motivation

INTERACT research stations generate data as a result of long-term environmental monitoring programmes and shorter term research projects. Currently more than 75 research stations located throughout arctic and northern alpine areas are part of the INTERACT network (Figure 1). Among the scientific disciplines practiced in the network are climatology, geoscience, biology, ecology, cryospheric science, and to some extent anthropology. These activities can be organised by the station itself or by external scientists. In addition, research stations often archive relevant data from external sources (usually meteorological observations, photos, reports, maps). Such heterogeneous data generating activities, combined with lacking structured data management practices at the stations, result in data archived at multiple locations for individual stations. Current INTERACT data repositories include research stations' archives, local archives (e.g. municipal authorities), national archives (e.g. meteorological institutes), archives of international, single discipline networks (e.g. CALM), EU repositories (e.g. SIOS Knowledge Centre), pan-Arctic/regional repositories (e.g. SAON), and global repositories (e.g. Pangaea, GTN-P). Far too often, research project data stays with the research project leader and is not shared according to SAON/IASC, EU, OECD, WMO and GEOSS recommendations. Furthermore, most stations lack the interoperability interfaces necessary to actively engage in national and international data exchange and management activities coordinated through international programmes (e.g. EU, WMO, ICSU, GEO etc.). Also, no unified interface for INTERACT datasets curretly exists that could help INTERACT achieve a domain data repository role. The consequence is underutilisation of existing and future monitoring capabilities, as well as INTERACT as a contribution to the scientific toolbox. However, through the application of accepted documentation and exchange standards, INTERACT can become a valuable asset in network gap analysis performed in various communities (e.g. WMO Observation Systems Capability Analysis and Review tool (OSCAR) surface supporting GAW and GCW). The improvements to data management practices would make INTERACT data **FAIR**: findable, accessible, interoperable, and re-useable.

The data management work package of INTERACT aims to **increase the data interoperability** among stations and towards external data consumers by defining needs for generating **common standards** and data dissemination strategies. The benefit of such a process is increased visibility and potentially impact for INTERACT stations.

The purpose of the data management plan is to describe the basic principles how the data generated by the project is handled during and after the project. This includes standards and generation of discovery and use metadata, data sharing and preservation and life cycle management; i.e. by following the principles outlined by the Open Research Data Pilot and The FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al. 2016). However, INTERACT is a heterogeneous community and full implementation of data management at stations is not in the budget. Thus the primary objectives of this Data Management Plan is to initiate a process that at some time will lead to a unified view of the INTERACT data. This is achieved through dialogue with station managers, description of best practises and linking stations and data centres where stations do not want to manage data themselves.

This document is a **living document** that will be updated during the project.

1.2. Organisation of the plan

This plan is based on the template provided by the UK Digital Curation Centre (DMP Online). This approach is recommended by OpenAIRE guidelines.

2. Admin details

Project Name	INTERACT
Funding	EU HORIZON 2020 Research and Innovation Programme
Partners	• Lund University (LU) SE • University of Sheffield (USFD) UK • University of Copenhagen (UCPH) DK • University of Oulu (UOULU) FI • Aarhus University (AU) DK • CLU srl (CLU) IT • Alfred Wegener Institute for Polar and Marine Research (AWI) DE • Norwegian Polar Institute (NPI) NO • Natural Environment Research Council (NERC) UK • Tomsk State University (TSU) RU • University of South Bohemia in Ceske Budejovice (USB) CZ • Swedish Polar Research Secretariat (SPRS) SE • Norwegian Institute for Agricultural and Environ. Research (NIBIO) NO • Stockholm University (SU) SE • University of Helsinki (UH) FI • Greenland Institute of Natural Resources (GINR) GL • Polish Academy of Sciences - Geophysics dept. IGF-PAS PL • University of Turku (UTU) FI • University of Oslo (UiO) NO • Natural Resources Institute Finland (LUKE) FI • Russian Academy of Sciences - Siberian Branch (IBPC) RU • M V Lomonosov Moscow State University (MSU) RU • Swedish University of Agricultural Sciences (SLU) SE • Zentralanstalt für Meteorologie und Geodynamik (ZAMG) AT • University of Innsbruck (LFU) AT • Yugra State University (YSU) RU • Faroe Islands Nature Investigation (JF) FO • Northeast Iceland Nature Center (RFS) IS • Centre for Northern Studies (CEN) CA • Polish Academy of Sciences - Geography Dept. (IGSO-PAS) PL • Consiglio Nazionale delle Ricerche (CNR) IT • University of Alaska Fairbanks (UAF) US • Sudurnes Science and Learning Center (SSLC) IS • Finnish Meteorological Institute (FMI) FI • CAFF International Secretariat (CAFF) IS • APECS - University of Tromsoe (UiT) NO • Aurora College - The Western Arctic Research Centre (AC) CA • Arctic Institute of North America (AINA) CA • Umbilical Design (UD-AB) SE • ÅF Technology AB (AF) SE • Norwegian Meteorological Institute (METNO) NO

	• Agricultural University of Iceland (AUI) IS • University of Groningen (UoG-AC) NL • International Polar Foundation (IPF) BE • Mapillary (MAP) SE • University Centre in Svalbard (UNIS) NO • The International Centre for Reindeer Husbandry (ICR) NO

3. Data summary

The INTERACT Data Management Plan addresses data describing >75 research stations (Figure 1) in cold regions of the Northern Hemisphere. A listing of the stations involved is provided in the proposal Section 4 and on the projects website (http://www.eu-interact.org). These stations obtain baseline- and monitoring data on a multitude of scientific disciplines practiced within the network. Through the integration of the independent research stations' data through a unified approach, a comprehensive coordinated view on the Arctic is achieved. Multitudes of stakeholders, scientists, modellers, government agencies, educators, and to some extent private citizens have a vested interest in accessing the various kinds of data collected at the stations that can provide historical records, serve in model validation, and provide critical indicators across the disciplines covered within the network.

The main objective of INTERACT is **to build capacity for identifying, understanding, predicting and responding to diverse environmental changes throughout the wide environmental and land-use envelopes of the Arctic**.

A prerequisite to achieve this is to coordinate the data collected at INTERACT stations and to make them available. Thus, INTERACT data management aims to integrate datasets in a unified system, simplifying discovery, access and utilisation of data for various stakeholders in the scientific community, as well as in operational communities (e.g. scientists, national and local decision makers, etc.).

INTERACT is truly interdisciplinary. With this perspective and as this activity on coordinated data management has just begun, no full overview of data types exist.

Concerning the encoding of data, self-explaining file formats (e.g. NetCDF, HDF/HDF5) combined with semantic and structural standards like the Climate and Forecast Convention are required to ensure interoperability at the data level. Implementation of this is however a time consuming process and will be done gradually.

Eventually, data can be integrated from different data centres with this approach.

The default format for INTERACT datasets is NetCDF following the Climate and Forecast Convention (feature types grid, timeseries, profiles and trajectories if applicable). However, not all data handled at INTERACT stations are covered by the Climate and Forecast Convention for standard names. INTERACT is currently in a process of analysing the data collected and potential ways for handling these data. This work must be based on external activities within the disciplines and in Arctic data management in general.

INTERACT has a huge legacy of data. Within this phase of INTERACT, an effort to identify legacy datasets and plan future handling of these will be initiated.

Data are generated by permanent instrumentation (monitoring) and field work (projects) at the INTERACT research stations.

The total amount of data is yet not known in detail currently. As the project progress, better understanding of the full capacity of INTERACT will be achieved.

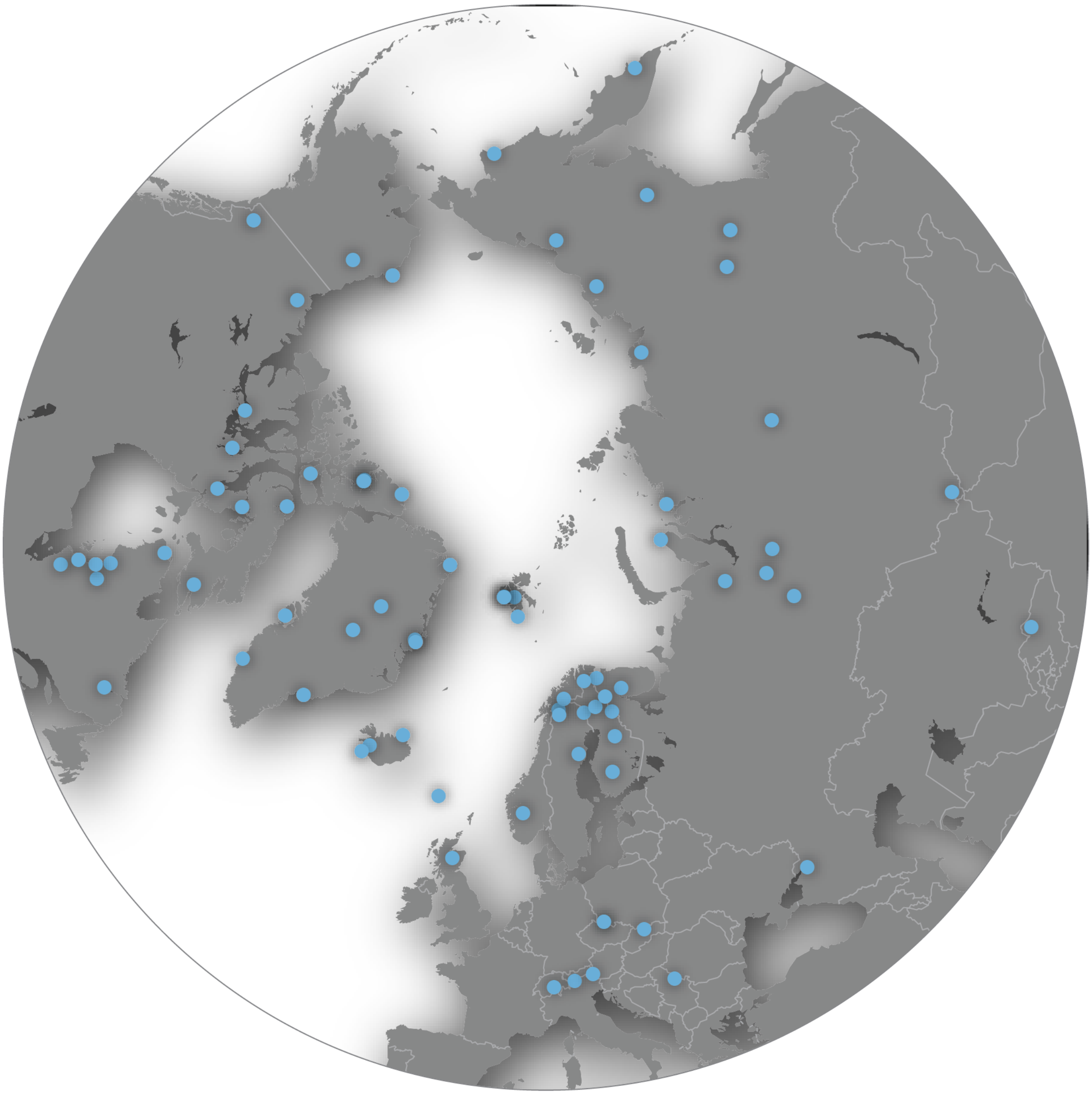

Figure 1: More than 75 research stations are participating in INTERACT

INTERACT data are useful for all users of INTERACT research stations, as well as projects, programmes and individual scientists undertaking scientific or monitoring work in the Arctic. Establishing a unified view on the data produced by INTERACT stations will improve the impact of INTERACT and the individual stations through promotion of the their capacity for various data consumers, ranging from individual scientists to regional or global monitoring programmes (e.g. AMAP, GCW and GAW).

3.1. Making data findable, provisions for metadata [FAIR data]

Improving the ability of internal and external data consumers to find and understand the data INTERACT stations are producing is essential to increase the impact of INTERACT, individual stations and researchers. Through exposure of the data produced by INTERACT in relevant discipline specific, regional and global catalogues, the knowledge and interest in INTERACT is increased. This can be done both individually by each station or by the INTERACT community.

INTERACT is following a metadata driven approach. This means that by utilizing internationally accepted standards and protocols for documentation and exchange of discovery and use metadata, interoperability with international systems and frameworks, including WMO's systems, Year of Polar Prediction (YOPP), WMO Global Cryosphere Watch (GCW) and many national and international Arctic and marine data centers (e.g. Svalbard Integrated Arctic Earth Observing System) is ensured.

INTERACT data management is distributed in nature, relying on a number of data centres with a long term mandate. This ensures preservation of the scientific legacy. While defining the approach of INTERACT data management, INTERACT is aligning efforts with SAON/IASC Arctic Data Committee. This implies documenting all datasets with standardised discovery metadata using either the Global Change Master Directory Directory Interchange Format or ISO19115 standards.

INTERACT promotes and encourages the implementation of globally resolvable Persistent Identifiers (e.g. Digital Object Identifiers - DOI) at each contributing data centre. Some have this in place, while others are in the process of establishing this. If DOIs are not supported, a local persistent identifier must be supported.

Concerning naming conventions, INTERACT requires that controlled vocabularies are used both at the discovery level and the data level to describe the content. Discovery level metadata must identify the convention used and the convention has to be available in machine readable form (preferably through Simple Knowledge Organisation System). The fallback solution for controlled vocabularies is the Global Change Master Directory vocabularies.

The search model of the data management system is based on GCMD Science Keywords for parameter identification through discovery metadata.

Versioning of data is required for the data published in the data management system. Details on requirements for how to define a new version of a dataset is to be agreed upon by the Data Forum.

The central node can consume and expose discovery metadata as GCMD DIF and ISO19115 records (using GCMD keywords for physical/dynamical parameters). Support for more formats is considered. For use metadata the Climate and Forecast convention is promoted. More specifications will be identified early in the project.

3.2. Making data openly accessible [FAIR data]

Being able to find relevant data is only the first step. Most data consumers are interested in the actual data. The requirements of data consumers vary. While ad hoc consumers (usually scientists) frequently consume whatever is found from a network of stations, consumers concerned with monitoring, or calibration and validation of numerical models, or remote sensing products will usually require harmonisation of the data to a common form before they invest in integration. In order to address this standardisation of file formats (encoding) and data access mechanisms is required.

The discovery metadata that can be collected will be made available through a web based search interface at https://interact.met.no. Some data may have temporal access restrictions (embargo period). An embargo period on data may be requested for different reasons, e.g. allowing Ph.D. students to prepare their work, or while data is used in the preparation of a publication. Even if data are constrained in the embargo period, data will be shared internally in the project. Any disagreements on access to data or misuse of data internally are to be settled by the INTERACT Steering Board.

A central data repository supporting the demonstrator will be made available. Within this demonstrator, data are made accessible using interoperability protocols using a THREDDS Data Server. This will support OpeNDAP, OGC Web Map Service for visualisation of gridded datasets, and direct HTTP download of full files. Standardisation of data access interfaces and linkage to the Common Data Model through OPeNDAP[2] is promoted for all data centres contributing to INTERACT. This enables direct access of data within analysis tools like Matlab, Excel[3] and R. The purpose of this demonstrator is to show how data may be shared in standard manner using Open Source Software. Most of the INTERACT data will however be managed by the stations or data centres the stations make agreements with. The purpose of the demonstrator is to increase the knowledge among stations on metadata and data interoperability and to encourage stations not sharing data today to start exploring possibilities.

Metadata and data for the datasets are maintained by the stations and responsible data centres, metadata supporting unified search is harvested and ingested in the demonstrator hosted by central node.

3.3. Making data interoperable [FAIR data]

Interoperability at the data level will be facilitated by following best practises within international data management and relevant standardisation efforts. This includes application of self explaining file formats utilising discipline specific controlled vocabularies for data annotation. Data will be made available through OPeNDAP with use metadata following the Climate and Forecast conventions e.g. for geophysical data. However, exceptions will occur due to the diversity of INTERACT data. Some of the disciplines covered by INTERACT, e.g. meteorology, are advanced in the context of use metadata, while others are lacking a unified, discipline specific approach. INTERACT must rely on discipline specific activities and larger network activities (e.g. GTN-P, GAW, GCW) to avoid duplication of efforts and reuse the solutions developed. In order to address this aspect, the Data Forum is established to promote the understanding of emerging data management requirements. Implementation within INTERACT will be a long and stepwise process.

Initially GCMD Science keywords will be used, mapping between GCMD Science keywords and CF standard names is supported (but needs to be updated). Other vocabularies are included (e.g. GBIF) as they are available and considered mature. In the current situation, interaction with the stations is needed to fully get an overview of the relevant standards and controlled vocabularies.

3.4. Increase data re-use (through clarifying licenses) [FAIR data]

The INTERACT data policy is not written yet, but INTERACT promotes free and open data sharing in line with the Open Research Data Pilot. Each dataset requires a license attached. The recommendation in INTERACT is to use Creative Commons attribution license for data (see https://creativecommons.org/licenses/by/3.0/ for details). However, INTERACT is spanning many nations and a more careful examination of the business models for various stations and funding regimes is required.

INTERACT data should be delivered in a timely manner, meaning without undue delay. Any delay, due or undue, shall not be longer than one year after the dataset is finished. Discovery metadata shall be delivered immediately.

INTERACT is promoting free and open access to data. Some data may have access constraints. Details will be evaluated during the project.

The quality of each dataset is the responsibility of the Principal Investigator. The Data Management System will ensure that information on the quality of the data is available in the discovery metadata.

INTERACT is primarily concerned with observational data. These data cannot be reproduced and must be reusable in the undefined future.

[2] http://apievangelist.com/2014/12/05/history-of-apis-noaa-apis-have-been-restful-for-over-20-years/

[3] https://www.opendap.org/support/faq/general/use-spreadsheet

4. Allocation of resources

In the current situation it is not possible to estimate the cost for making INTERACT data FAIR. Part of the reason is that this work is relying on existing functionality at the contributing data centres and that this functionality has been developed over years. The project is also still in the process of establishing an overview of the current situation among the 79 research stations involved.

Within the first period of INTERACT, a questionnaire has been filed to stations asking for details on existing data management. This is still being analysed, but preliminary results indicate challenges establishing a preliminary data management system as a demonstrator for INTERACT. Over 50 % of the stations surveyed indicated established data management routines. Thus, instead of starting with stations, INTERACT will start with selected data centres that host data for INTERACT stations. Most of these data centres are active in relevant data management activities.

The following data centres are so far identified:

Data centre	Contact	Comment
AWI/Pangaea	Boris Biskaborn and Boris Radosavljevic	PANGAEA stores a high number of datasets from various stations and all data from Samoylov station. INTERACT metadata will be published and made accessible in PANGAEA, parallel to publication in ESSD
Italian Arctic Data Centre	Angelo Viola	Handling data from: • CNR station Ny-Ålesund
Norwegian Meteorological Institute/Arctic Data Centre	Øystein Godøy	A subsystem[4] will be used to integrate information from the data centres maintaining INTERACT data. Through this the central hub for INTERACT data will be established.
NordicanaD[5]	Luc Cournoyer Christine Barnard	Handling data from: • CEN WK
Norwegian Polar Institute/Norwegian Polar Data Centre	Stein Tronstad	Handling data from: • Sverdrup station in Ny-Ålesund
British Antarctic Survey/UK Polar Data Centre		Handling data from: • NERC Arctic Research Station (Ny-Ålesund) • ECN Cairngorms
NSF Arctic Data Center		Handling data from: • Toolik Research Station
AVAA, Finland		Handling data from: • Värriö Research Station
Research Data Descriptions Discovery Service of Natural Resources Institute Finland		Handling data from: • Kainuu Fisheries Research Station • Kolari Research Station

Not all contact points identified above are directly involved in INTERACT, but their institutions are and the data centres are handling INTERACT data. For some archives, contact points are to be identified. This table will be further developed and is only to be considered as a preliminary version.

[4] Will be available August 2017.
[5] Data available through Polar Data Catalogoue which has interoperability interfaces for metadata.

Once INTERACT data management is fully established, each data centre is responsible for accepting, managing, sharing and preserving relevant datasets. Concerning interoperability interfaces the following interfaces are required for the first version of the system:

1. Metadata
    1. OAI-PMH serving either CCMD DIF or the ISO19115 minimum profile with GCMD Science Keywords.
2. Data (will also use whatever is available and deliver this in original form, for those data no synthesis products are possible without an extensive effort)
    1. OGC WMS (actual visual representation, not data)
    2. OPeNDAP for data streaming/download, including format conversion

However, it should be understood that this is a best effort basis to show the benefit for the INTERACT community, at least initially. Thus, the activities are aligned with the efforts of the SAON/IASC Arctic Data Committee.

In the current situation there is no overview of the costs of long term preservation of data as this is the responsibility of the contributing data centres and the business model for these differs. This information will be updated.

5. Data security

Data security relies on the existing mechanisms of the contributing data centres. INTERACT recommends to ensure the communication between data centres and users with secure HTTP. Concerning the internal security of the data centre, INTERACT recommends the best practises from OAIS.

The central node relies on secure HTTP, but not all contributing data centres support this yet. As this effort is for demonstration initially, this section will be addressed following discussions in the Data Forum.

6. Ethical aspects

INTERACT is handling a wide variety of data. Some data may be ethically sensitive. In the IASC context this is especially related to humans and resources (e.g. fisheries, birds and mammals). As the INTERACT Data Policy still is under development, this will be addressed in later versions of the document.

7. Other

This is not applicable in the current situation, but other considerations (e.g. funder, institutional, departmental or group policies on data management, data sharing and data security) may become applicable in later versions of the plan.