EU grant agreement No. 773421

Deliverable No. 10.3

Nunataryuk Data Management Plan

Project No. 773421

Version 1, April 2018

Start date of project: 2017/11/01

Due date of deliverable: 2018/04/30

Lead partner of deliverable: AWI

Authors: Anna Irrgang (AWI), Alexander Smirnov (AP)

Submission of Deliverables

Work Package	WP10
Deliverable No	10.3
Deliverable title	Nunataryuk Data and Knowledge Management
Version	1
Status	Draft
Dissemination level	PU – Public
Lead Beneficiary	AWI
Contributors	Arctic Portal (AP)
	Alfred Wegener Institute (AWI)
Due Date	30 April 2018
Delivery Date	30 April 2018

This project has received funding from the European Union's Horizon 2020 Research & Innovation programme under grant agreement No. 773421.

Table of Contents

2 EXECUTIVE SUMMARY

The main goal of the Nunataryuk project is to determine the impacts of thawing land, coast and subsea permafrost on the global climate and on humans in the Arctic and to develop targeted and co-designed adaptation and mitigation strategies. For this purpose, a high diversity of data will be collected and produced within different work packages. The purpose of the Data Management Plan is to describe the data that will be created and to present a concept of how the data will be shared and preserved. The goal of Nunataryuk data management is to create a web-based Nunataryuk data portal which provides a unified search interface to all gathered data sets generated within the project in order to maximize the visibility and impact of data generated within the Nunataryuk project.

This Data Management Plan (DMP) is based on the H2020 FAIR Data Management Plan template designed to be applicable to any H2020 project that produces, collects or processes research data. The purpose of the DMP is to describe the data that will be produced, collected or processed during the project, as well as the plans for data sharing and data preservation.

Nunataryuk follows a metadata-driven approach where a physically distributed number of data repositories are integrated using standardized discovery metadata and interoperability interfaces for metadata and data storage and publication. The Nunataryuk data portal will provide a unified search interface to all gathered data sets. Further, Nunataryuk will host a data management system directly coupled to the Global Terrestrial Network for Permafrost (GTN-P), where many of the data generated in the project will be stored. Nunataryuk promotes free and open access to data in line with the European Open research Data Pilot (OpenAIRE). Within this plan an overview of the data collection procedures is provided as well as an initial outline of dissemination.

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

3 INTRODUCTION

3.1 Background and motivation

Within the Nunataryuk project, a vast amount and diversity of data will be produced. The purpose of the DMP is to document how the data generated within the project is handled during and after the project. It describes the basic principles for data management within the project. This includes standards and generation of discovery and use metadata, data sharing and preservation and life cycle management. This DMP is a living document that will be updated during the project in time with the periodic reports. Nunataryuk is following the principles outlined by the Open Research Data Pilot (OpenAIRE) and The FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al. 2016[1]).

3.2 Organization of the plan

This DMP is based on the H2020 FAIR Data Management Plan template[2] designed to be applicable to any H2020 project that produces, collects or processes research data. This is the same plan as OpenAIRE is referring to in their guidance material.

[1]Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).
[2]http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

4 Administration details

Project Name: Nunataryuk

Funding: EU HORIZON 2020 Research and Innovation Programme

Partners:

- Alfred Wegener Institute Helmholtz Center for Polar and Marine Research (Germany)
- Stockholms Universitet (Sweden)
- VU University Amsterdam (Netherlands)
- Le Centre National de la Recherche Scientifique (France)
- Université Laval (Canada)
- Max Planck Institute for Meteorology, Hamburg (Germany)
- University of Oulu, (Finland)
- Technical University of Denmark (Denmark)
- NORDREGIO (Sweden)
- Stefansson Arctic Institute (Iceland)
- University of Vienna (Austria)
- B•GEOS (Austria)
- Consiglio Nazionale delle Ricerche (Italy)
- University of Oslo (Norway)
- University of Lisbon (Portugal)
- The International Institute for Applied Systems Analysis (Austria)
- University of Hamburg (Germany)
- Université libre de Bruxelles (Belgium)
- Norwegian University of Science and Technology (Norway)
- University of Versailly saint-Quentin en Yvelines (France)
- Grid Arendal (Norway)
- Natural Resources Canada - Geological Survey of Canada (Canada)
- INFORMUS GmbH (Germany)
- ACRI-He (France)
- Universite Pierre Et Marie Curie (France)
- Helmholtz Zentrum Potsdam Deutsches Geoforschungszentrum (Germany)
- Kommune Kujalleq (Greenland)

- Arctic Portal (Iceland)

5 Data summary

The primary goal of Nunataryuk is

to investigate the impacts of thawing coastal and subsea permafrost on the global climate, and develop targeted and co-designed adaptation and mitigation strategies for the Arctic coastal population. Nunataryuk brings together world-leading specialists in natural science and socio-economics to:

- *develop a quantitative understanding of the fluxes and fates of organic matter released from thawing coastal and subsea permafrost*
- *assess which risks are posed by thawing coastal permafrost to infrastructure, indigenous and local communities and peoples' health, and from pollution*
- *use this understanding to estimate the long-term impacts of permafrost thaw on global climate and the economy*.

Therefore, a number of datasets will be generated. These will include:

- Datasets to quantify thawing permafrost and its impact on storage and vulnerability of organic matter and contaminants on land
- Datasets of lateral fluxes of organic matter from coastal erosion and watersheds draining into the Arctic Ocean
- Datasets on the quantitative constraints for the vulnerable subsea permafrost system
- Datasets on quantified trends in the signature of organic matter fluxes in coastal waters
- Datasets on health and pollution risks associated with permafrost thaw for wildlife and humans living in the coastal Arctic
- Datasets on the quantified effect of permafrost thaw on Arctic infrastructure by means of site investigations and local-scale modelling

5.1 Data overview

In order to get an overview over the variety and amount of data which are aimed to be generated during the Nunataryuk project, a short data survey was carried out among the Principal Investigators for ach work package in March 2017. The chapters 5.1.1 and 5.1.2 represent some of the survey results.

5.1.1 Types and formats of data generated/collected

Nunataryuk will generate a variety of data which will include:

1. Geospatial data:
    - Lateral carbon fluxes
    - Ocean color
    - Temperatures of soil and snow
    - Model output
    - Organic carbon in the Arctic Ocean shelf sediments
    - Elevations, distances, coordinates (GPS measurements)
    - Airborne LiDAR, Hyperspectral measurements
    - UAV surveys
    - Point clouds, thematic maps, coastline, vegetation, geomorphology
2. Multimedia data:
    - Videos
    - Photos
    - Audio recordings
    - Science blogs
3. Empirical data:
    - Interview recordings, transcripts, field notes
    - Socioeconomic data on Arctic coastal settlements
4. Field measurements:
    - Hydrological observations
    - Aquatic carbon samples
    - pH measurements
    - Optical and biogeochemical data

- Data on physical properties and temperatures of soil, snow, and hydrodynamic conditions
- Soil temperature measurements
- Near surface geophysical data
- CH_4 concentration measurements

5. Laboratory experiments:
   - Post-processed aquatic carbon samples (e.g. isotope information)
   - Optical and biogeochemical data
   - Physical properties of soils
   - Sediment and subsea permafrost organic carbon properties
   - Triple-isotope analyses of CH_4
   - Soil Carbon, Nitrogen and their isotopes
   - Soil Organic Matter quality data (MS data)
   - Microbial activity
   - Inorganic contaminants
   - Thermal and salinity experiments for thermal model validation
   - Lab-scale coastal erosion experiment (soil and water temperature, wave heights and frequency, imagery for DEM creation, measurements of mechanical erosion)
   - Experimental data (soil laboratory incubations)
6. Other:
   - Data and information from literature (incl. grey literature like white papers, government documents, newspaper articles, etc.)

The main data formats are expected to be:

- Microsoft Excel (XLS)
- Shapefile (SHP)
- Comma-separated values (CSV)
- Text (TXT)
- NetCDF (NC)
- MPEG-4 video format (MP4)
- GeoTIFF (TIF)
- Extensible Markup Language (XML)

- Microsoft Word (DOC)
- Joint Photographic Experts Group (JPG)
- Portable Network Graphics (PNG)
- MPEG 2.5 audio format (MP3)
- Portable Document Format (PDF)
- Hierarchical Data Format (HDF5)
- Raw data (RXP, RAW)
- Various proprietary binary formats (BIN)

A high variety of data formats will be generated and worked with in the Nunataryuk project (Fig. 1). The most popular formats are expected to be XLS (16% of total amount of datasets), SHP (13%), CSV (12%), TXT (12%) and netCDF (5%).

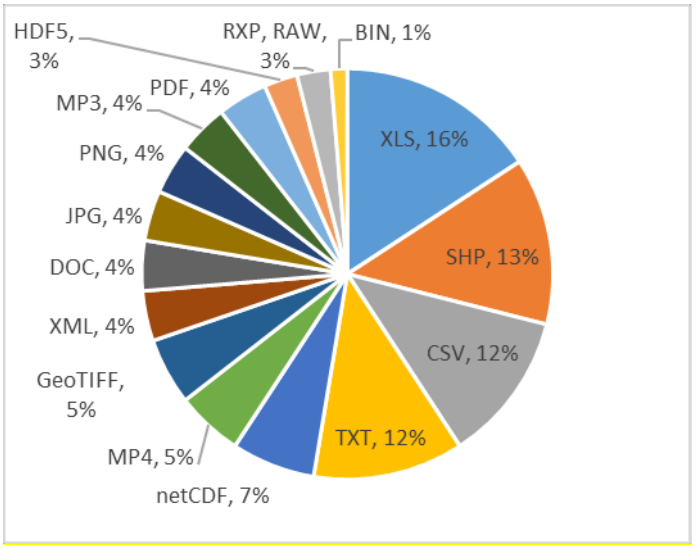

Figure 1. Data formats generated in the project

5.1.2 Origin of the data

The majority of the data used in the project will be generated within the project (81.8 %, Fig. 2). However, some data will be reused in accordance with the FAIR data re-use policy (18.2%).

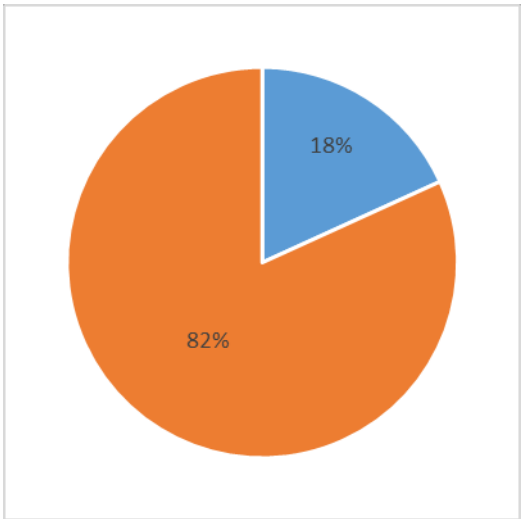

Figure 2. Distribution of data origin.

Orange field marks newly generated data, blue field marks reused data.

It is estimated that approximately 10 Terabytes of data will be generated within the Nunataryuk project.

A major goal of the Nunataryuk data management is to make data generated within the project visible and useful for regional and global monitoring programs, Arctic researchers, Arctic communities and individuals. Therefore, a Nunataryuk data portal will be established, which will provide a unified view on the data produced by the Nunataryuk project. This approach is essential in order to increase the visibility of the Nunataryuk project and benefit from the generated data.

6 FAIR data

6.1 Making data findable, including provisions for metadata [FAIR data]

Nunataryuk is following a metadata driven approach, utilizing internationally accepted standards and protocols for documentation and exchange of discovery and use metadata. This ensures interoperability at the discovery level within international systems and frameworks. The Nunataryuk project will host a data management system directly coupled

to the Global Terrestrial Network for Permafrost (GTN-P). Existing and new datasets will be documented in a standardized manner for data discovery and delivered through the data management system. GTN-P is part of the Global Climate Observing System (GCOS), of the Global Cryosphere Watch (GCW) and the World Data System (WDS) and complies with all geoinformation standards. Part of the Nunataryuk data will therefore naturally feed into the Global Earth Observation System of Systems (GEOSS) Common Infrastructure.

Nunataryuk promotes the implementation of Persistent Identifiers at each contributing data center. Some have this in place, while others are in the process of establishing this. Although application of globally resolvable Persistent Identifiers (e.g. Digital Object Identifiers) is not required, it is promoted by the Nunataryuk data management. Concerning naming conventions, Nunataryuk 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 format. The fallback solution for controlled vocabularies is the Global Change Master Directory (GCMD) vocabularies. The search model of the data management system is based on GCMD Science Keywords for parameter identification through discovery metadata. The Nunataryuk data management system can consume and expose discovery metadata provided in ISO19115. GCMD keywords must be used to describe physical and dynamical parameters.

6.2 Making data openly accessible [FAIR data]

Nunataryuk will participate in the Pilot on Open Research Data in Horizon 2020 (OpenAIRE). All discovery metadata will be available through a web-based search interface available through the central project website (www.Nunataryuk.org). Some data may have temporal access restrictions (embargo period). These will be handled accordingly. Valid reasons for an embargo period on data are primarily for educational reasons, allowing Ph.D. students to prepare and publish their work (2-3 years) and for publishing research papers (around 1 year). Even if data will be 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 Nunataryuk Executive Board. Data will be openly accessible by the data management system directly coupled to the Global Terrestrial Network for Permafrost (GTN-P).

Most of the datasets produced by the project will also be stored in the data repository PANGAEA (https://pangaea.de/). PANGAEA is a data publisher for Earth and environmental science and hosted jointly by the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (AWI) and the Center for Marine Environmental Science (MARUM) at the University of Bremen. PANGAEA provides long-term archiving of data, data publication and dissemination, as well as scientific data management. One major advantage of PANGEA is that it provides each dataset with a bibliographic citation and a Digital Object Identifier (DOI) allowing it to be identified, shared, published and cited. Is is expected that more than 45 % of the project's data will be provided with a DOI identifier, 23 % will not and the remaining 32 % are yet undefined.

6.3 Making data interoperable [FAIR data]

In order to be able to reuse data, standardization is important. This implies both standardization of the encoding/documentation, as well as the interfaces to the data. Further up in the document, it is referred to documentation standards widely used by the scientific communities. This includes encoding gridded data output as NetCDF files, following the Climate and Forecast convention or the WMO GRIB format. NetCDF files following the CF convention are self-describing and interoperable. Application of the CF conventions implies requirements on the structure and semantic annotation of data (e.g. through identification of variables/parameters through CF standard names). Irregular data will be mostly encoded in XLS, CSV and TXT formats which are open and convenient in terms of data interoperability.

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

Nunataryuk promotes free and open data sharing in line with the Open Research Data Pilot (OpenAIRE). Each dataset needs a license attached. The recommendation in Nunataryuk is to use Creative Commons attribution license for data (ee https://creativecommons.org/licenses/by/3.0/ for details). Nunataryuk data should be delivered in a timely manner meaning without un-due delay. Any delay, due or un-due, shall not be longer than one year after the dataset is finished. Discovery metadata shall be delivered immediately. Nunataryuk is promoting free and open access to data. Some data may have constraints (e.g. on access or dissemination) and may be exclusively available for

project participants. Details will be evaluated during the project. The quality and information about the quality of each dataset is the responsibility of the Principal Investigator.

7 Allocation of resources

In the current situation it is not possible to estimate the cost for making Nunataryuk data FAIR. Part of the reason is that this work is relying on existing functionality at the contributing data centers and that this functionality has been developed over years. The cost of preparing the data in accordance with the specifications and initial sharing is covered by the project. Maintenance of this over time is covered by the business models of the data centers. 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 centers and the business model for these differs. This information will be updated in further versions of the DMP.

8 Data security

Data security relies on the existing mechanisms of the contributing data centers. Nunataryuk recommends ensuring the communication between the data management system and users with secure HTTP. Concerning the internal security, Nunataryuk recommends the best practices from the Open Archival information System (OAIS). The technical solution will vary between data centers, but most data centers have solutions using automated check sums and replication.

9 Ethical aspects

Nunataryuk is transdisciplinary in nature, involving a large socio-economic component, and intrinsically user and stakeholder driven. Therefore, ethical principles are considered central in the design of the project and their proper processing will be essential for the successful implementation and dissemination of the project. While the project will not raise any highly sensitive data, the involvement of human beings and collection of personal data has a potential to raise general ethical concerns. In order to minimize the potential, the project will meet the highest established ethical standards in science and research and follow, inter

alia, the European Charter for Researchers, The Code of Conduct for the Recruitment of Researchers and EC regulations personal data processing. The research groups involved will follow, where applicable and available, international and national ethical research principles, national legal regulations on personal data as well as any applicable local legislation. Also, voluntary informed consent of the research participants will be obtained in all cases. The project will not involve persons unable to give informed consent apart from occasional interviews/questionnaires with children, in which cases the informed consent will be obtained from their parents or legal guardians. Data will be collected, stored, protected and disposed of according to the applicable national and local regulations. In addition, all research, independent from the field site location, will comply with the Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. All publication and dissemination of the projects data will be done in a manner respecting the research participants right to privacy and no link to actual persons will be included in such materials.