H2020-MSCA-ITN-2018

FutureArctic – 813114

A glimpse into the Arctic future: equipping a unique natural experiment for next-generation ecosystem research’ — ‘FutureArctic

Data Management Plan

Deliverable 4.6

Project Information

Project Acronym:	FutureArctic
Project Full Title:	A glimpse into the Arctic future: equipping a unique natural experiment for next-generation ecosystem research' — 'FutureArctic
Call:	H2020-MSCA-ITN-2018
Grant Number:	813114
Project URL:	https://futurearctic.eu/

Abstract

Climate change will affect Arctic ecosystems more than any other ecosystem worldwide, with temperature increases expected up to 4-6°C. While this is threatening the integrity and biodiversity of the ecosystems in itself, the larger ecosystem feedbacks triggered by this change are even more worrisome. During millions of years, atmospheric carbon has been stored in the Arctic soils. With warming, the carbon can rapidly escape the soils in the form of CO2 and (even worse) the strong greenhouse agent CH4. Despite decades of research, scientists still struggle to unveil the scale of this carbon exchange, and especially how it will interact with climate change. An overarching question remains: how much carbon will potentially escape the Arctic in the future climate, and how will this affect climate change? FutureArctic embeds this research challenge directly in an inter-sectoral training initiative for early stage researchers, that aims to form "ecosystem-of-things" scientists and engineers at the ForHot site. The FORHOT site in Iceland offers a geothermally controlled soil temperature warming gradient, to study how Arctic ecosystem processes are affected by temperature increases as expected through climate change. FutureArctic aims to pave the way for generalized permanently connected data acquisition systems for key environmental variables and processes. We will initiate a new machine-learning approach to analyse large high-throughput environmental data-streams, through installing a pioneer "ecosystem-of-things" at the ForHot site. FutureArctic will thus channel, building on a timely project in the ForHot area, an important evolution to machineassisted environmental fundamental research. This is achieved through the dedicated training of researchers with profiles at the inter-sectoral edge of computer science, artificial intelligence, environmental science (both experimental and modelling), social sciences and sensor engineering and communication.

Document information

Version number :	v1.0
Description :	Final first version
Date of first version:	14/02/2020
Date of last update:	14/02/2020
Author:	Joke Van den Berge
DMP url:	https://futurearctic.be/data-access/ Future updates of the DMP can be found here

Contents

1. Data summary

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

Data will be collected in Work packages 1 - 3 to achieve the respective objectives. ESRs are embedded into a workflow (Fig. 1) that integrates **ecosystem scale environmental and earth science empirical research** (*WP1*, ESRs 1-8)) and **technological research on permanent ecosystems sensors and autonomous UAV remote sensing**, pushing current limits of ecosystem sensors to new levels of performance (*WP2*, ESRs 9-12). This supports the **machine assisted analysis of all data** through the development of novel algorithms for complex environmental data and for efficient data management (*WP3*, ESR 13-14). The WP3 analysis feeds back into the ecological models used by WP1 ESRs, giving important inputs on interactions between multiple ecosystem compartments and environmental variables. WP3 also feeds back into the performance of WP2 sensors. ESR15 (WP3) looks at the project from a birds-eye view, assessing societal and science-philosophical aspects of the novel analysis structure we propose.

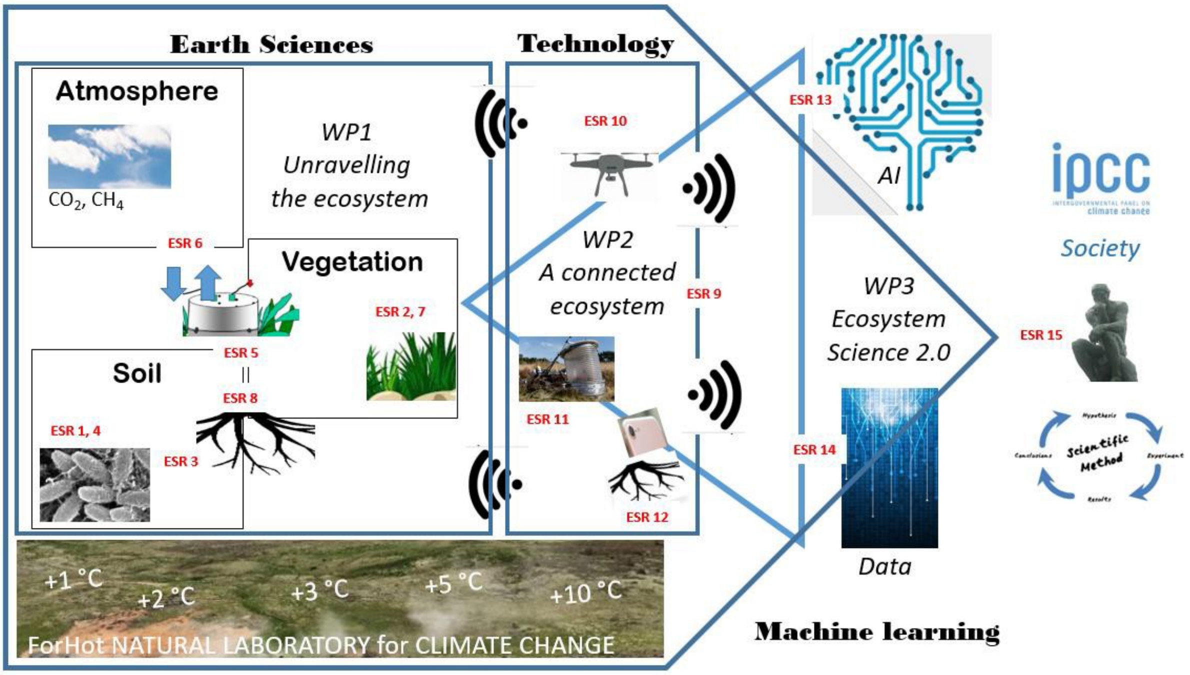

Fig. 1 The project workflow of FutureArctic

WP 1 Unravelling the ecosystem (ESRs 1-8)

Objective 1: improving process understanding in the ForHot ecosystem, unravelling ecosystem complexity and interactions through focused research projects.

In the soil compartment, the focus lies on the assessment of **root growth and phenology** (**ESR 3**) and **soil microbial community physiology, composition and functioning** (**ESR 1, ESR4**), with multiple interactions between the ESRs to asses root-rhizobiome-microbiome interaction.

ESRs 2 and **7** will focus on the **functioning of the plant community**, with a specific focus on plant and vegetation traits, community composition and interaction with environmental controlling variables (ESR2) and plant phenology and plant stress adaptation (ESR7).

ESR8 focuses on the **functioning of the food-web as a whole**, where ecosystem health assessment is based on the analysis of metabolites emitted by organisms and shifts in elemental composition (elementome).

ESRs 5 and 6 will focus on the **ecosystem carbon balance**, with detailed gas exchange measurements (ESR 6), cooperation with ESR 11.

A crucial aspect of FutureArctic is that all ESRs in WP1 and WP2 will focus on the same existing ecosystem scale warming experiment, and will coordinate sampling efforts in the tightest way possible.

WP2: A connected ecosystem (ESRs 9-12)

Objective 2: develop ready-to-market online and permanent ecosystem sensors and technology and develop the prototyping platforms for connecting the multiple sensors to a central database.

ESR9 is at the interface of WP1, 2 and 3, focusing on the development of a modular, ruggedized development kit for **fast wireless sensor prototyping in a harsh environment**.

ESR 10 focuses on novel **UAV hyperspectral assessments** and will find strong synergies in cooperation with ESRs 2 and 7.

ESR 11 will focus on technological developments for **permanently coupled flux chamber-lysimeter assessments** and will cooperate with ESRs 5 and 6.

ESR 12 will develop **new imaging technology for root phenology and root growth assessment and for the identification of root taxa by hyperspectral imaging**, in specific synergetic interaction with ESRs 1,3 and 4.

WP3: Ecosystem Science 2.0 (ESRs 13-15)

Objective 3: use machine-learning to unravel complex ecosystem interactions that are difficult to detect using traditional focused experimental and empirical research in ecosystem and earth sciences.

ESRs 13 and 14 focus on the development of algorithms and machine-learning technologies to allow **AI-based analysis of complex patterns** at the ecosystem scale (ESR13) and to allow an **automated management of online collected data and optimize the storage and analysis of ecosystem data both on the edge and in the cloud** (ESR14). WP3 analyses will feed back into the hypotheses and experiments developed in WP1.

In a unique birds-eye assessment, **ESR15** will assess **societal** (e.g. implications for IPCC assessments) **and science-ethical consequences of the machine-assisted environmental research**.

What types and formats of data will the project generate/collect? What is the origin of the data? What is the expected size of the data?

WP 1 Unravelling the ecosystem				
ESR	**Data type**	**Data format**	**Origin**	**Size**
1	- sequencing data: fastq.gz files of sequence reads for each sample (= tarball of fastq files) - environmental descriptors of samples	fastq.gz .txt files	Field surveys and in-situ experiments	20 GB
2	- drone images - preprocessed images - data extraction via QGis/ArcGIS Pro - statistical analysis data - on-ground measurements - machine learning using python	.arw, .jpg, .tif .shp .qgs, … .r scripts .xlsx, .csv, .txt .py	Field survey	5-15 Tb
3	- root traits (mass, shape dimensions (length, diameter, volume)), chemical parameters - root images - root associated microbial data (sequencing)	.xls, .csv .jpg fastaq.gz	Field survey, lab and field incubations	1-2 Tb
4	- Soil chemical analysis - Soil molecular analysis - iChip setup and testing - Growth and turnover measurements of iChip Isolates - Physiological measurements of iChip isolates - NanoSIMS and FISH measurements - Modelling	.xls FASTQ .pdf .xls .xls .tiff .csv	Field surveys, in situ experiments, lab incubations and computers simulations	50 GB
5	soil, plant and gas samples	.xls, .csv	Field surveys, in situ experiments, lab incubations	30 GB
6	- Sensor and analyzer data from automatic gas measurement chambers - Flux calculations	ASCII .Rscripts	Field surveys, lab incubations	20-50 Gb
7	- Soil T sensors in 10 cm – 1h resolution - Manual NDVI/PRI measurements - Plant water potential and gas exchange - Harvest measurements - Statistical analysis data	.xls, .csv .xls, .csv .xls, .csv .xls, .csv R scripts	Measurements, field surveys	25-30 GB
8	- Leaf phenology data (eco-physiological data about timing of seasonal growth) - Metabolomics, i.e. concentration levels and concentration changes of a 800-1000 small organic molecules)	.xls, .csv RAW (orBruker or MZxml), .xls, .csv	Measurement results, lab analysis	100 Gb per 100 samples (metabolytes)

WP 2: A connected ecosystem				
ESR	**Data type**	**Data format**	**Origin**	**Size**
9	- Sensor data of the installed remote environmental sensors	MongoDB and Timeseries database entries	Wireless sensors installed in the field	10 Gb
10	- Sensor data, RGB images, Thermal (LWIR) images, multispectral images, LiDAR point cloud - Software for automated, data processing and analysis	.jpg, .tiff, .las etc. .py	Field surveys	300-1000 GB
11	- Analysis of conceptual model will provide graphics which visualize the processes in the soil that influence soilwater nutrient contents - Soil water sensor - Measurement of soil-water retention curve and hydraulic conductivity.	.ppt .xls, .csv .xls; .csv	Measurement results, field surveys, test data from new equipment	1-10 GB
12	- RGB images, spectral pattern/curves of RGB image sensors, spectral pattern determined with hyperspectral or monochromatic imaging devices; raw files and processed images, metadata of captured image - Accompanying information on soil moisture (e.g. TDR probes) and texture, soil pH, nutrient availability and C, soil and air temperature (thermocouples), - Statistically processed data of aforementioned spectral pattern, incl. models and machine leaning approaches, neural networks etc - Designs and technical detail drawings (CAD) of various minirhizotron imaging devices; - Software to operate minirhizotron devices (semi-) automatically; - Software to organize and analyse MR images autmatically for root leght, root class, root taxa and soil moisture	.jpg, .tiff, etc. .txt, .csv, etc txt, csv, xlsx etc .xls, .Rscripts, SAS, SPSS output file formats CAD OS Windows, Arduino, RasberryPi	Lab experiments, Field set-ups aim to test the developed equipment under arctic conditions and to validate the methodology versus established, manual MR techniques (ESR3)	3.5 TB

WP 3: Ecosystem Science 2.0				
ESR	**Data type**	**Data format**	**Origin**	**Size**
13	No data	/	/	/
14	No data	/	/	/
15	- Interviews with researchers and the corresponding verbatim transcripts - Videos from the field sites - pictures from field work - Atlas.ti files for coding the transcripts from interviews and other collected material - Informed consent files (given by interview partners to allow the recording of interviews) - Filed notes	.doc; .wav; MP3; MP4; AVI, WMV. Jpg; tiff hpr.8 .doc .doc	Interviews, pictures, videos with researchers in the ITN network during field work and at the labs; inform consent is a prerequisite; transcripts and analysis files	2-8 GB

Will you re-use any existing data and, if so, how?

The research initiated over the last 5 years at the ForHot site provides crucial information for the setup of the FutureArctic common sampling network, and also provides important baseline data for WP1 and WP3. Usage of data is described in the Memorandum of Understanding with the partners of the ForHot consortium. The ForHot data will be made available via the FutureArctic Invenio Database solution.

ESR	Re-use of data
1	ForHot database - Temperature data of each plot/transect through time (as a predictor) - Plant species composition of plots (as a predictor)
2	/
3	Own database and ForHot database: - Soil data: organisms, chemistry, physical parameters, aggregates, water, temperature - Plant data: biomass, growth dynamic, chemistry…
4	ForHot database - Temperature data of each plot/transect through time - Plant species composition of plots - Microbial Biomass and community composition Microbial processes (growth rates, N mineralization, and similar)
5	Yes, own database and ForHot database
6	Meteorological data
7	ForHot database: - NDVI measurements (2013-2019) - Biomass data 2013-2018 - Soil temperature and soil water 2013-2019 - GPP data 2013-2017

8	ForHot database and own database: - elemental composition of soil, microbes, litter and plants - metabolomic data of plants and soil, plant - community structure (species composition) - biomass and growth - soil enzyme activities - C- and N-cycle - phenology data
9	/
10	/
11	Yes, overview will be added to next version
12	- maps of the ForHot research area incl. mapped temperature gradients to set a design for testing the operation of various, automated minirhizotron imaging devices. - a taxonomic list of plant species present (and their abundance) at the ForHot research site to select three representative species for hyperspectral screening of roots and the development of a root taxa "detection algorithm"
13	/
14	/
15	/

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

The data sets will be shared within the consortium to reach the objectives as described in the Grant agreement.

After the data are publicly available, the data can be used by independent researchers and other interested parties to understand better the contents and conclusions of the scientific publications, which base their findings on the data. Furthermore, independent researchers can use the files to produce figures and publications, showing comparisons of their own results and the FutureArctic results. Data can be used in larger datasets for meta-analysis.

2. FAIR data

2.1 FAIR data: Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata?

The data will be stored in the Invenio platform, hosted by partner IMEC. This Invenio solution is based on the de facto industry-standard Zenodo. This platform is searchable for Metadata.

Are the data produced and/or used in the project identifiable and locatable by means of a standard identification mechanism?

The data will be identifiable and locatable with Digital Object Identifiers (DOIs) that are used in the Invenio platform.

What naming conventions do you follow?

The filename of a dataset contains a clear, concise and short name that identifies the contents. Individual files are placed in subfolders according to this structure:

FOLDER:
LLLLL-YYMMDD is the name of the folder for a measurement campaign. LLL is the abbreviation of the Consortium member institute at which the data set is created. See Table 1 for list of abbreviations.

The folder contains at least the following FILES:

- README.TXT - brief and clear description of the folder contents, authors, other useful information (e.g. details about measurement campaign)
- FILES with datasets, results, …
- METADATA_FILENAME.TXT – Metadata for each file.

The file and folder naming convention will evolve according to the project needs and with growing experience.

Table 1: Organisations with their abbreviations

Institute/Company	Abbreviation	Institute/Company	Abbreviation
Universiteit Antwerpen	UA	Landbunadarhaskoli Islands	LBHI
Eigen Vermogen Van Het Instituut Voor Landbouw- En Visserijonderzoek	ILVO	Centro De Investigacion Ecologica Yaplicaciones Forestales Consorcio	CREAF
Tartu Ulikool	UTARTU	Interuniversitair Micro-Electronica Centrum Vzw	IMEC
Universitaet Wien	UNIVIE	DMR A/S	DMR
Universitaet Innsbruck	UIBK	SVARMI EHF	SVARMI
Kobenhavns Universitet	UCPH	Vienna Scientific Instruments Gmbh	VSI

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

Each dataset will at least be tagged with the following keywords:

1. FutureArctic
2. EU H2020
3. climate change

In addition, appropriate keywords will be added. This keywords can include:

- ForHot
- Site name
- measurement type
- …

What is your approach for clear versioning?

We will probably not need versioning conventions. If needed, versioning conventions will be added to the next version of the DMP.

What metadata will be created?

The following metadata will be provided for each file in a METADATA_FILENAME.txt

- **Project Identifier** Project Title, grant provider, grant number
- **Title** meaningful name of the measurement series
- **Description** Description of the experiment, description of plot and sample numbering (Sample ID), time frame of the measurements, other relevant info to understand the dataset.
- **Documentation** description of the dataset content, all variables (column headers), all calculations
- **Format** file type
- **Identifier** DOI reserved for this data set
- **Link** to where data are stored. Path of the folder in the common storage system.
- **Issue date** date of first issue
- Last modification date
- **Contact** name + e-mail
- **Institue/Company:** Name of the institute/company.
- **Quality control:** name ESR + name supervisor
- **Data Policy**: Indicate under which license data is licensed.

The metadata convention will evolve according to the project needs and with growing experience.
A Sample ID system, to identify measurements on shared samples, will be added to the DMP once it is developed. This should be done before the first measurement campaign (July 2020).

2.2. FAIR data: Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If some data is kept closed provide a rationale for doing so. When will the data be made available for re-use? If applicable, specify why and for what period a data embargo is needed.

Data will be added every six months to the Invenio platform. Data can be indicated as:

- closed
- restricted access: open for beneficiaries of the FutureArctic project
- open for public: All data will be indicated as open, after publication of the results or at the latest three years after the end of the project if the data are not published.

Each year at the annual meeting, a check is carried out to determine which data may be made openly available. Three years after the project ends all data will be made openly available.

How will the data be made accessible?

The data will be stored in the Invenio platform, hosted by partner IMEC. This Invenio solution is based on the de facto industry-standard Zenodo. This platform is freely accessible.

What methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

No specific methods and software tools are needed to access the data.

Where will the data and associated metadata, documentation, and code be deposited? Have you explored appropriate arrangements with the identified repository?

The data will be stored in the Invenio platform, hosted by imec. Arrangements have been made. The service is free of charge as imec is a beneficiary in the project.

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

If there are restrictions on use this will be indicated in the INVENIO platform. The supervisor will be marked in the INVENIO platform as the contact person to request access.

2.3. FAIR data: Making data interoperable

Are the data produced in the project interoperable? What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

We collect the best practices of the beneficiaries and we strive to make maximum use of the best practices within consortium.

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

We will use as much standard vocabularies that are used in the consortium. Project specific vocabularies will be mapped.

2.4. FAIR data: Increase data re-use (through clarifying licenses)

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

The data is owned by the beneficiary that collected/created the data. Each beneficiary indicates how the data is licensed. The license will be indicated in the Invenio Platform and in the Metadata Files.

These are five model licenses that can be used:

- **A Creative Commons Zero12 License**, in which the institution renounces its intellectual property rights, as much as legally possible. This allows the user to reuse the data for any purpose, without any obligation regarding attribution.
- **Creative Commons CC BY 4.0** in which data can be reused for any purpose, with an obligation for attribution (https://creativecommons.org/licenses/by/4.0/).
- **Free Open Data License**: under this license the institution does not relinquish their intellectual property rights, but the data can be reused for any purpose, for free and under minimal restrictions.
- **Free Open Data License for Non-Commercial Reuse**: to comply with the principle of Open Data, the data must be available under minimal restrictions for both commercial and non-commercial re-use. A distinction between types of use can be made if the institution so wishes. For commercial reuse a fair compensation may be required, while non-commercial re-use will be made free of charge. This license governs free non-commercial re-use.
- **Open Data License with Equitable Remuneration for Commercial Use**: When a distinction is made on the basis of the commercial nature and a fee is asked for reuse, this license is the counterpart of the Free License for Non-Commercial Reuse.

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

The data is reusable. Embargoes will disappear at the latest three years after the end of the project. This embargo is set to give the ESRs the chance to publish their results before they are openly accessible.

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

It is intended that the data remains re-usable for 20 years after the end of the project.

Are data quality assurance processes described?

Quality of the data sets, metadata and measurement setup and procedure description are the responsibility of the ESR and the supervisor, which are both mentioned in the metadata.

3. Allocation of resources

Who will be responsible for data management in your project?

Dr. Joke Van den Berge, is responsible for implementing the DMP, follow up of the data storage and metadata production process. She will also update the DMP when conditions change and will make sure the DMP is effectively applied during the project. The infrastructure for storage and backup will be introduced by beneficiary imec. The project manager will facilitate data archiving and data sharing.

What are the costs and potential value of long term preservation?

Beneficiary imec covers the costs for hosting the Invenio solution. If the Invenio solution fails the data will be transferred to Zenodo.

What are the costs for making data FAIR in your project? How will these costs be covered?

The project manager is paid by the FutureArctic consortium. Beneficiary imec covers the costs for hosting the Invenio solution.

4. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing?

For interviews, field observations, as well as for taking pictures or making videos, we need informed consent from the participants (other researchers in FutureArctic). In this consent form participants can chose whether they want their identity to be disclosed or not. If they give us permission to disclose their identity they can further chose if they want/do not want to be quoted verbatim in the final PhD thesis and related publications that will build on these data. Depending on the choices the personal information will be removed from the transcript of any interviews (interviews will be pseudonymised). The informed consent forms, data on researchers interviewed, audio files as well as images that should not be disclosed will be stored in a separate password protected folder in order to comply to data protection.

5. Data security

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

The data is stored in a secure environment, hosted in the imec data centres. The data storage solution used, Invenio, is being used in industry-quality solutions such as Zenodo.org, where its security has been validated. The FutureArctic deployment will use HTTPS for secure data transfer. Data recovery is tackled by data backup policies in place at the imec data center.

Sensitive data, both the recording and the (pseudonymized) transcript (see choices above) will be stored separately in a password-secured server at the University of Vienna, in folders only accessible to ERS15 and the supervisors. Necessary precautions will be made to assure information of research participants will be handled in accordance with relevant data regulations. Data recovery is tackled by data backup policies in place at the servers of the University of Vienna.