iCUPE Data Management Plan (ver 1)

Alexander Mahura, Tuukka Petäjä, Hanna K. Lappalainen (University of Helsinki); Steffen M. Noe (Estonian University of Life Sciences); Stefano Nativi, Paolo Mazzetti (Consiglio Nazionale delle Ricerche), and DS Leaders[1]

Helsinki, Finland, 15 November 2017

WP5: Data provision, interoperability and facilitation of data and services

Task 5.1: iCUPE data management plan / Deliverable 5.1.1: iCUPE data management plan (ver 1)

Version 1

Introduction

The iCUPE WP5 ("*Data provision, interoperability and facilitation of data and services*") facilitates the data provision and services from the iCUPE project (www.atm.helsinki.fi/icupe) to the end-users, decision-makers, and stakeholders. Data obtained, integrated and accessed in WPs 1-4 are distributed in WP5 as data products and assessments. Data (ground-based, time-series, column, etc. observations) will be harmonized and assessed through developed novel methods, proxies and observables. These will be delivered through interoperability tools and services and will be available to iCUPE partners and any other end-users, decision-makers, and stakeholders (in agreement with existing data policies and open source principles).
To achieve the ERA-PLANET (European Network for Observing our Changing Planet; www.era-planet.eu) overall objectives, and to pursue cross-thematic interoperability and contribute effectively to GEOSS (Global Earth Observation System of Systems; www.earthobservations.org/geoss.php), the iCUPE project will implement the best practices and recommended approaches of ERA-PLANET. This will allow project to contribute to GEOSS via the GCI (GEOSS Common Infrastructure) and to utilize relevant Copernicus data and Core Services and EU capabilities in the EO domain. iCUPE will promote and implement the use of open specifications (i.e. international standards, community specifications) for data sharing and will foster technological development to deliver more timely and high quality data and information, in compliance with the GEOSS Data Management Principles.

[1] see responsible persons DS Leaders in Appendix "iCUPE deliverables as datasets (DS)" of this Data Management Plan; appendix has contact information of persons responsible for delivering the datasets, and this information is available only at the internal iCUPE project website.

The **iCUPE Data Management Plan (DMP)** (Task 5.1) will assure interoperability between the ERA-PLANET projects/ strands and with other activities carried out as part of GEO Strategic Plan (2017-2019) (i.e., GEO Initiatives, Flagships, Foundational tasks). As part of making iCUPE research data findable, accessible, interoperable and re-usable (FAIR), iCUPE DMP will include information: on handling of obtained research data (during and after the end of the project); on types of data to be collected, processed, analyzed, etc.; on applied methodological approaches and standards; on whether data will be shared and/or made open access; and on how data will be curated and preserved (including after the end of the project). The most important aspects for data management include: discoverability (data and metadata should be discoverable); accessibility (data should be accessible in online services); usability (encoding, traceability, documentation, quality); preservation (preservation, verification); and curation (review and reprocessing, persistent and resolvable identifiers).
Following the signed iCUPE Consortium Agreement /Section 11 on Data Management/, the appropriate and secure use of material and data of the project will be enabled according to the application of common standards. In the iCUPE project, the DMP will follow the "*Guidelines on FAIR Data management in Horizon 2020*" (version July 2016; [2]) and according to the ERA-PLANET Data Management Plan (Deliverable 4.5). The iCUPE DMP will be updated during the lifetime of the project.

1. Data Summary

- *State the purpose of the data collection/generation & explain relation to objectives of the project*
- *Specify the types and formats of data generated/collected*
- *Specify if existing data is being re-used (if any)*
- *Specify the origin of the data*
- *State the expected size of the data (if known)*
- *Outline the data utility: to whom will it be useful*

1.1 Purpose of data collection/ generation & relation to project objectives

The iCUPE addresses the overarching objective of ERA-PLANET to simplify access to information required by decision-makers and brings together and strengthen the European national and regional research and innovation programs in the EO domain. In particular, the iCUPE connects to the thematic strand 4 on polar areas and natural resources by integrating national and international monitoring and assessment activities in relation to ecosystem and environment quality in Arctic and Antarctic regions.

The work in iCUPE answers to the ERA-PLANET horizontal objectives and overarching goals of GEOSS and Copernicus integration by facilitating integration of the environmental pollution and sources and their transformation and impacts. The iCUPE platform takes on-board national agendas and improves co-alignment of national activities, and in particular, in the Arctic environmental observations. The

[2] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

iCUPE project endorses open data policies and Key Enabling Technologies (KET) via horizontal work within the three other ERA-PLANET strands. The iCUPE work provides ground-truthing for satellites in relation to snow, precipitation and aerosol observations.

The iCUPE will: 1) synthesize data from comprehensive long-term measurements, intensive campaigns and satellites, collected during the project or provided by on-going international initiatives; 2) relate the observed parameters to impacts;, and 3) deliver novel data products, metrics and indicators to the stakeholders concerning the environmental status, availability and extraction of natural resources in the polar areas. The obtained results/ data will be useful for policy development and for improving and clearly communicating multidisciplinary understanding of status of the polar environment and pollution dynamics in the future.

1.2 Types and formats of data generated/ collected

The iCUPE will apply an integrated approach through combining and providing several types of in-situ, ground-based, airborne observations, satellite remote sensing and multi-scale modeling data. The produced datasets (DSs) for Arctic/Polar regions and for selected locations (to be specified for each dataset separately) will include information/ data on: in-situ, ground-based, remote sensing and airborne measurements; anthropogenic contaminants in snow and in ice cores; atmospheric Hg(II) and Hg isotope observations; organic contaminants in air; ground-based measurements for particle number, black carbon mass and ozone concentration; organic contaminants in snow and water; satellite derived cryospheric measurement data; snow spectral reflectance; aerosol vertical profiles from ground-based and satellite observations; blueprint for novel proxy variables integrating in-situ and satellite remote sensing data; novel optical remote sensing products on snow and on vegetation and gas flaring mapping; precipitation in the high-latitudes.

The methods and approaches for collecting mentioned above data are described in the iCUPE project Description of Work (DoW) as part of the project research plan. These standard collection methods and validated protocols are commonly used in the research field.

All DSs will be immediately available for the iCUPE consortium partners following a principle of “Internal use until publication”, and then these DSs will be publicly available for all other potential users (e.g. decision-makers, stakeholders, end-users and other researchers, whom are not directly involved into iCUPE project activities).

A series of brief descriptions (1 page summary including data examples, contact information, etc. and extracts from the data available online together with the metadata) of planned DSs will appear by end summer 2018 (e.g. in advance before the 1st deliverable expected as the dataset, by M16 or Dec 2018) at the ICUPE Datasets web-page (https://www.atm.helsinki.fi/icupe/index.php/datasets). This will allow potential users to learn and test applicability of DSs and to consider opportunities for establishing collaboration with the DSs owners (responsible for delivering DSs) as well as other iCUPE partners.

The formats and specifications for each individual dataset will be provided in the next version of DMP (*based on information provided by responsible persons for delivering the specific DSs*). Where it is applicable, the data formats may be migrated in case of new technologies will become available and are proved to be robust enough in order to ensure digital continuity and continued availability of data.

1.3 Existing data reuse (if any)

The iCUPE activities to a high degree is based on data currently residing in existing data and information repositories and these will be used to produce new compilations and other derived datasets as products. The possibility of data reusing will be also considered, where it is applicable; and it will allow avoiding duplication of work already done. This is to be clarified during the first year of the iCUPE project (*and to be available in next version of DMP based on information provided for each DS by responsible persons for delivering such DS*).

1.4 Origin of data

Data originates from various national data and information repositories as well as some will be taken from on-going and planned observations and measurements during the project lifetime. Compilations of data will be produced by iCUPE, but at some degree will be also linked with other ERA-PLANET strands/ projects. This is to be clarified during the first year of the iCUPE project (*and to be available in next version of DMP based on information provided for each DS by responsible persons for delivering such DS*).

1.5 Expected data size

The expected sizes of datasets (DSs) to be delivered as products are not known yet (*to be clarified during lifetime of the iCUPE project and to be available in next versions of DMP based on information provided for each DS by responsible persons for delivering such DS*). The expected size will depend on extend and nature of the data that will be made available.

1.6 Data utility

The data generated by the iCUPE project will be useful for decision-makers, policy-makres, stakeholders, end-users as well as for other researchers (whom are not involved in the iCUPE project). In particular, iCUPE data can be widely utilized by decision-makers, governmental organizations, regional and local authorities, national environmental agencies and ministries, national weather and air pollution control services at national and European levels, etc. Among these are the research communities working on tasks linked with the air quality and climate modelling, and numerical weather prediction. Moreover, international in-situ observational networks such as PEEX, WMO-GAW, AMAP, INTAROS, and others (monitoring tasks and filling gaps in polar regions) are potential users. And of course, concerned citizens (whom are interested in environmental problems) are also potential users, including both school and universities educational systems promoting various programmes for better understanding of environment and sustainable development of society in a changing climate (and especially in most vulnerable polar regions). In addition, a detailed information will be also available in the iCUPE Stakeholder Engagement Plan (Deliverable 6.1.1).

2. FAIR (Findable, Accessible, Interoperable and Re-usable) Data

2. 1. Making data findable, including provisions for metadata

- *Outline the discoverability of data (metadata provision) & outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?*
- *Outline naming conventions used*
- *Outline the approach towards search keyword*
- *Outline the approach for clear versioning*
- *Specify standards for metadata creation (if any).*

Q1: Are the data produced and/or used in the programme discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

All data will be discoverable with the metadata. Many will also be identifiable and locatable by means of standard identification mechanisms. The iCUPE will be encouraged to consider Digital Object Identifiers (DOIs) for produced datasets.

Q2: What naming conventions do you follow?

The proposed for dataset, metadata, and template names, the iCUPE will define naming convention consisting the following key parts: i) prefix indicating if it is dataset, metadata or template; ii) root composed by: short and meaningful name of dataset/template & acronym or short name of the data provider (i.e. organization; for example: iCUPE - by default for templates); and iii) suffix indicating date of the last upload into the repository in YYYYMMDD (year-month-day format).

Q3: Will search keywords be provided that optimize possibilities for re-use?

The metadata system will provide opportunities for tagging the datasets and their content with keywords. In particular, the dataset information reported into the metadata will be published, where specific filters, based on the metadata elements, will allow to refine the search across datasets (e.g. search dataset by key words, by temporal or spatial coverage/location of data, by selected parameters or group of parameters, etc.).

Q4: Do you provide clear version numbers?

It is expected that many of delivered datasets will have version numbers. E.g. after the project ended (or during lifetime of the project), for example, for dynamic in nature datasets: new quality controlled measurement data can be added to already existing dataset represented as a time-series of observations. Although, for selected datasets some providers may not define the versions. This will be clarified during the 1st year of the project. In general, the versioning management of data and files stored into the repository, will be realised via the naming convention and use of the date as a suffix (indicating the latest version of the file uploaded into the repository).

Q5: 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.

The data will be documented following metadata standards. Each dataset documentation will include measurement stations/site basic information, explains terms, variable names, codes (or abbreviations used), etc. (questionnaire to be distributed among the persons/ partners responsible for the datasets). Considering possibility of DS usage in the future, a set of information needed to find, use and interpret the data and description of documentation's types (that will accompany DS) will be prepared. Metadata will provide standardized structured information with explanation of purpose, origin, time references, spatial (geographic) location, creator, access conditions and terms of dataset usage.

Consistency and quality of iCUPE data will be controlled and documented (on how data were collected). Data quality control will ensure that no data will be lost or accidentally changed. Such procedure is an integral part of DS creation and it takes place during data collection, entering or digitization, and checking. Quality control measures include e.g. standardized methods and protocols for making observations, alongside recording forms with clear instructions, calibration of instruments, etc. Missing data codes will be defined that actual data values fall within the range of expected values. Processed datasets will be reviewed by the iCUPE DMP responsible persons.

2.2. Making data openly accessible

- *Specify which data will be made openly available? If some data is kept closed provide rationale for doing so*

- ➢ *Specify how the data will be made available*
- ➢ *Specify 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)?*
- ➢ *Specify where the data and associated metadata, documentation and code are deposited*
- ➢ *Specify how access will be provided in case there are any restrictions*

Q1: 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.

All produced datasets will be public of nature and shared. No restrictions are applied. Except, the raw data (used to produce datasets) will be available through request to the responsible person/ partners.

Q2: How will the data be made accessible (e.g. by deposition in a repository)?

Basic user requirements document will be posted on the iCUPE project website with instructions on how to access and use datasets.

Q3: What methods or software tools are needed to access the data?

Datasets will be accessible via HTTP and FTP by downloading/ ftp-ing files containing data and files with descriptions of datasets.

Q4: Is documentation about the software needed to access the data included?

Information about possible (preferably, publicly available) software/ tools to access datasets will be also included in description (as basic instructions on how-to-do and with corresponding web-links to technical aspects) of datasets produced.

Q5: Is it possible to include the relevant software (e.g. in open source code)?

A pointer to documentation on relevant standards can be included in open source code.

Q6: 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.

Datasets will be available through the iCUPE project website.

Q7: Have you explored appropriate arrangements with the identified repository?

Several options to be explored.

Q8: If there are restrictions on use, how will access be provided?

As the iCUPE produced datasets are all public in nature, there are no restriction on their usage. The iCUPE data to be stored are encouraging an unlimited and open data policy for non-commercial use. Data will be available and cited in publications. Interested persons will be able to contact with partners-owners of DSs (with CC to PI) for accessing data. For research and educational purposes, access to these data is unlimited and provided without a charge. By using such data, the person should accept that an offer of co-authorship will be made through personal contact with the owners whenever substantial use of such data is made. In all cases, an acknowledgement should be made to the owners and to the project name when these data are used within a publication.

Q9: Is there a need for a data access committee?

The iCUPE Project Office can take a role of the Data Access Committee, when it might be necessary. For example, for providing access to collaborators (i.e. researchers, whom are not involved into the iCUPE project) to specific dataset after discussion and agreement with the owner/ partner, whom produced DS.

Q10: Are there well described conditions for access (i.e. a machine readable license)?

To be specified and described at the later stage of the project.

Q11: How will the identity of the person accessing the data be ascertained?

Issues regarding authentication, authorization and accounting will be dealt with on general terms. The quality checked newly generated datasets (DS) including the metadata descriptions together with the harmonized database of long-term time-series produced in the iCUPE will be curated by the data producing partner, and saved and archived by national activities. The data will be accessible via iCUPE virtual platform (WP5).

2.3. Making data interoperable

- ➢ *Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.*
- ➢ *Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?*

Q1: 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)?

Produced datasets will be interoperable allowing data exchange and reuse. Documentation for each dataset will be produced.

Q2: What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

iCUPE will use well-established European and international standards for storage, exchange and dissemination of produced data. Every dataset will have metadata.

Q3: Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

iCUPE will use standard vocabularies to an extent that they exist (or might be also novel ones developed in the project).

Q4: 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?

Although this is not anticipated to be an issue, but if it is found unavoidable, iCUPE will be required to document a usage of uncommon or to generate specific ontologies or vocabularies. In a case, if mapping to more commonly used ontologies will be possible, then such mapping will be required to establish.

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

- *Specify how the data will be licenced to permit the widest reuse possible.*
- *Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed.*
- *Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.*
- *Describe data quality assurance processes.*
- *Specify the length of time for which the data will remain re-usable.*

Q1: How will the data be licensed to permit the widest re-use possible?

Datasets will be publicly available. Information to be available at the later stage of the project. To be decided by owners/ partners of the datasets.

Q2: 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.

It is not envisaged that iCUPE will seek patents. The data collected, processed and analyzed during the project will be made openly available following deadlines (for deliverables as the datasets) and once the corresponding peer-reviewed papers will be published. All datasets are expected to be publicly available by the end of the project.

Q3: 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 iCUPE general rule will be that data produced after lifetime of the project will be useable by third partied. For shared information, standard format, open source software, and proper documentation will guarantee re-usability by third parties. Regarding the iCUPE digital research data, the ownership of raw data is on responsible partner for delivering dataset.

Q4: How long is it intended that the data remains re-usable?

The data are expected to remain re-usable (and maintained by the partner/ owner) as long as possible after the project ended, at least, 5 year period.

Q5: Are data quality assurance processes described?

Data quality assurance processes will be described before the public release of the dataset.

3. Allocation of resources

Allocation of resources includes costs for making data FAIR (Findable, Accessible, Interoperable and Re-usable), responsibilities for data management, and resources for long-term data preservation. The iCUPE data management will be realized by Steffen Noe (WP5 leader, EULS), Alexander Mahura (UHEL), Stefano Nativi (CNR), Paolo Mazzetti (CNR) in close collaboration with the iCUPE Project Office as well as with active involvement of WP leaders and key researchers, whom are responsible for delivering data. Costs of making data FAIR are already included in this and other ongoing relevant projects as part data production and analysis. Time has been already allocated in WP5 for reach team by involvement of all partners, and it covers costs of preparing datasets and documentation for archiving.

Data will be stored at the coordinator's (UHEL) repository, and will be kept maintained, at least, for 5 years after the end of the project (with a possibility of further prolongation for extra years). UHEL repository will be managed and supported by a team of experts, and it is free of change. UHEL will be setting up and upgrading, when it is needed, the hardware and software components of the storage repository; creation, maintenance and upgrading database for users accessing datasets; collecting users requests for access to and download of data; co-creation of the data repository's folders/subfolders for each dataset and keeping document type (e.g. data, metadata, templates); capacity management of hardware and software components.

4. Data security

The iCUPE consortium partners endorse the open data policy as stated in the project DoW – all deleiverables as datasets are "P" public of nature. The long-term storage of iCUPE project data (in form of datasets, DS) and legacy will be planned in more details during the project. Several options are considered, such as topical data storages, EBAS database (ebas.nilu.no), smartSMEAR (www.atm.helsinki.fi/smartSMEAR), GEOSS portal (www.geoportal.org), World Data Centre of Aerosols (www.gaw-wdca.org), for mercury observations we could connect to GEO flagship Global Observation System for Mercury (GOS4M). The aerosol and trace gas observations will be continued in collaboration of ACTRIS Research infrastructure and its data services. In the future, PEEX offers continuation in the Arctic observations and integration of in-situ and satellite data in the Arctic context. For applicable data, another option could also be Copernicus Climate Change Service (C3S) that offers a possibility for operational and quality controlled information sharing. These aspects will be developed further in the evolving iCUPE data management plan).

For duration of the iCUPE project, data will be also stored at the project website (https://www.atm.helsinki.fi/icupe/index.php/datasets; initially - at internal, and then - at public domain) with corresponding links to access and description information. It will be managed and supported by a team of experts at the University of Helsinki (UHEL) and subject to the university's policies ensuring long-term security of the datasets, including version control and secure backups (i.e. there is no issue with the DSs recoverability). By depositing datasets, the iCUPE will ensure that the research data will be migrated to new formats, platforms, and storage media as required by good practice. The individual DOI's [=persistent identifier] will be generated enabling access to the iCUPE datasets via persistent links. The transfer of data is via a zip-process of distribution (at internal website – with access password separately distributed). Note that there are no iCUPE data of sensitive nature.

5. Ethical aspects

The iCUPE scientific focus is to provide novel insights and observational data on global grand challenges with a polar focus. The project is examining the environmental questions also having high ethical impact for the sustainable development of the polar regions. The project partners are in bringing forward knowledge about climate variability, climate change and adaptation to climate change. The iCUPE research approach will follow EU standards of ethical principles in its tasks and outcomes ensuring equality, quality and integrity in all conductions.

As an important part of the work also includes observations in Russia and use of such data available via collaboration between the Institute of Atmospheric Optics (IAO, Siberian Branch, Russian Academy of Sciences, SB RAS), CNRS, CNR and UHEL partners. For such iCUPE datasets, the data acquirement, storage, openness and sharing will be also based on requirements and guidelines of the participating Russian institutions. These partners are addressing the importance to adhere the ethical norms in research related to aspects of knowledge, truth, avoidance of error, falsifying, misrepresenting research data, promoting truth and minimizing error. The existing collaboration with Russian partners is based on well-established contacts through bi-lateral projects and contacts.

Another aspect is personal data (i.e. name-surname, e-mail address, telephone, and skype) of researchers responsible for producing and maintaining (after the end of the project) the iCUPE datasets (DSs). This information will be internally available for the project partners during lifetime of the project for more efficient work and discussions on creation, evaluation, testing, quality control, integration, etc. of DSs into the web-based system. For others (e.g. potential decision-makers, stakeholders, end-users as well as other researchers), only e-mail contact information will be available at the iCUPE Datasets web-page (https://www.atm.helsinki.fi/icupe/index.php/datasets).

Storage and access to iCUPE project DSs will be allowed. Dataset owner will have access to data and consortium partners will have access as far as it is needed to follow the project research plan. Intellectual Property Rights, in case when needed, will be agreed following the iCUPE Consortium Agreement.

6. Other issues

Concerning the iCUPE data management, the iCUPE project does not involve any issues raising security issues: i.e. there are no results obtained (including datasets to be delivered) raising such issues, and there are no EU-classified information as background or expected project results.

Information on suitable repositories for datasets produced by iCUPE can be also located using the Registry of Research Data Repositories (www.re3data.org) and Zenodo (zenodo.org) with providing tools to link publications and data. Information on research data management is also available from the Digital Curation Centre (www.dcc.ac.uk/dmponline) and ScienceMatters (www.sciencematters.io). In addition, the Research Data Alliance provides the Metadata Standards Directory (rd-alliance.github.io/metadata-directory) that can be searched for discipline-specific standards and associated tools, and the EUDAT B2SHARE (b2share.eudat.eu) tool includes a built-in license wizard that facilitates the selection of an adequate license for research data.

Appendix: “iCUPE deliverables as datasets (DS)”

Appendix has contact information of persons responsible for delivering the datasets, and this information is available only at the internal iCUPE project website.