Version 1.1
2020

COAT Data Management Plan

Climate-ecological
Observatory for Arctic Tundra

Contents

How to cite this document:

Jepsen, J.U., De Stefano, M., Frassinelli, F., Yoccoz, N.G., Vang, R. & Soininen, E. (2020). COAT Data Management Plan Version 1, COAT Climate-ecological Observatory for Arctic Tundra, www.coat.no. Available online at: https://data.coat.no/dmp/coat-data-management-plan_v1. https://doi.org/10.48425/0035406

Change log

Date:	Version:	Authors:	Template:	Quality assurance:
xx.03.2020	Version 1	J.U. Jepsen, M. De Stefano, F. Frassinelli, N.G. Yoccoz, R. Vang, & E. Soininen	H2020	Torkild Tveraa (NINA)
Change log				
Date:	Version:	Revised by:	Revisions made:	Quality assurance:
15.11.2020	Version 1.1	M. De Stefano	Minor updates	

Abbreviations

COAT: Climate-ecological Observatory for Arctic Tundra (www.coat.no)

SIOS-KC: Svalbard Integrated Arctic Earth Observing System – Knowledge Centre (www.sios-svalbard.org)

UiT: UiT Arctic University of Norway (www.uit.no)

NINA: Norwegian Institute for Nature Research (www.nina.no)

MET: Norwegian Meteorological Institute (www.met.no)

NPI: Norwegian Polar Institute (www.npolar.no)

AU: University of Århus, Denmark (www.au.dk)

UNIS: University Centre in Svalbard (www.unis.no)

RCN: Research Council of Norway (www.forskningsradet.no)

DCAT: Data Catalog Vocabulary

DC: Dublin Core

RDF: Resource Description Framework

CSW: Catalog Service for the Web

FAIR: Findable, Accessible, Interoperable, Reusable

COAT Data Management Plan

1. Data Summary

COAT (Climate-ecological Observatory for Arctic Tundra, www.coat.no) is an adaptive, ecosystem based, observation system for Arctic tundra in Norway. It aims to unravel how climate change affects arctic tundra food webs, what is the condition of tundra ecosystems, and to enable robust science-based management. The baseline of the COAT approach is conceptual models that specify the biotic links between monitoring targets (often species or species assemblages) and expected impacts of climate change and management actions. These models guide the monitoring design and the structure of the statistical models that estimate impacts and derive predictions based on the monitoring data and management policies. COAT is a collaboration between UiT, NINA, MET, NPI, UNIS and AU.

The COAT Data Portal hosts ecological and climatic primary monitoring data from COATs monitoring sites in high-arctic (Svalbard) and low-arctic (Varanger) tundra, as well as certain prioritized data from outside these regions. In addition, the COAT Data Portal contains a number of secondary (derived) data sets, termed *COAT state variables*, which are calculated from primary data originating either from within COAT or requested from other data repositories. The state variables are the primary input to the quantitative analysis and predictive modelling performed in COAT.

The majority of data in the COAT Data Portal are plot-based measurements (e.g. made on a fixed geographical point) or individual-based measurements (e.g. made on individual plants or animals) stored in simple text formats (.txt, .csv, .asc with UTF-8 encoding). In addition, the collection includes image based data, including remote sensing products (.tiff, .jpg, .png), acoustic data (.mp4) and vector-based data (.shp, .json). The primary users of the COAT Data Portal are researchers involved in COAT, as well as national and international collaborators. The Data Portal will also be useful for members of the wider scientific community who wish to access COATs long-term monitoring data. Importantly, due to the explicit consideration of management actions in COAT, the COAT Data Portal will provide state variables, and model predictions targeted directly towards local managers of, for instance, harvested species.

2. FAIR data

2.1. Making data Findable, including provisions for metadata

The COAT Data Portal is both a metadata registry and a data repository, and provides tools for publishing, sharing, and finding data. Metadata are based on a custom profile, partially converted and exposed with endpoints in different metadata standards (DCAT, DC/RDF, CSW/ISO 19115) and exchange protocols (OAI-PMH). Information on the available data sets is also provided as structured data (JSON-LD snippets) for improved findability on the web (i.e. google search). Data is organized in data sets, which can contain an unlimited number of data resource files in different formats, including ancillary information. Each data set is registered with a full set of metadata elements, providing identification, keywords, thematic classification, geographic location, temporal reference, version, licensing, references and citation (Annex A). Each resource also has its own metadata, partly automatically generated. COAT data sets are built following a custom naming convention, described in Annex B. A thesaurus of keywords was designed to provide a controlled vocabulary of terms and a thematic classification for search filtering. A mandatory metadata element is the geographical location, providing information about the geographical source of the data and storing coordinates, which enables spatial search capabilities. Each data set is also classified by an INSPIRE Topic Category Code, for improved interoperability. Data portal users can easily create data set versions, and choose which to keep private or publish. Published versions will be permanently available for citation purposes, and always findable. Persistency is guaranteed by the use of DOI unique identifiers. Each data set version is assigned a Datacite DOI.

2.2. Making data openly Accessible

Data and metadata in the COAT Data Portal are by default shared under the Creative Commons Public Licence (CC BY 4.0; see 2.4 for details and exceptions). Published data are accessible either directly via the COAT Data Portal web interface (www.data.coat.no[1]), or they can be harvested via the API using standard scripting in for instance R or Python. COAT maintains a github repository (https://github.com/COATnor) that contains example scripts documenting how COAT data can be requested via the API. The github repository also contains R scripts that document how each state variable is calculated based on COAT primary datasets. Links to these specialized scripts are made directly from the metadata of the relevant datasets. The identity of persons accessing the data from outside COAT is not monitored. Statistics on the number of downloads of each public dataset are kept for reporting purposes.

2.3. Making data Interoperable

Data interoperability is reached inside the COAT project by defining a common standard for data set structures and shared naming conventions. Data sets collected during the project or imported from external sources are formatted according to the COAT conventions, described in Annex B.

Most of the collected data is stored by convention in common open formats (.csv, .txt, .asc with UTF-8 encoding), structured according to project's conventions, and both data and metadata are exposed with recognized standards for metadata discovery and data access.

Mappings to internationally recognized ontologies are not planned at the moment.

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

Data and metadata in the COAT Data Portal are by default shared under the Creative Commons Public Licence (CC BY 4.0), unless they are limited by privacy or licence restrictions from data providers outside COAT. Examples of restricted data are i) data (coordinates, photos) which can be used to establish the exact breeding locations of certain species of conservation concern (dens of arctic foxes, nests of raptors), ii) data (coordinates, photos) which can be used to establish the exact locations of certain expensive field sensors, or iii) data in which data owners outside COAT have a commercial interest. State variables *derived* from restricted data are still shared under CC BY 4.0, even if certain underlying data sources are not.

Before being published under CC BY 4.0. data files could be subject to an *optional embargo of a maximum of 2 years*. This means that for a running time series, all years, except the two most recent years if in embargo, are publicly available at any one time. The motivation for this is to allow COAT researchers and students time to publish prior to release of the full data set. If a published dataset includes data in embargo, a portal user is informed about the embargo and redirected to a previous version including only public data.

The data flow in COAT is described in Annex C. It details how data and associated scripts are transferred, quality checked and deposited in repositories all the way from notebooks, field tablets or paper forms to the published datasets and derived state variables in the COAT Data Portal.

The maintenance of the COAT Data Portal, and hence re-use of COAT data, is currently guaranteed for a minimum of 5 years after the end of the COAT Infrastructure Implementation phase (2016-2020), hence until the end of 2025. This is in accordance with the contract with the funding bodies (see 3. Allocation of resources). The ambition in COAT, however, is long-term maintenance on the scale of decades, both for the observation system and the COAT Data Portal.

[1] To be made operational by the end of 2020.

3. Allocation of resources

Substantial resources are invested in COAT data management, in order to make COAT data public according to FAIR principles, and to provide COAT researchers with a tool for efficient data management, access and documentation. In the implementation phase (2016-2020) COAT Data Management involves the following main activities:

i) Design and implementation of a central CKAN-based repository, the COAT Data Portal, including routines for metadata management and versioning.
ii) Developing routines for file structure, file- and dataset naming conventions, and taxonomies for key-words, localities, and species names.
iii) Collection of primary data sets from COAT member institutions and formatting these according to conventions.
iv) Developing routines for data quality checking.
v) Developing and harmonising data collection protocols for all data sets.
vi) Developing scripts for calculation of COAT state variables from primary datasets.
vii) Establishing a COAT github repository for documentation, and for sharing data collection protocols and scripts for data management and analysis.

The work is made possible by a contribution of ~3.5 mill NOK from RCN, as part of a larger research infrastructure allocation to COAT Varanger 2016-2020 (project grant number 245638/F50). An extension to include COAT Svalbard data is covered by a contribution of 1.9 mill NOK from RCN as part of a larger research infrastructure allocation to SIOS-KC (www.sios-svalbard.org) 2018-2022 (project grant number 269927). The latter also cover added cost related to interoperability between the COAT Data Portal and SIOS-KC. In addition, COAT member institutions contribute own funding particularly related to the cost of formatting, harmonising and documenting the data prior to uploading these to the COAT Data Portal.

The COAT Data Portal is developed by the COAT Data Management implementation team. This team is led by COAT researchers Nigel G. Yoccoz (UiT), and Jane Uhd Jepsen (NINA), and consists additionally of IT personnel from NINAs environmental data section (Roald Vang, Matteo De Stefano, Francesco Frassinelli) and the COAT project coordinator (Eeva Soininen, UiT). The implementation team receives input from other COAT personnel when needed. NINA is responsible for hosting and technical maintenance of the COAT Data Portal both during the implementation phase (2016-2020) and once it is operational (2021->).

Included in the terms for receiving support for research infrastructure from the Research Council of Norway, is the responsibility of the receiving institutions for maintaining the infrastructure for a minimum of 5 years after the end of the project period. This means that the costs for updating the datasets in the portal, and the technical maintenance, is guaranteed until the end of 2025 by contributing institutions (updating the data sets) and NINA (technical maintenance). Continued funding is sought by COAT personnel to ensure the long-term preservation of the COAT Data Portal.

4. Data security

Data is stored on NINA servers and transferred to users using an encrypted connection (HTTPS) to improve confidentiality of sensitive information. Authentication mechanism relies on Uninett Feide login system, which has been considered stable and safe. Custom code developed to support additional features with respect to what is provided by the software used (named CKAN) could pose potential security risks (like embargo and compressed archive creation). CKAN codebase and third-party modules could pose some risks too, as no security audit has been carried at the moment of writing. No warranty is provided to the users of the software.

Backups of files and databases are executed regularly based on the following table:

Media	**Retention**	**Cycle**	**Storage**
Snapshot	30 days	Daily	In house
Snapshot	90 days	Weekly	In house
Disk	6 months	Monthly	In house
Tape	5 years	Quarterly	Out of house

5. Ethical aspects

COAT manages certain sensitive data sets containing information on the location of breeding sites of protected species (see 2.4). Any such data are withheld from publication in the COAT Data Portal. The COAT data portal does not include personal data which require informed consent.

6. Annexes

The COAT data management plan contains the following 3 annexes:

Annex A: List of metadata fields.

Annex B: File formatting instructions for COAT tabular data.

Annex C: Description of data flow from raw data to state variables.

Annex A List of metadata fields.

Metadata field	Description	Example	DCAT equivalent (dcat:dataset)
Title*	Dataset title following COAT naming conventions	V_ungulates_pellets_diagonals_polmak	dct:title
Module*	COAT monitoring module to which the dataset is associated	Tundra-Forest Module	dct:publisher
Description	A text field with short description of key dataset attributes not computed in metadata fields	The data set contains the number of pellet groups observed along the diagonals in 24, 30x30 m plots in Polmak. Both diagonals are used (corner A - corner C, and corner B - corner D). Pellet groups are counted 1 meter on each side of the diagonal. Most observations are from reindeer and moose. Other herbivores (hare, ptarmigan) are included, but rarely observed.	dct:description
Keywords	A list of one to multiple keywords selected from a fixed taxonomy	Birch-forest, experimental-exclosure, forest, pellet-counts	dct:keyword
Topic Category Code*	INSPIRE TCC	Biota	
Embargo end date*	Date for making the current version of the dataset public. Max. embargo 24 months.	2021-10-21	
License*	Defaults to Creative Commons Attribution (CC BY 4.0). Any exceptions are handled on a case by case basis.	CC-BY-4.0	
Contact person*	Name of COAT contact person for the dataset	Jane Uhd Jepsen	dcat:contactPoint
Email address*	Email address for contact person	jane.jepsen@nina.no	vcard:fn
Title of position	Title of contact person	Senior researcher	
Organization name*	Affiliation of the dataset's Contact person, or Institution responsible for data dataset creation	NINA – Norwegian Institute for Nature Research	
Associated parties	List of other related institutions or organizations	UiT-Arctic University of Tromsø	
Persons	Names of other persons associated with the dataset (students, key collaborators)	Ole Petter L. Vindstad	
Temporal extent – start date	Start date for the dataset	2011-08-01	Dct:temporal Dct:PriodOfTime Schema:startDate
Temporal extent – end date	Current end date for the dataset	2019-08-01	Dct:temporal Dct:PriodOfTime Schema:endDate
Geographical location*	Name(s) of location where data have been	Varanger Varanger - Rakkonjarga	

	collected. A list of one to multiple location names selected from a fixed taxonomy	Varanger - Rakkonjarga - Polmak	
Scientific name	Name(s) of species concerned in the dataset. A list of one to multiple species names selected from a fixed taxonomy	Cervidae - Alces alces (moose/elg) Cervidae - Rangifer tarandus (reindeer/reinsdyr)	
Resource citation*	Citation for the dataset. Generated automatically	Jepsen et al., 2020, V_ungulates_pellets_diagonals_polmak_v1: COAT project data. Available online: https://data.coat.no/dataset/V_ungulates_pellets_diagonals_polmak v1	
Associated scripts	Link to associated data analysis scripts. Placed in the COAT github repository	https://github.com/COATnor/data_analysis_scripts/XXXXXX	
Associated study protocol	Link to the field protocol describing the study design and methodology used collecting the dataset. Placed in the COAT github repository	Jepsen & Vindstad, 2020, Polmak protocol v1. Available online: https://data.coat.no/protocols/Polmak protocol v1	
Bibliographic citation	Reference to publications using the dataset		
Funding source	Funding source(s) contributing to the dataset	Norwegian Research Council, Nordforsk, Fram Centre	

*Mandatory fields

Annex B File formatting instructions for COAT tabular data

Most of COATs data are in a tabular format. This includes all data which can be stored in a tabular format (.txt/.csv/.asc) format. We divide tabular data into two types, and we have defined a generic format for each of these two types:

Plot-based data: the sampling units are fixed in space (e.g. a plot, quadrat, transect etc), and the data contain 1-n variables measured at each of these points.

Individual-based data: the sampling units are individuals not fixed in space (e.g. animals) and the data contain 1-n variables measured on each of these individuals.

What is a data set in the COAT data portal?

A dataset in the COAT data portal is a collection of files *which are all represented by the same set of metadata.* If the collection of files you have at hand cannot be represented by the same set of metadata, you should consider splitting them into several collections (i.e. several data sets).

Each dataset in the portal is allocated a unique identifier (URL), and all search capabilities are on the metadata level of the data sets (see further info below).

A dataset could typically be *X* number of files containing measurements of the same variables in *X* number of years. Or *Y* number of files containing measurements of the same variables in *Y* number of localities.

A dataset can contain auxiliary files which contain information not directly part of the data set. For instance, an auxiliary file with spatial coordinates of all sampling sites in a standardized format is mandatory for all plot-based data. An example of a non-mandatory auxiliary file is a file containing a set of station-level variables measured at each plot/site/transect/quadrat. Typically, these variables would be such as terrain characteristics, habitat classes, land cover etc. Another example is a file defining first and last year when given sampling sites have been included in the study design. This is useful when the study design has changed over time.

Data set naming

The name of a dataset (e.g. a collection of files, see above) should always start with a standardized prefix indicating which region (Svalbard, Varanger) and which monitoring target they refer to. After the prefix the data set owner can define a name that is informative of the content. For instance, all data sets that describe the monitoring target *rodents* on Varanger are given the prefix '*V_rodents_*' while all data sets that describe the monitoring target *arctic fox* on Svalbard are given the prefix '*S_arcticfox_'*. A data set from *Varanger* which describes the monitoring target *forest* and contains data on tree structure could hence be called '*V_forest_treestructure'* or possibly '*V_forest_treestructure_Polmak'* if there is a need to distinguish between this data set and another from a different region or design. Below is a complete list of monitoring target (and prefixes).

Target	**Locality**	**Type**	**Datasetname_prefix**
Vegetation	Svalbard, Varanger	Biotic	S_vegetation_/V_vegetation_
Ungulates	Svalbard, Varanger	Biotic	S_ungulates_/V_ungulates_
Ptarmigan	Svalbard, Varanger	Biotic	S_ptarmigan_/V_ptarmigan_
Arctic fox	Svalbard, Varanger	Biotic	S_arcticfox_/V_arcticfox_
Red Fox	Varanger	Biotic	V_redfox_
Geese	Svalbard	Biotic	S_geese_
Rodents	Varanger	Biotic	V_rodents_
Insect defoliators	Varanger	Biotic	V_insect_defoliators_
Insect communities	Varanger	Biotic	V_insect_commun_
Specialist predators (rodents)	Varanger	Biotic	V_rodent_special_predators
Specialist predators (ptarmigan)	Varanger	Biotic	V_ptarmigan_special_predators
Generalist predators	Varanger	Biotic	V_general_predators_
Bird communities	Varanger	Biotic	V_bird_commun_

Tall shrub	Varanger	Biotic	V_tall shrub_
Meadow	Varanger	Biotic	V_meadow_
Forest	Varanger	Biotic	V_forest_
Heath	Varanger	Biotic	V_heath_
Forest_understorey	Varanger	Biotic	V_forest_understorey_
Snowbed	Varanger	Biotic	V_snowbed_
Timing of snow melt	Svalbard, Varanger	Climatic	S_timing_snowmelt_/V_timing_snowmelt_
Snow depth	Svalbard, Varanger	Climatic	S_snowdepth_/V_snowdepth_
Snow structure	Svalbard, Varanger	Climatic	S_snowstructure_/V_snowstructure_
Ground ice	Svalbard, Varanger	Climatic	S_groundice_/V_groundice_
Timing of icing	Svalbard, Varanger	Climatic	S_timing_icing_/V_timing_icing_
Weather	Svalbard, Varanger	Climatic	S_weather_/V_weather_
Energy Balance	Svalbard, Varanger	Climatic	S_energy_balance_/V_energy_balance_

File naming

The naming of the files within a dataset should inform about their content. However, there are no fixed naming conventions, and it is up to each data owner to ensure that file naming is consistent and informative within each data set. For some datasets it might be relevant to include locality names while for others it will be more relevant to name the file according to main variable and year.

The only file where there is a strict naming convention is the mandatory file with plot coordinates. This should be named according to the data set name + 'coordinates'. Example: for a dataset named 'V_forest_treestructure', the coordinate file should be named 'V_forest_treestructure_coordinates.csv'.

Special characters: avoid special characters (such as '()','#','&', '{}', ':', ';', '*') in all file names.

Coordinate files

All plot-based data sets must be accompanied by a file containing plot coordinates. The files have standard format to facilitate finding data from the same localities from the data portal.

Standard coordinates in COAT are decimal degrees and UTM zone 33. We provide the latter to facilitate the integration with national level map-based data where UTM33 Euref89 is the standard.

For some datasets there might be uncertainties attached to which datum was used. In most COAT relevant cases, any uncertainty as to whether the datum is Euref89 or WGS84 is irrelevant, as the difference is in the order of centimetres. However, if there is uncertainty as to whether the datum is ED50 or Euref89/WGS84, the difference can be substantial. In such cases, the data owner should judge whether this has implications for the use of the data, and specify any uncertainties in the Description field of the metadata. For example: "*Uncertainty related to datum: For the years 2000-present datum is Euref89. For the years preceding 2000, datum is suspected to be ED50, but this is uncertain. The potential displacement of coordinates due to this uncertainty is in the order of XX meters*".

The coordinate file should contain the following five columns in the given order:

[sn_siteID]: this is the waypoint of the plot. Must be the same name as used in the sn_siteID column in the data files (see section on *Columns names in data files*)

[e_dd]: X coordinate of the plot in longitude decimal degrees (WGS84 unless otherwise stated)

[n_dd]: Y coordinate of the plot in latitude decimal degrees (WGS84 unless otherwise stated)

[e_utm33]: X coordinate of the plot in UTM zone 33 (WGS84 unless otherwise stated)

[n_utm33]: Y coordinate of the plot in UTM zone 33 (WGS84 unless otherwise stated)

Column names in data files

As a general rule-of-thumb simple tables in COAT should be formatted in a long format, rather than in a wide format. This makes it easier to combine and plot data in R and facilitates the use of standardized column names. For instance a dataset with abundances measured on X number of species should have one column indicating species and one column indicating abundance, instead of X columns indicating abundance for species 1..X.

https://www.theanalysisfactor.com/wide-and-long-data/

http://www.cookbook-r.com/Manipulating_data/Converting_data_between_wide_and_long_format/

LIKE THIS:				NOT LIKE THIS:			
Plot	Species	Abundance		Plot	Species1	Species2	Species3
1	Species 1	5		1	5	12	0
2	Species 1	7		2	7	20	1
1	Species 2	12		..			
2	Species 2	20					
1	Species 3	0					
2	Species 3	1					
…							

Standardized column names have been defined to describe the spatial sampling hierarchy, the temporal sampling and the most commonly used value columns. All column names include a prefix: sn_(for spatial nested variables), sc_(for spatial crossed variables), t_ (for temporal variables), v_ (for variables containing other observations).

A complete list of these and a definition for each can be found at *Box/COAT/Data Management/Datatypes/Simple tables/Generic format data tables/Simple tables column definitions.xlsx* (internal users only). The file also contains information on which standard columns should be included in all datasets.

Spelling and general text formatting in data files

- Capital letters: used only for NA, not for any other purpose
- Scandinavian letters: replace Scandinavian letters ø, æ, å with ‘o’, ‘ae’, ‘aa’.
- Special characters: avoid special characters (such as ‘()’,’#’,’&’, ‘{}’, ‘:’, ‘;’, ‘*’) in all text columns.
- Separating words in ‘notes’/’comments’ columns: use space, underscore or comma.

Missing data in data files

Always indicate missing data, also in text columns, by NA (capital letters only). All observations that have no comments, should have value “NA” in the comment column.

Date formats in data files

Dates should always be given as YYYY-MM-DD for example 2018-12-31.

Locality names in data files

All locality names in files should conform to the standard lists of locality names found in *Box/COAT/Data Management/Taxonomy/Locality/Locality taxonomy COAT.xlsx* (internal users only). The list defines the spelling of all place names found at the top-four levels of the spatial hierarchy; sn_region, sn_subregion, sn_locality and sn_section. Note that the names confirm to the general rules of text formatting (no capital letters, no Scandinavian letters etc). Anyone in need of the proper spelling (for instance for plotting purposes) can consult the sheet “correct spelling” in the same file.

Species names in data files

All species names in files should conform to the standard taxonomy lists found on *Box/Data Management/Taxonomy/…* (internal users only). There are separate subfolders with lists for birds, insects, mammals, and plants. Each contains scientific names and common names in Norwegian and English where available. In addition, each species is given a unique abbreviated scientific name (3+3 letters, ex. *Branta leucopsis* = *bra_leu*, subspecies are 3+3+3 ex. *Lagopus muta hyperborea* = *lag_mut_hyp*). In order to make data sets as comparable and 'mergeable' as possible, always use the abbreviated name in species columns (e.g. the column 'v_species').

Taxonomic groups that include several species should also be included in the taxonomy lists. See table below for an example at genus level.

Family	Latin	Norwegian	English	Abbreviation
Soricidae	Sorex minutus	dvergspissmus	pygmy shrew	sor_min
Soricidae	Sorex sp	spissmus	shrew	sor_sp

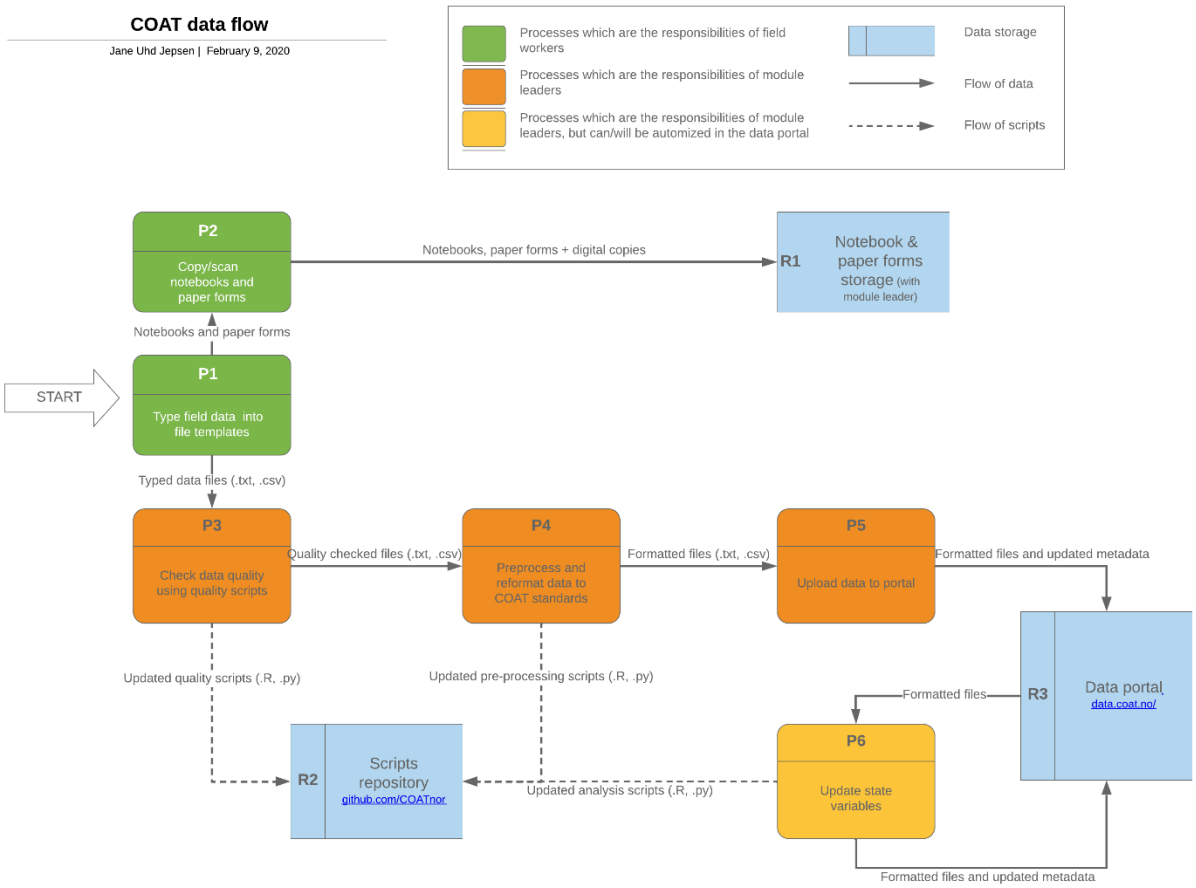
Try to use as detailed taxonomic grouping as possible. However, if the group is defined in a very specific manner, and difficult to include in this standard taxonomy table, include extra information in the metadata.

If you find any species missing, please add them to the list or inform Eeva Soininen (eeva.soininen@uit.no) or Jane Uhd Jepsen (jane.jepsen@nina.no).

Functional groups names in data files

All functional group names in files should be explained in auxiliary metadata files.

Annex C Description of data flow from field data to state variables



The data flow in COAT is described as a set of processes (P1 – P6) which involve either manual actions performed by COAT personnel, or automized actions performed within the data portal. Each process receives data (in-going arrows), make some changes to the data, and pass the data on (out-going arrows).

The COAT data management system consists of three different types of repositories or storage spaces: one for the original field recordings (notebooks, paper forms etc), one for the final data sets (the COAT Data Portal) and a github repository for all scripts used to transform the raw field data to their final format. In the github repository, scripts are divided into three types depending on whether the scripts perform initial quality checks of the data, mandatory pre-processing or reformatting of the data to meet COAT standards, or some form of analysis of the data in their final format.

The responsibilities of the different processes rests either with the field workers (ultimately the field leader) or with the data responsible for each module (the module leader or some person to which this responsibility has been delegated by the module leader). Ultimately, however, it is the responsibility of the module leader to ensure that all data from a given module are transferred to the COAT Data Portal following COAT standards. In the following we describe each repository and process in further detail.

Repositories (R)

R1: Notebooks and paper forms as they are received in from the field, are stored with the module leader after the data has been typed. Notebooks and paper forms should be kept in ordered collections ordered by year and/or site, so that it is possible to check suspected errors in the typed data against the field records. May also include storage of raw data files typed directly in the field, converted to pdf-files.

R2: All scripts related to COAT datasets should be placed on the COAT github repository (github.com/COATNor). Scripts should have a header explaining which datasets they relate to, the author of the script and they should be sufficiently documented with comments to allow other people to use them. We separate between three different kinds of scripts:

Data management scripts (https://github.com/COATnor/data_management_scripts). This contains scripts related to upload/download of data from the portal.

Data pre-processing scripts (https://github.com/COATnor/data_preprocessing_scripts). This should contain scripts for pre-processing of data sets. This includes all routines for quality checking of data prior to analysis.

Data analysis scripts (https://github.com/COATnor/data_analysis_scripts). This should contain scripts used for plotting and analysing the data, including estimation of state variables in the data portal.

The pre-processing scripts (quality check and formatting) are primarily for internal use within the module/within COAT. Scripts used for upload/download or data analysis and estimation of state variables are relevant for external users as well, and should be written with this in mind. The use of Rmarkdown reports to document data quality checking and data summaries is encouraged, and Rmarkdown reports can also be placed in the script folders.

Whenever a script is updated to a new version, it is recommended to create a new release in github, which guarantees a unique correspondence between a specific dataset version and the related scripts on github. See:

https://help.github.com/articles/creating-releases/

R3: The COAT Data Portal is the central data repository containing all datasets which are ready to use. Datasets can exist in the portal both as private and public versions. Internal users (from within COAT) can access both private and public datasets including all older versions. External users (anyone) can only access published datasets.

Processes (P)

P1: Process 1 starts when the data are brought in from the field either in notebooks, on paper forms, or, in some cases, on field tablets. The process is complete when all data are typed/converted into a predefined file format, checked for completeness and obvious typing errors, and delivered to the module leader.

When: ideally during field work (e.g. in the evenings) so that any mistakes made in recordings can be cleared up immediately. If this is not logistically possible, typing should be done immediately after the field work has been completed.

P2: Notebooks and paper forms should be scanned or photocopied at the end of the field work in order to minimize the risk of losing the original records. The process is complete when the notebooks, paper forms as well as the digital copy has been delivered to the module leader. Raw files from field tablets should be converted to pdf-files and delivered to the module leader.

When: immediately after completing the field work.

P3: The typed data should be quality checked prior to any preprocessing or analysis. The content of the quality check may depend on the nature of the data, but typically consists of checking for missing plots, duplicate plots, unlikely/non-permitted values which may be typing errors, consistency in spelling of locality and species names, etc. The use of Rmarkdown reports to document the quality checking is encouraged. The process is complete when the data can be considered complete, free of detectable errors, with consistent spelling and naming, and when updated quality checking scripts have been uploaded to the github repository.

When: As soon as possible after field work, and in time to reach the deadline for uploading the final datasets to the COAT Data Portal (see P5).

P4: In some cases data needs a certain amount of pre-processing before they can be used. This can for instance be adjustment of nomenclature against Artsdatabankens species lists (relevant for invertebrates). All preprocessing should be done using scripts (no manual adjustments). Further, most datasets will need reformatting to adhere to COAT generic file formatting. The process is complete when all data are reformatted and the updated scripts have been uploaded to the github repository.

When: As soon as possible after field work, and in time to reach the deadline for uploading the final datasets to the COAT Data Portal (see P5).

P5: The reformatted data should be uploaded to the COAT Data Portal as soon as they are ready. The process is complete when all datasets and their associated metadata are updated and a new version of the dataset has been published.

When: The deadline for uploading data collected during spring and summer is September 30th each year. This deadline ensures that all new data are include in the 4th quarterly backup to tape (see main document, section 4). An extended deadline for data collected during autumn is December 1st each year.

P6: The COAT State Variables are published in the COAT Data Portal as separate datasets, derived from primary data. The estimation of state variables from primary data is sometimes trivial (e.g. perhaps just a sum or a mean over primary data), and sometimes complex (e.g. estimated using multiple datasets and complex statistical models). However, whether trivial or complex, the estimation of all state variables must be documented in a script, uploaded to in the github repository. The updating of state variables can be automized within the data portal if updated scripts are available. The process is complete when updated analysis scripts are available on github, and a new version of all state variables are published.

When: As soon as a new version of the primary datasets are available in the data portal.