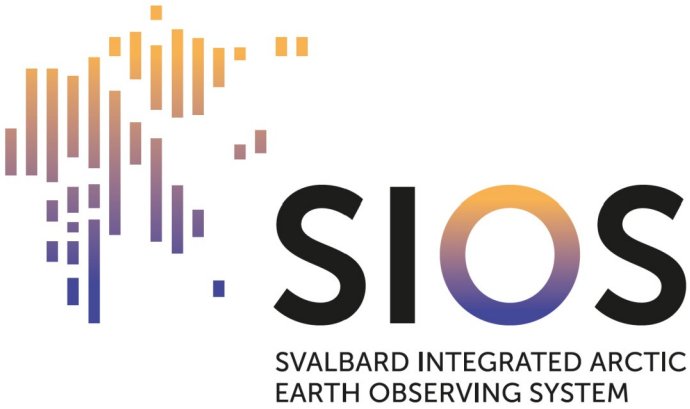

Guidelines for metadata and data sharing

21 June 2019

Revision history

Date	Comment	Responsible
2019-06-21	Added reference to discovery metadata webform (previously xls-sheet)	Christiane Hübner
2019-04-10	Added information on non standardised data and contributing data centres.	Øystein Godøy
2019-03-29	Minor updates and additions.	Øystein Godøy
2019-03-15	First draft	Øystein Godøy

Table of Contents

1 Introduction

Environmental and climate changes are currently observed at a global scale and in particular in the Arctic. In order to give better estimates of the future changes, the Arctic has to be monitored and analysed by a multi-disciplinary observation system which is suited to validate and gradually improve Earth System Models. The best chance to achieve significant results within a relatively short time frame is found in regions with a large natural climate gradient, and where processes sensitive to the expected changes are particularly important.

Svalbard and the surrounding ocean areas fulfil all these criteria: Svalbard is located in a region with a very large climate gradient, being alternately influenced by cold central Arctic or mild marine climate conditions at time scales of weeks to years. It is also located in the region with the strongest inflow and outflow processes between the Arctic and lower-latitude oceans. In addition, Svalbard is the only region in the world (and has the facilities) where one can study and quantify one of the remaining unknowns in the climate puzzle: the extraterrestrial and especially solar influence on climate.

The vision for the Svalbard Integrated Arctic Earth Observing System (SIOS) is to be a regional observational system for long term acquisition and proliferation of fundamental knowledge on global environmental change (GEC) within an Earth System Science (ESS) perspective in and around Svalbard. SIOS will systematically develop and implement methods for how observational networks are to be construed and thus become a leader regarding observational systems in the Arctic and Polar regions. The SIOS Data Management System (SDMS) Data Portal is the entry point to SIOS datasets, including the datasets used in the SESS reports. It offers a web interface that contains information about datasets (metadata). These metadata are harvested on a regular basis from data centres contributing to SIOS. These data centres manage the data on behalf of the owners/providers of the data.

The ambition of SIOS is to manage all data according to the FAIR principles. This implies that data are Findable, Accessible, Interoperable and Reuseable. This implies standardisation of information and interfaces at different levels.

2 Workflow

The basic work flow for data collected or used is:

1. Planning of data collection (including datasets reused from existing data centres)
    1. Ideally data documentation standards and quality assurance procedures are defined as part of the observation protocol for the data and implemented prior to data collection.
2. Data collection
    1. During data collection, the procedures and standards defined in the planning is applied.
    2. Application of standardised observation protocols and documentation standards at this level simplifies work at later stages and in particular during data documentation and publication.
3. Quality control

    1. All quality control procedures should be documented.
4. Data documentation
    1. According to the SIOS data policy all data should be self-describing according to standards. If no standard is available a detailed product manual must be provided and linked to the dataset when published. Datasets shall be reuseable in the long term, requiring open standards and formats to be used.
5. Data publishing
    1. Data publishing is done through one of the contributing data centres.
6. Long term data preservation
    1. Long term data preservation is the responsibility of the host data centre.

3 Data documentation

3.1 Introduction

The purpose of metadata is to ensure that data are easily found and used by users not necessarily familiar with the data. In order to achieve this we need data to be properly documented using discovery metadata and use metadata.

The purpose of discovery metadata is to make data searchable in the SDMS. Discovery metadata are similar to library index cards basically describing who measured what, where, when, where the data is located, how the data can be accessed and what can be done with them.

The purpose of use metadata is to describe the actual content of the data through standardised naming conventions (ontologies) for variables, units, missing values.

SIOS has some guidelines on these issues. Whenever data have been generated using the SIOS infrastructure, both discovery metadata and the actual data have to be submitted to a data centre integrated with the SIOS Data Management System. When non SIOS data are used in SIOS projects (e.g. SESS reports) and the data are already maintained in a proper data centre, discovery metadata have to be provided to the SIOS Data Management System.

3.2 Discovery metadata

SIOS operates a data discovery interface on the SIOS website. This is available at https://sios-svalbard.org/metadata_search. In order to feed this, discovery metadata are required.

The (generic) discovery metadata elements that SIOS requires are the following:

Element	Description
Title	A brief title for the dataset.
Abstract	A short summary of the process of developing the dataset, the content and potential linkages etc.
Principal investigator (PI)	Name of the PI.
PI email	Email to the PI
PI institution	Affiliation of the PI
PI address	Postal address of the PI

Dataset start time	IS8601 reference for the dataset YYYY-MM-DD
Dataset end time	IS8601 reference for the dataset YYYY-MM-DD
Dataset northernmost latitude	Geographical northernmost position of the dataset (rectangular box)
Dataset southernmost latitude	Geographical southernmost position of the dataset (rectangular box)
Dataset easternmost longitude	Geographical easternmost position of the dataset (rectangular box), positive east of Greenwich.
Dataset westernmost longitude	Geographical westernmost position of the dataset (rectangular box), negative west of Greenwich.
ISO Topic category	See list at the end or https://gcmd.nasa.gov/add/serfguide/iso_topic_category.html .
GCMD Science keywords	Select parameter descriptions from https://gcmdservices.gsfc.nasa.gov/static/kms/sciencekeywords/sciencekeywords.csv?ed_wiki_keywords_page, this provides a comma separated file which can be searched for parameters. The list is hierarchical, starting e.g. by atmosphere, ocean, cryosphere, climate indicators, biosphere, biological classification etc
Dataset production status	See code list at the end.
Access constraint	See code list at the end.
Activity type	See code list at the end.
Platform	Brief description, e.g. name of the Research Vessel, Sverdrup station, Hopen weather station, etc.
Related information (URL)	An URL to e.g. landing page for the dataset.
Related information (type)	e.g. “Landing page” or “Project homepage”.
Data access (URL)	An URL to the dataset.
Data access (type)	A specification of the data access protocol, e.g. HTTP (for direct download), OPeNDAP, OGC WMS.
Data center (URL)	URL to the data center hosting the data.
Dataset citation	A citation string for the dataset.
Dataset version	Version number for the dataset if available in multiple versions. Else use 1.
Project long name	Project name for the subproject (e.g. the SESS project), multiple names to be separated by comma.
Project short name	Acronym for the project.
Quality statement	Keywords indicating the quality control of the dataset (Not quality controlled, Quality controlled by statistical methods, Quality controlled by manual inspection)
License	SIOS recommends Creative Commons BY which ensures proper attribution of data providers.

If this information is not provided by the host data centre for existing datasets the PI of the project has to fill this information into following webform: https://sios-svalbard.org/metadata-collection-form (login required). The information can then be further processed by the SIOS Data Management System. If the host data centre provides this information as GCMD DIF or ISO19115, the PI must provide the reference to this information to the SIOS Data Management system for inclusion in the searchable index.

Once discovery metadata are available, the datasets in question can be discovered in the SIOS Data Management System.

3.3 Use metadata

3.3.1 Introduction

Discovery metadata mentioned above are the first step towards FAIR data, addressing the issue of Findability. However, SIOS also need to address Accessibility, Interoperability and Reuseability. In order to achieve this standardisation is essential. For some disciplines/parameters there are self-describing standards that are applicable while for other parameters this is not available. The SIOS approach is to use standards wherever possible and to extend documentation where no standard exist (or is not mature enough).

3.3.2 Biology

For biological information SIOS recommends that information is encoded according to the Darwin Core Archive format. Darwin Core is a flexible standardised FAIR compliant framework for managing biodiversity data. It is not possible to establish services on top of the data like for NetCDF/CF mentioned below, but the benefits of the community standardisation effort far exceeds the shortcomings and datasets may be contributed to GBIF as well as SIOS. Darwin Core Archive format is essentially a collection of comma separated variables text files with a simple description file (meta.xml) describing how the other files are to be interpreted. Files are collected in ZIP or GZIP files.

3.3.3 Meteorology, oceanography, glaciology

For these domains usage of NetCDF according to the Climate and Forecast convention is recommended. For SIOS core data in these domains, NetCDF/CF is required. The reason for this is that it provides a standardised documentation that is machine readable on which SIOS can build services. SIOS has established services that allows users to convert comma separated files to NetCDF/CF and to validate the CF compliance of the files. The compliance checker allows checking of global attributes according to the Attribute Convention for Dataset Discovery. If NetCDF/CF is provided with ACDD attributes, the SIOS Data Management System can extract discovery metadata directly from the dataset, publish and preserve the data.

For NetCDF/CF it is important to use the featureType attribute of CF for all data that is not gridded. The following types of data has been identified within SIOS:

- Time series at weather stations should e.g. follow the approach described in http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.html#_single_time_series_including_deviations_from_a_nominal_fixed_spatial_location
- Profiles, e.g. CTD or radiosoundings, should follow the approach described in http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.html#_single_profile
- Timeseries of profiles should follow the approach described in http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.html#_time_series_of_profiles_at_a_single_station
- Trajectories e.g. drifters, should follow the approach described in http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.html#_single_trajectory

- Trajectories of profiles should follow the approach described in http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.html#_profiles_along_a_single_trajectory

For all the examples provided above, please remember to add ACDD elements.

As a general guideline, datasets should be kept at the highest possible granularity, e.g. do not combine e.g. multiple weather stations or CTD stations in one data file, to ensure flexibility when building services on top of the data.

A very important feature of the CF standard is to choose the appropriate CF standard name for the variable. CF standard names are searchable at http://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html. If no appropriate standard name is found for the variable in question, please contact the SIOS Data Management Service for guidance and leave the standard name element open. Instead, use a descriptive long name.

An example of a CF compliant file structure containing quality controlled weather information is provided below.

```
netcdf                    SN99938                    {
dimensions:
       time     =     UNLIMITED     ;     //     (65003     currently)
variables:
       double                 time(time)                 ;
              time:standard_name         =         "time"         ;
              time:long_name     =     "Time     of     measurement"     ;
              time:calendar          =          "standard"          ;
              time:units  =  "seconds  since  1970-01-01  00:00:00  UTC"  ;
              time:axis            =            "T"            ;
       double                  latitude                 ;
              latitude:standard_name       =       "latitude"       ;
              latitude:long_name        =        "latitude"        ;
              latitude:units        =        "degree_north"        ;
       double                 longitude                 ;
              longitude:standard_name       =       "longitude"       ;
              longitude:long_name        =        "longitude"        ;
              longitude:units        =        "degree_east"        ;
       float           air_pressure_at_sea_level(time)          ;
              air_pressure_at_sea_level:long_name = "Air pressure at sea
level
"                                  ;
              air_pressure_at_sea_level:standard_name                    =
"air_pressure_at_sea_l
evel"                                  ;
              air_pressure_at_sea_level:unit       =       "Pa"       ;
       float           surface_air_pressure_2m(time)           ;
              surface_air_pressure_2m:long_name = "Air pressure at station
lev
el"                                  ;
              surface_air_pressure_2m:standard_name                      =
"surface_air_pressure"                              ;
              surface_air_pressure_2m:unit       =       "Pa"       ;
       float             air_temperature_2m(time)             ;
              air_temperature_2m:long_name   =   "Air   temperature"   ;
              air_temperature_2m:standard_name   =   "air_temperature"   ;
              air_temperature_2m:unit         =         "K"         ;
       float          air_pressure_at_sea_level_qnh(time)          ;
```

```
               air_pressure_at_sea_level_qnh:long_name   =   "Air   pressure
(QNH)"                                    ;
               air_pressure_at_sea_level_qnh:standard_name                 =
"air_pressure_at_s
ea_level"                                    ;
               air_pressure_at_sea_level_qnh:unit     =     "Pa"     ;
       float              wind_speed_10m(time)             ;
               wind_speed_10m:long_name    =    "Mean    wind    speed"    ;
               wind_speed_10m:standard_name     =     "wind_speed"     ;
               wind_speed_10m:unit       =       "m       s-1"       ;
       float           wind_from_direction_10m(time)           ;
               wind_from_direction_10m:long_name  =  "Wind  direction"  ;
               wind_from_direction_10m:standard_name                       =
"wind_from_direction"                          ;
               wind_from_direction_10m:unit      =      "degree"      ;
       float             relative_humidity(time)             ;
               relative_humidity:long_name  =  "Relative  air  humidity"  ;
               relative_humidity:standard_name  =  "relative_humidity"  ;
               relative_humidity:unit        =        "1"        ;

//                   global                    attributes:
               :wigos           =           "unknown"           ;
               string      :station_name      =      "KVITØYA"      ;
               :wmo_identifier          =          "01011"          ;
               :date_created   =   "2019-01-31T12:02:02.221824+00:00"   ;
               :time_coverage_end     =     "2019-01-31T12:00:00"     ;
               string   :title   =   "Observations   from   station   KVITØYA
SN99938"                                    ;
               :metadata_link         =         "https://oaipmh.met.no/oai/?
verb=GetRecord&meta
dataPrefix=iso&identifier=SN99938"                          ;
               :acknowledgment       =       "frost.met.no"       ;
               :comment = "Observations based on data from frost.met.no" ;
               :institution  =  "Norwegian  Meteorological  Institute"  ;
               :featureType         =         "timeSeries"         ;
               :id          =          "metno_obs_SN99938"          ;
               :references            =             ""            ;
               :geospatial_lat_min       =       "80.105800"       ;
               :Conventions       =       "ACDD-1.3,CF-1.6"       ;
               :creator_name  =  "Norwegian  Meteorological  Institute"  ;
               :keywords         =         "observations"         ;
               :history  =  "2019-01-31T12:02:02.221824+00:00:  frost  write
netcdf
"                                      ;
               :creator_url       =       "https://met.no"       ;
               :geospatial_lon_max       =       "31.464300"       ;
               :summary  =  "Surface  meteorological  observations  from  the
observa
tion network operated by the Norwegian Meteorological Institute. Data are
receiv
ed and quality controlled using the local KVALOBS system. Observation
stations                                                                  a
re normally operated according to WMO requirements, although specifications
are
not followed on some remote stations for practical matters. Stations may
have                                                                      mo
```

```
re     parameters     than     reported     in     this     dataset."     ;
                :geospatial_lon_min       =       "31.464300"       ;
                :geospatial_bounds   =   "POINT(31.464300   80.105800)"   ;
                :geospatial_lat_max       =       "80.105800"       ;
                :creator_email      =      "observasjon@met.no"      ;
                :geospatial_bounds_crs       =       "latlon"       ;
                :source    =    "Meterological    surface    observation    via
frost.met.no"                                  ;
                :time_coverage_start    =    "1996-01-01T03:00:00"    ;
                :wigos_identifier        =        "unknown"        ;
}
```

3.3.4 Other datasets

If the above mentioned standards are insufficient the fallback solution is to document and publish data in a form that is possible to preserve and decode over time. Essentially this implies using a text based format (GeoJSON, XML) for instrument data and JPEG for images. Text based formats should have proper metadata embedded, but in addition a comprehensive product manual is required in PDF format. The data and the product manual is combined using ZIP, GZIP or tar. All documentation should be written for people without detailed knowledge on the dataset, but with domain knowledge to understand the content.

4 Data deposition

SIOS has a number of partner repositories that are contributing to the SIOS Data Management System. These are listed below. Please contact the SIOS Data Manager for guidance on which partner repository to connect with for data deposition.

Data centre	N[1]	URL	Contact	Comment
Arctic Data Archive System	JP	https://ads.nipr.ac.jp/	Hironori Yabuki	Discovery metadata interoperability interfaces are available.
Arctic Data Centre	NO	http://arcticdata.met.no/	Øystein Godøy	Discovery metadata interoperability interfaces are available. This subsystem is a WMO Information System Data Collection and Production Center and is in the process of supporting DOI. Interoperability interfaces for meteorology, oceanography and glaciology datasets are supported.
AWI/PANGAEA	DE	http://pangaea.de/	Please involve the SIOS Data Manager if depositing data here.	Discovery metadata interoperability interfaces are available. This subsystem is a member of ICSU World Data System and offers DOI.
Italian Arctic Data	IT	http://arcticnode.dta.cnr.it/cnr	Angelo	Discovery metadata

1 Nation.

Data centre	N	URL	Contact	Comment
Center (IADC)		/index.php	Viola	interoperability interfaces are available. Interoperability interfaces for meteorology, oceanography and glaciology datasets are supported.
NILU	NO	http://ebas.nilu.no/	Markus Fiebig	Discovery metadata interoperability interfaces are available. Interoperability interfaces for atmospheric composition datasets are under implementation.
Norwegian Marine Data Centre	NO	http://www.nmdc.no/	Terry Hannant	Discovery metadata interoperability interfaces are available. Interoperability interfaces for atmospheric composition datasets are under implementation.
Norwegian Polar Institute	NO	http://data.npolar.no/	Stein Tronstad	Discovery metadata interoperability interfaces are available. This archive supports DOI and is also supporting Darwin Core Archives.
Norwegian Infrastructure for Research Data	NO	https://archive.sigma2.no/	Please involve the SIOS Data Manager if depositing data here.	Discovery metadata are maintained external to this system, the system has DOI support and application servers for meteorological, oceanographical and glaciological data are under implementation.

5 Existing datasets

Frequently project supported by SIOS do reuse existing data. In this context data already hosted by a proper data centre adhering to the FAIR guiding principles mentioned above and serving data online should not be duplicated but rather being pointed towards in the SIOS Data Management search interface. In order to achieve this discovery metadata are required, either through a manual process (filling out a spreadsheet) or by identification of the proper discovery metadata (GCMD DIF or ISO19115 as mentioned above) by the host data centre. It is the responsibility of the SIOS project to gather this information and to communicate it to the SIOS Data Manager. It is important to remember that data should be available online.

6 Code lists

6.1 ISO Topic Category

farming	rearing of animals or cultivation of plants. For example, resources

	describing irrigation, aquaculture, herding, and pests and diseases affecting crops and livestock.
biota	naturally occurring flora and fauna. For example, resources describing wildlife, biological sciences, ecology, wilderness, sea life, wetlands, and habitats.
boundaries	legal land descriptions.
climatologyMeteorologyAtmosphere	atmospheric processes and phenomena. For example, resources describing cloud cover, weather, atmospheric conditions, climate change, and precipitation.
economy	economic activities or employment. For example, resources describing labor, revenue, commerce, industry, tourism and ecotourism, forestry, fisheries, commercial or subsistence hunting, and exploration and exploitation of resources such as minerals, oil, and gas.
elevation	height above or below sea level. For example, resources describing altitude, bathymetry, digital elevation models, slope, and products derived from this information.
environment	environmental resources, protection, and conservation. For example, resources describing pollution, waste storage and treatment, environmental impact assessment, environmental risk, and nature reserves.
geoscientificinformation	earth sciences. For example, resources describing geophysical features and processes, minerals, the composition, structure and origin of the earth's rocks, earthquakes, volcanic activity, landslides, gravity information, soils, permafrost, hydrogeology, and erosion.
health	health services, human ecology, and safety. For example, resources describing human disease and illness, factors affecting health, hygiene, mental and physical health, substance abuse, and health services.
imageryBaseMapsEarthCover	base maps. For example, resources describing land cover, topographic maps, and classified and unclassified images.
intelligenceMilitary	military bases, structures, and activities. For example, resources describing barracks, training grounds, military transportation, and information collection.
inlandWaters	inland water features, drainage systems, and their characteristics. For example, resources describing rivers and glaciers, salt lakes, water use plans, dams, currents, floods, water quality, and hydrographic charts.
location	positional information and services. For example, resources describing addresses, geodetic networks, postal zones and services, control points, and place names.
oceans	features and characteristics of salt water bodies excluding inland waters. For example, resources describing tides, tidal waves, coastal information, and reefs.
planningCadastre	land use. For example, resources describing zoning maps, cadastral surveys, and land ownership.
society	characteristics of societies and cultures. For example, resources describing natural settlements, anthropology, archaeology, education, traditional beliefs, manners and customs, demographic data, crime and justice, recreational areas and activities, social impact assessments, and census information.

structure	man-made construction. For example, resources describing buildings, museums, churches, factories, housing, monuments, and towers.
transportation	means and aids for conveying people and goods. For example, resources describing roads, airports and airstrips, shipping routes, tunnels, nautical charts, vehicle or vessel location, aeronautical charts, and railways.
utilitiesCommunications	energy, water and waste systems, and communications infrastructure and services. For example, resources describing hydroelectricity, geothermal, solar, and nuclear sources of energy, water purification and distribution, sewage collection and disposal, electricity and gas distribution, data communication, telecommunication, radio, and communication networks.

6.2 Dataset production status

Code	Description
Planned	Refers to data sets to be collected in the future and are thus unavailable at the present time. For Example: The Hydro spacecraft has not been launched, but information on planned data sets may be available.
In Work	Refers to data sets currently undergoing production or data that is continuously being collected or updated. For Example: data from the AIRS instrument on Aqua is being collected continuously.
Complete	Refers to data sets in which no updates or further data collection will be made. For Example: Nimbus-7 SMMR data collection has been completed.
Obsolete	A new version of the dataset has been generated. The new version should be used, this is kept for back tracing.

6.3 Access constraint

Code	Description
Open	These data are freely available to everyone and can be made available without any restrictions. This is the recommended and required status for SIOS. Justifications are required if this is not used.
Registered users only (automated approval)	These data are available for users as long as they register with name, affiliation, and a valid email address. The verification process can be automated.
Registered users only (manual approval required)	These data are available for users as long as they register with name, affiliation, and a valid email address. The verification process must be manual.
Restricted to a community	These data are available for users within a restricted community. This community determines the authorization mechanism to utilise whether this being IP-address, community specific users names or other.
Restricted access to metadata	Information on these data MUST NOT be exposed externally. This implies that neither metadata can be exposed externally.

6.4 Activity type

Code	Description
Aircraft	Observations made during a flight trajectory. The observations can include remote sensing instruments, drop sondes or in situ measurements. Both manned and unmanned vehicles are covered by this term. The output is typically a trajectory, but could also be profiles or points.
Space Borne Instrument	Observations or analysed products based upon data from a space-borne instrument (typically onboard a satellite). The nature of the output is typically gridded of type imagery or profiles.
Numerical Simulation	Data are generated by the use of a numerical simulation of the atmosphere, the ocean, the climate or similar. Statistical analysis is not covered by this.
Climate Indicator	This indicates a dataset that has been generated by analysis of some data with the emphasis on being representative in a climate context (e.g. consistent in time). Furthermore, a climate indicator is a "compact" representation of the feature studied (e.g. the temporal evolution of area covered by sea ice in the Arctic). Climate Indicators are frequently linked to GCOS requirements.
In Situ Land-based station(Land station) (Field Experiment)	This is used to tag datasets generated from a site located on land. This can be a permanent (e.g. a SYNOP or TEMP station) or a temporary site (e.g. a field experiment).
In Situ Ship-based station(Cruise)	This is used to identify datasets generated during cruises. Typically it describes a full dataset generated in a context, possibly describing both ocean and atmospheric conditions.
In Situ Ocean fixed station(Moored instrument)	This is used to describe ocean stations that are fixed in space. Typically this is moorings, anchored buoys, oil rigs etc.
In Situ Ocean moving station(Float)	This is used to describe ocean stations that are moving around. Typically this is gliders and drifting buoys.
In Situ Ice-based station(Ice station)	This is used to tag datasets generated from a site located on drifting sea ice or some other ice sheet (possibly on land but moving). It typically describes a temporary site (e.g. a field experiment). It would also be used to describe ships frozen in ice and drifting e.g. across the Arctic as well as Ice Tethered Platforms (ITP) and Ice Mass Balance Buoys (IMBB).
Interview/Questionnaire(Interview) (Questionnaire)	This is not much used within environmental science, but comes in useful sometimes. It is used to cover the results of interviews and questionnaires especially in interdisciplinary science.
Maps/Charts/Photographs (Maps) (Charts) (Photographs)	This is used to tag datasets containing imagery or PDF documents. This could e.g. be a time lapse photographic session of a specific site illustrating e.g. snow cover or cloud cover. It can also be used to tag documents or maps describing the nature of a field station. It would then require datasets to be linked (which currently is not supported).