Identification and citation in SEADATANET

Context of identification and citation in SEADATANET

Summary of SEADATANET requirements for identification and citation

Detailed requirements

IDENTIFICATION

1. What granularity do your RI's data products have:

a) Content-wise (all parameters together, or separated e.g. by measurement category)?

Products gather parameter category, e.g. temperature and salinity of the water column are in a single product. Observation data sets also managed by the infrastructure are managed individually per sampling features (e.g. vertical profile, time series, trajectories, …), per observing platform.

b) Temporally (yearly, monthly, daily, or other)?

Seadatanet provides compilation of data sets over decades for climatological study purpose.

c) Spatially (by measurement station, region, country or all together)?

The products group observations spatially, by sea basin (e.g. Black Sea, Baltic Sea, Arctic Ocean, North Atlantic Ocean, ...)

2. How are the data products of your RI stored - as separate "static" files, in a database system, or a combination?

The products are separated static files. The input observations also managed by the infrastructure are managed heterogenously in databases or files. The harmonization is done thanks to web services on top of the datasets.

3. How does your RI treat the "versioning" of data - are older datasets simply replaced by updates, or are several versions kept accessible in parallel?

For product version is managed (DOI are minted). For observation only the best (latest) copy is managed.

4. Is it important to your data users that

a) Every digital data object is tagged with a unique & persistent digital identifier (PID)?

Yes

b) The metadata for data files contains checksum information for the objects?

Yes if it enables to detect quality control updates in the dataset.

c) Metadata (including any documentation about the data object contents) is given its own persistent identifier?

No

d) Metadata and data objects can be linked persistently by means of PIDs?

Yes

5. Is your RI currently using, or planning to use, a standardized system based on persistent digital identifiers (PIDs) for:

a) "Raw" sensor data?

A central system called Common Data Index (CDI) delivers identifiers for observations. However the identifier is delivered 3 to 5 years after observation is done. Local identifiers are also used by the data centres of the network. We are looking into using UUID for observations or observing platforms. If one UUID is associated with each platform, then the identification of observations dataset is going to be eased. We are looking into OGC/PUCK standard interfaces to get these unique identifiers.

b) Physical samples?

IGSN is sometimes used for geological sampling.

c) Data undergoing processing (QA/QC etc.)?

No

d) Finalized "publishable" data?

Yes product have UUID and DOIs

6. Please indicate the kind of identifier system that are you using - e.g. Handle-based (EPIC or DOI), UUIDs or your own RI-specific system?

See above

7. If you are using Handle-based PIDs, are these handles pointing to “landing pages”? Are these pages maintained by your RI or an external organization (like the data centre used for archiving)?

DataCite DOIs are defined after automated UUID, for example: http://dx.doi.org/10.12770/2a5c1396-f832-4500-8faa-8cfeeded1ebb

8. Are costs associated with PID allocation and maintenance (of landing pages etc.) specified in your RI’s operational cost budget?

No, cost are shared by differents infrastructures and small regaring the overall budget.

CITATION

9. How does your “designated scientific community” (typical data users) primarily use your data products? As input for modelling, or for comparisons?

Comparison, local study

10. Do your primary user community traditionally refer to datasets they use in publications:

a) By providing information about producer, year, report number if available, title or short description in the running text (e.g. under Materials and Methods)?

--

b) By adding information about producer, year, report number if available, title or short description in the References section?

--

c) By DOIs, if available, in the References section?

Would like to push for using DOIs.

d) By using other information?

--

11. Is it important to your data users to be able to refer to specific subsets of the data sets in their citation? Examples:

a) Date and time intervals

b) Geographic selection

c) Specific parameters or observables

For SeaDatNet strictly speaking the priority is to enalbe citation of the whole datasets which is static. If we extend the scope to marine data management, subsetting is important (x,y,t, observed properties) but, in case of citation, no as much as snapshot tag, or quality assessment method applied to dataset which can evolve continuously. There is a challenge to improve this.

12. Is it important to be able to refer to many separate datasets in a collective way, e.g. having a collection of “all data” from your RI represented by one single DOI?

Datasets are compiled together in a specific process to harmonized data for specific usage (e.g. cliamtology). Then the product is cited and has its own DOI. Not a priority to cite on-demand compilation yet.

13. What strategy does your RI have for collecting information about the usage of your data products?

a) Downloads/access Authentication of data access (user directory).

Log analysis.

b) Visualization at your own data portal .

Non authenticated Log analysis

c) Visualization at other data portals

Non authenticated. Log and refererer analysis

d) References in scientific literature

“Manual” survey done with the support a library team.

e) References in non-scientific literature

“Manual” survey done with the support a library team.

f) Scientific “impact”

“Manual” survey done with the support a library team.

14. Who receives credit when a data set from your RI is cited?

Hereafter is what should be done, not what is available

a) The RI itself

yes

b) The RI’s institutional partners (all or in part, depending on the data set contents)

yes, with weight or specific feedback (dashboards) related to the contribution of each.

c) Experts in the RI’s organization (named individuals)

no (data managers or IT experts should be transparent, neutral in the process of data delivery)

d) “Principal investigators” in charge of measurements or data processing (named individuals)

yes, they are the one contributing to the RI by providing data

e) Staff (scientists, research engineers etc.) performing the measurements or data processing (named individuals)

yes, they are the one contributing to the RI by providing data

15. What steps in tooling, automation and presentation do you consider necessary to improve take up of identification and citation facilities and to reduce the effort required for supporting those activities?

Standardisation of data citation by the publishers. (on-going) Develop a full framework from user registration to usage analysis (download, view) via proper business oriented activity log and “customer relationship management”.

Formalities (who & when)

Go-between	?? Info added by topic coordinator Maggie Hellström
RI representative	??
Period of requirements collection	Nov 2015 - December 2015
Status	Requirements collection completed