Identification and Citation for AnaEE

Context of identification and citation in AnaEE

AnaEE is still in its preparatory phase, and therefore - as pointed out by the AnaEE representatives - it should be noted that many of the questions on this topic could only be answered in a very preliminary way.

Summary of AnaEE's requirements for identification and citation

Detailed requirements

The following information was contributed via e-mail by the RI representatives directly to the topic coordinator Maggie Hellström.

IDENTIFICATION

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

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

The data are collected into distributed site data bases. Some of them may be gathered at the national level. A querying interface allows to get data flexibly at different level (from a parameter to the whole data of given site/experiment or even from different sites). We have two kind of data sets :

- from long term experiment where the data are collected in a site data base.

- from short term experiment as in controlled conditions in ECOTRON where the data are gathered in a project data base.

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

Yearly, monthly, daily, hourly and sometimes at higher temporal resolution

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

By measurement station or a network of stations. We don't produce data products representative of an area whatever the scale.

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

Mainly in Data Base information systems.

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? How do you identify different version of the same dataset?

Not yet addressed. We start the data production at the France level. It is intended to expose the latest updates on the data base system. However published data will be versioned according to the update of their content.

4) Is it important to your data users that:

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

Not yet. It is intended to have PID at the data set level (a site , an experiment). Not mature for finer descriptions (eg parameter , variable …). However we are working on the annotation of the data using a ontological approach which would lead to unique identification of every parameter.

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

not yet applicable

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

Not yet. Will be using DOI

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

not yet applicable

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

a) "Raw" sensor data?

Not yet decided. However there will be a strong probability that raw data will be stored together with processed data in order. The aim is to make reprocessing possible by users.

b) Physical samples?

Not yet implemented. It is planned to annotate persistently the different objects on which observations are made (soil sample, soil layer, plot, tree, animal…)

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

Not yet implemented. It is intended to define different levels of processing (L0, L1, L2, L3 …) and have an array with quality code. Some of the level (not necessarily all) will need to have a PID

d) Finalized “publishable” data?

Not yet decided.

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?

Not yet . Our plan is to use DOI for published data set and or own specific system for the description at the parameter level..

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

Not yet decided.

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

Not yet adressed

CITATION

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

Both

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)?

Yes in material and method, with appropriate reference and appropriate acknowledgement

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

See previous

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

Not widely yet, But could be used

d) By using other information?

No other known practices

e) By providing the data as supplementary information, either complete or via a link

Yes

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

yes

b) Geographic selection

yes

c) Specific parameters or observables

yes

d) Other

Data quality, accuracy,

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?

Yes at a site level or for an experiment that produced several datasets. Not necessarily to the whole RI

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

Not yet fully defined

It is expected to have a registration of users (account in the Information System), download tracking, identification in scientific publication, citation (DOI, publication/report transmission …)

a) Downloads/access requests

Yes, access requests

b) Visualization at your own data portal

Not yet defined

c) Visualization at other data portals

No.

d) References in scientific literature

Yes

e) References in non-scientific literature

Yes if easily collected

f) Scientific "impact"

To be defined

14) Who receives credit when a dataset from your RI is cited?

a) The RI itself

Yes

b) The RI's institutional partners (all or in part, depending on the dataset contents)

Yes

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

no

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

yes.

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

yes

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?

How to deal with incremental datasets?

How to link annotation on ontology and PID?

Formalities (who & when)

Go-between	?? Questionnaire response received by topic coordinator Maggie Hellström
RI representative	Christian Pichot and André Chanzy

Period of requirements collection	March 2016
Status	Information gathered, no analysis done yet