Identification and citation in EMSO

Context of identification and citation in EMSO

Summary of EMSO requirements for identification and citation

Detailed requirements

The following information was contributed via e-mail by the RI representative 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)?

Yes, e.g. the MOIST system which holds some INGV EMSO data delivers data/metadata according to instrument types

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

Yes, and in combination with c) e.g. the EMSO data stored at Ifremer/Eurosites provides data per site and in yearly or monthly snapshots or per deployment. See also via the new sextant publishing tool: http://sextant.ifremer.fr/record/02c4b294-f3e6-4760-bc67-f401a583f475/

Same at PANGAEA which holds data per site and deployment or recovery period or splitted into e.g. monthly snapshots.

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

Yes, see above b)

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

Depends on the data archive in the distributed system. We have all options there, database, filesystem or combination

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?

There is no commonly implemented solution to this problem in EMSO, but in general the approach is to keep archived dataset static. We at PANGAEA currently work on a solution together with DataCite within the THOR project.

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?

This would be interesting fo the responsible archive to keep integrity but I never heard this demand from our users or from users of other system.

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

Yes

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

DOIs used in EMSO (PANGAEA/Ifremer) are linked to the data as digital object not to the metadata. We do not treat metadata and data as different identification objects

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

a) "Raw" sensor data?

No

b) Physical samples?

Not really discussed in EMSO. For PANGAEA: Yes (but only for a very small fraction of the data -> ISSN)

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

NO

d) Finalized “publishable” data?

Yes, DOIs are already used by Ifremer (Sextant) and PANGAEA

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?

DOI

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)?

Yes landing page are in place which are maintained by the responsible data center

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

NO, costs (in as far these occur) are covered by the responsible data archives

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:

Please note: it is not possible to give sufficiently based answers to the questions below as no analysis on the current habits of data usage has been performed. EMSO would like to have data properly cited but actually we do not know in how far this is done..

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?

--

d) By using other information?

--

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

--

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

No, not yet

a) Date and time intervals

--

b) Geographic selection

--

c) Specific parameters or observables

--

d) Other

--

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 in as far this is offered by the data archive. E.G PANGAEA provides collections of data sets within a single DOI

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

There is no common approach to this issue, usage tracking is performed by each data center individually

a) Downloads/access requests

--

b) Visualization at your own data portal

--

c) Visualization at other data portals

--

d) References in scientific literature

--

e) References in non-scientific literature

--

f) Scientific "impact"

--

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

This depend on the offered data citation but usually this is PI or author centric

a) The RI itself

--

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

--

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

Yes

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?

I do not know what you mean here

Formalities (who & when)

Go-between	?? Questionnaire response received by topic coordinator Maggie Hellström
RI representative	Robert Huber
Period of requirements collection	Nov 2015 - Dec 2015
Status	Information gathered, no analysis done yet