|  |
| --- |
| **Requirements survey topics:**   1. General questions 2. ***Identification and citation*** 3. Curation 4. Cataloguing 5. Processing 6. Provenance 7. Optimization 8. Community support |

**ENVRIplus Theme 2:**

Requirements information gathering exercise

***ICOS (Integrated Carbon Observation System)***

*RI representative(s):*

* *Margareta Hellström,*

*ICOS Carbon Portal & Lund University*

This version from January 27, 2016.

**1. Identification and citation**

**IDENTIFICATION**

1. What granularity do your RI’s data products have:
   1. Content-wise (all parameters together, or separated e.g. by measurement category)?

The observational data (1-dimensional time series) typically are considered per measurement category, i.e. as sets of observables that belong together (concentrations, fluxes, meteorological or a combination of all).

The elaborated products are typically collections of multi-dimensional (time and space) parameter arrays, that are co-stored in a single (NetCDF) data object.

* 1. Temporally (yearly, monthly, daily, or other)?

Traditionally, ICOS-type observational data have been considered as yearly datasets - but the data itself mainly consists of time series at half-hourly or hourly resolution.

For elaborated products, datasets may cover periods ranging from some months to several years. The temporal resolution varies accordingly.

* 1. Spatially (by measurement station, region, country or all together)?

By measurement station, and possibly also by country.

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

Raw observational data: as individual files, and incorporated into databases at the Thematic Centers.

Intermediate stage observational data: mainly incorporated into databases.

Final (QA/QC:ed) observational data products: aggregated into static files (for archiving) and incorporated into databases at the Thematic Centers and at the Carbon Portal.

Elaborated data products: (delivered as) static files, but may be transferred into databases.

1. 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?

Depends on the circumstances. All datasets will be given a persistent unique identifier. For datasets stored as static files, versions are distinguished by the given dataset “name” and PID. The landing page of an object will contain info on parent (if any) or “replacement” (or child) object relationships. For datasets stored in “dynamic” databases, every update (addition, deletion or edit) will be time-stamped - making it possible to extract the “state” of the dataset at any given point in time.

1. Is it important to your data users that:
   1. Every digital data object is tagged with a unique & persistent digital identifier (PID)?

Yes.

* 1. The metadata for data files contains checksum information for the objects?

Yes, very. This is helpful to identify tampering.

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

Not critical, but desirable.

* 1. Metadata and data objects can be linked persistently by means of PIDs?

Very important.

1. Is your RI currently using, or planning to use, a standardized system based on persistent digital identifiers (PIDs) for:
   1. “Raw” sensor data?

Yes, from the European Persistent Identifier Consortium (ePIC).

* 1. Physical samples?

No. Instead, currently only ICOS-internal schemes for sample identification are considered for use by the responsible Central Facilities. The IGSN (International Geo Sample Number) system may be implemented in the future for physical samples (mainly from Ecosystem Stations).

* 1. Data undergoing processing (QA/QC etc.)?

No, only ICOS-specific schemes are currently in use. However, ePIC PIDs may be implemented later if it is found useful for provenance tracking.

* 1. Finalized “publishable” data?

Yes, DOIs from DataCite.

1. 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?

All of the above!

1. 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, the PIDs will point to landing pages maintained by the ICOS Carbon Portal. (Content dynamically created from content in the ICOS data object metadata database.)

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

PID minting: At this time, ICOS will not have to pay for minting PIDs - the costs for ePIC PID minting will be carried by EUDAT, and the costs for DataCite DOI minting is covered by a grant from the Swedish Research Council to SND, the Swedish DataCite partner. This situation may change in the future (after 2018).

PID “maintenance”: This falls under the existing metadata management budget item of the ICOS Carbon Portal.

**CITATION**

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

ICOS observational data are mainly used by environmental and climate scientists as input (driving parameters and/or for tuning &verification) to various models - atmospheric inversion models, or ecosystem vegetation models. Another important use is for direct comparison and/or benchmarking with researchers’ own datasets.

1. Do your primary user community traditionally refer to datasets they use in publications:
   1. By providing information about producer, year, report number if available, title or short description in the running text (e.g. under Materials and Methods)?

This is the traditional way.

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

This is also common.

* 1. By DOIs, if available, in the References section?

This is not very common (yet).

* 1. By using other information?

No answer provided (unclear question).

* 1. By providing the data as supplementary information, either complete or via a link

If required by the publishers. Often only partial datasets are made available in this way.

1. Is it important to your data users to be able to refer to specific subsets of the data sets in their citation? Examples:
   1. Date and time intervals

Yes, this is very important. Also qualitative temporal “filters”, such as daytime or nighttime, or seasonal markers, may be useful.

* 1. Geographic selection

In principle, yes. If datasets contain e.g. observations collected at different locations (even within a given observation station footprint), it may be very important. Also elevation (or depth) with respect to the ground level may be significant.

* 1. Specific parameters or observables

Yes, this is very important.

* 1. Other

Some uses may require the possibility to specify a specific instrument or measurement configuration, or the application of conditions (on the value of a quality flag, for example).

1. 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?

This will clearly become more important, also for ICOS data users.

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

ICOS is planning to use different measures for this...

* 1. Downloads/access requests

Yes, these will be logged together with the respective queries.

* 1. Visualization at your own data portal

Yes, all requests will be (anonymously) logged.

* 1. Visualization at other data portals

In principle, this is desirable - but may be difficult to implement in practice. Requests for data from e.g. ICOS-operated mapping and/or coverage services will of course be logged. But it will not be easy, or even possible, to track visualizations that rely on replicas of ICOS datasets that are stored at & served from other data portals (after they downloaded the data in question).

* 1. References in scientific literature

Yes, using (traditional) bibliometric methods. But this may not work 100% until all “citation indices” accept to track data object DOIs.

* 1. References in non-scientific literature

In principle, yes - but unless full-text searches are implemented on large corpora, this cannot be expected to result in any “complete” record.

* 1. Scientific “impact”

Yes, but this is again difficult to quantify.

1. Who receives credit when a dataset from your RI is cited?

As ICOS is still in the start-up phase, no official ICOS datasets have yet been published - and therefore no citations have been made yet. ICOS is however planning to set up its data object metadata database in such a way that every person and organizational entity involved in collecting and processing the (observational) data is associated with the respective data products. Based on this information, it should always be possible to extract accurate “credit lists”.

* 1. The RI itself

Yes.

* 1. The RI’s institutional partners (all or in part, depending on the dataset contents)

Yes.

* 1. Experts in the RI’s organization (named individuals)

At least in some cases.

* 1. “Principal investigators” in charge of measurements or data processing (named individuals)

Yes.

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

In some cases, but probably not all of the time.

1. 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?

The handling and curation of data objects in ICOS is performed in close collaboration between the Thematic Centers and the Carbon Portal. A key component is the exchange of metadata, and the setting up of one central “data object metadata database”. This relies heavily on provisioning ICOS vocabularies and ontologies that allow for (machine-to-machine) interoperability, with both internal and external users. A second crucial component is a system that ensures that unique persistent identifiers are automatically allocated to all data objects that should be “citable” (in the workflow, provenance description or later by end users). Thirdly, in order to implement the RDA recommendations for supporting dynamic data citation, all data products intended for end user consumption should be stored in a versioned and timestamped manner. This allows for unambiguously referring to (subsets of) data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store.

The best way to implement the above points in a way that will benefit not only ICOS but also a wider community, will be further investigated under the framework of the ENVRIplus Implementation Case “Dynamic data identification & citation”, that will be undertaken together with ANAEE, IAGOS and LTER.