|  |
| --- |
| **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*

* *Benjamin Pfeil, ICOS Ocean Thematic Centre & Geophysical Institute, University of Bergen*

This version is from January 27, 2016.

**5. Provenance**

**CASE 1: Atmospheric Thematic Center**

The ATC has chosen not to provide specific answers to this part.

**CASE 2: Ecosystem Thematic Center**

The ETC has chosen not to provide specific answers to this part.

**CASE 3: Ocean Thematic Center (Benjamin Pfeil, University of Bergen)**

1. Do you already have data provenance recording in your RI?

Yes

* 1. If so, please provide more information:

Will be implemented in 2016

1. Where/when do you need it, e.g., in the data processing workflows, data collection/curation procedures, versioning control in the repositories etc.?

All of the above stages plus quality control if it was not intended to be under one of the above

1. What systems are you using?

Will be implemented

1. What standards are you using?

No answer provided.

* 1. Advantages/disadvantages

No answer provided.

* 1. Have you ever heard about the PROV-O standard?

No answer provided.

1. Do you need provenance tracking?

No answer provided.

* 1. If so, which information should be contained?

No answer provided.

1. What information do you need to record regarding the following:
   1. Scientific question and working hypothesis?

Not relevant

* 1. Investigation design?

yes

* 1. Observation and/or measurement methods?

yes

* 1. Observation and/or measurement devices?

yes

* 1. Observation/measurement context (who, what, when, where, etc.)?

yes

* 1. Processing methods, queries?

yes

* 1. Quality assurance?

yes

1. Do you know/use controlled vocabularies, e.g. ontologies, taxonomies and other formally specified terms, for the description of the steps for data provenance?

NERC vocabulary server

1. What support, e.g. software, tools, and operational procedures (workflows), do you think is needed for provenance tracking?

No answer provided.

1. How does your community use/plan to use the provenance information?

No answer provided.

* 1. Do you have any tools or services in place/planned for this purpose?

Yes through the SOCAT automation that will be adjusted to

**CASE 4: Carbon Portal view (Margareta Hellström, Lund University)**

1. Do you already have data provenance recording in your RI?

No, at least not in a standardized, “overarching” form. The three Thematic Centers (Atmosphere, Ecosystem and Ocean) each have slightly different strategies and approaches. However, ICOS is considering different options for harmonizing provenance tracking across the whole RI, as it is recognized as being very important for data interoperability.

* 1. If so, please provide more information:

No answer provided.

1. Where/when do you need it, e.g., in the data processing workflows, data collection/curation procedures, versioning control in the repositories etc.?

Basically in all parts of the data life cycle, where data either passes from one organizational sub-entity to another (e.g., from observation station to Thematic Center), where some action is performed on the data (e.g., calibration, gap filling, quality control) and when the properties and/or relationships of data objects change (a new version is created, necessitating adding info about parents & child objects to the respective metadata))

1. What systems are you using?

This is still in the design & development stage. But some kind of ontology-based database is foreseen to keep all ICOS data object metadata in a central place.

1. What standards are you using?

None yet

* 1. Advantages/disadvantages

No answer provided.

* 1. Have you ever heard about the PROV-O standard?

Yes. We are currently investigating how much of this standard that we should or could incorporate into the ontology we are in the process of designing and implementing for ICOS. (Updated on 2016-01-26.)

1. Do you need provenance tracking?

? How is provenance tracking different from provenance recording?

* 1. If so, which information should be contained?

Yes, definitely. Our observational data products comprise data collected at 100+ individual stations, each equipped with hundreds of sensors, so it is very important to record exactly where data values come from. Also the entire chain of processing steps - including calibrations, quality control, gap filling, as well as any aggregation (e.g., mean values over collections of sensors in space or time) - must be traceable. Some variables, like half-hourly fluxes, are calculated from high-frequency observations using complex data processing. Here, too, it is important to include information about which software versions and parameter sets were applied. Similar considerations are applicable to any “elaborated” data products (produced by ICOS or by members of our designated user communities) that are based on ICOS observations - the models applied, and the data sets used, must be clearly specified to allow full reproducibility.

1. What information do you need to record regarding the following:
   1. Scientific question and working hypothesis?

None.

* 1. Investigation design?

This will be covered by a series of technical publications, which will be available under Open Access and shared at ICOS web sites. The documentation will cover all instrumentation specifications, measurement and observation protocols, data evaluation methodology, etc.

* 1. Observation and/or measurement methods?

See above.

* 1. Observation and/or measurement devices?

All individual variables in ICOS data products will have associated metadata entries describing where and how the measurements were made (at which height, using what sensor, etc.).

* 1. Observation/measurement context (who, what, when, where, etc.)?

See above.

* 1. Processing methods, queries?

All individual variables in ICOS data products will be accompanied by information about the processing steps undertaken etc. If specific queries and/or selection criteria (e.g. on other variables) are required, this will be indicated.

* 1. Quality assurance?

The ICOS Thematic Centers apply several “layers” of quality assurance (QA) and quality control (QC) on the variables they process. (For a large number of the variables, both the evaluated value and the associated “quality flag” will be reported for each time step covered by the dataset.) The methodology (“flagging technique”) used will be indicated in the provenance.

1. Do you know/use controlled vocabularies, e.g. ontologies, taxonomies and other formally specified terms, for the description of the steps for data provenance?

We are looking into this now, to see if we can use existing ones “out of the box”, or need to make many ICOS-specific changes. We are also very interested to see if we can apply semantic sensor network (SSN) ontologies to capture information.

1. What support, e.g. software, tools, and operational procedures (workflows), do you think is needed for provenance tracking?

To make the process as automated as possible, we will create a number of scripts for collecting information and for passing this on to the relevant metadata repositories. These will be integrated into the control scripts for data transfer, data processing and data curation - both at the Thematic Centers and at centrally at the Carbon Portal.

1. How does your community use/plan to use the provenance information?

For users of ICOS data products, it will be important to be able to follow the data “back in time” and see all the steps that happened from raw data collection, via quality control and aggregation, to useful product. For elaborated data products, knowing which ICOS data (and other information) was used is also very important.

* 1. Do you have any tools or services in place/planned for this purpose?

Not yet. But we are planning to provide users with web interfaces that will present all types of metadata for ICOS data objects available via the Carbon Portal website - including provenance.