Identification and citation in IS-ENES2

Identification

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

- **Content-wise (all parameters together, or separated e.g. by measurement category)?**
- **Temporally (yearly, monthly, daily, or other)?**
- **Spatially (by measurement station, region, country or all together)?**

We store a time series of each variable in a simulation run at given sampling frequency (yearly, monthly, day, sub-daily). Spatially we cover 1) the globe by gridpoints 2) regions like Europe, Africa…

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

Metadata catalogue; data on disk by variable (ESGF), some data on tape (LTA).

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?**

ESGF: several versions are kept in parallel on some reference nodes. Versions applied at the dataset level and contain several files pertaining to given variable or set of variables. New version are store in new directory. LTA: Version info is part of MD.

4. **Is it important to your data users that**

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

Yes.

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

Yes, it does already.

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

Some Metadata only.

- **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:

- **"Raw" sensor data?** n/a
- **Physical samples?** n/a
- **Data undergoing processing (QA/QC etc.)?** Yes.
- **Finalized "publishable" data?** Yes.

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?**

EPIC and DOI.

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)?**

Landing pages maintained by DKRZ

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

Yes

Citation

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

As climate model input, for analysis and for comparison.

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

- **By providing information about producer, year, report number if available, title or short description in the running text (e.g. under Materials and Methods)?**
- **By adding information about producer, year, report number if available, title or short description in the References section?**
- **By DOIs, if available, in the References section?**
- **By using other information?**
- **By providing the data as supplementary information, either complete or via a link**

DOIs are available for the most important data products like CMIP5 and CORDEX. Data is ready to be cited in the reference section, but it is not yet usual to do so.

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

- **Date and time intervals**
- **Geographic selection**
- **Specific parameters or observables**

We recommend citing a dataset collection and specifying the used subset in the text. The above-mentioned subsets are possible in any combination as well as combining specific subsets over multiple dataset collections i.e. citation entities.

4. 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?

See iii: A collection is suitable to be used in a reference list to keep the balance between data and paper citations.

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

- **Downloads/access requests**
- **Visualization at your own data portal**
- **Visualization at other data portals**
- **References in scientific literature**
- **References in non-scientific literature**

Scientific "impact"

Downloads/access requests: by number and volume with continental information on user origins (for DKRZ visualised on the DKRZ-Website). References in scientific literature, Scientific "impact": establish data references as part of the scientific record.

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

- **The RI itself**
- **The RI's institutional partners (all or in part, depending on the dataset contents)**
- **Experts in the RI's organization (named individuals)**
- **"Principal investigators" in charge of measurements or data processing (named individuals)**
- **Staff (scientists, research engineers etc.) performing the measurements or data processing (named individuals)**

The creator(s) as specified by the data originator; creators might be persons or institutions.

7. 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?

Not mentioned above is the identification of creators by PIDs like ORCID or the relation/connection to a scientific publication. Earth System Sciences data is of high volume; therefore data is hosted at established archival centers. Certificates like Data Seal of Approval (DSA) and World Data System (WDS) approval are of growing importance. Usually we have so-called 'stand-alone' data publications not directly connected to or supplemented to an article. Most of the data users publishing articles are not identical with the data creators.

We currently work on a stable and reliable possibility to cite dynamic data (CMIP6) in a federated data infrastructure.

Formalities (who & when)

Go-between	Yin Chen
RI representative	Sylvie Joussaume <sylvie.joussaume@lsce.ipsl.fr> Francesca Guglielmo <francesca.guglielmo@lsce.ipsl.fr>
Period of requirements collection	Oct -Nov 2015
Status	Completed