Identification and citation in LTER

Context of identification and citation in LTER

Summary of LTER requirements for identification and citation

Detailed requirements

1. **1. Identification and Citation**
   a. Identification
   i. What granularity do your RI's data products have:

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

Depends on the type of data. Both options are available. Time series (e.g. SOS) are normally providing one parameter. File based data provision is normally providing a block of related parameters in one file.

Sufficient MD need to be provided.

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

depending on the data provider; data on different levels of temporal aggregations should be provided, e.g. monthly means across all sites; more detailed depending on the needs

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

both options could be; spatial aggregation or separation. MD need to specify also the spatial link to the research site.

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

Depending on the data type. Static for file based data (e.g. temporal slices) or dynamic (e.g. SOS). Both options are supported

iii. How does your RI treat the "versioning" of data - are older datasets simply replaced by updates, or are several versions kept accessible in parallel?

currently different versions are kept by the data provider. The data portal (DEIMS) points to the most recent version of the data. versioning will be implemented (e.g. via B2SHARE) in the next version

iv. Is it important to your data users that

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

yes, even is not implemented at the moment we are heading for

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

there are technical MD which are kept by the data repository (e.g. B2SHARE); here at the moment we separate them from 'content MD'

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

not implemented at the moment. Need to be discussed

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

yes

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

- "Raw" sensor data?

vi. Yes, versioning of the data important by keeping the PID

- Physical samples?

yes

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

yes, versioning of data

- Finalized "publishable" data?

yes

vii. 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?

currently B2SHARE handle will be used

for data outside B2SHARE a solution needs to be found

viii. 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)?

in development within EUDAT2020

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

currently not, as EUDAT2020 is offering that in the project framework. Cost will be specified if in a operational phase.

1. **b. Citation - to be done, planned: 20160118**

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

Data from the LTER sites are usally used for modelling approaches or direct site comparisons. As many of the LTER sites are also part of regular monitoring programmes data are also reported to the respective programme centers (e.g. UNECE ICP Integrated Monitoring, UNECE ICP Forest). In this case data are used in the reporting to environmental programmes.

ii. 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?

If no further information is provided (e.g. DOI or data paper), data are referenced in the acknowledgement section. If data play an important role sometime co-authorship is offered to the data providers. In addition data used in the publication are described in the article as part of the material.

The is no common solution for the referencing and citation of the data entities at the moment. Depending on the local policy of the LTER site PIDs might be used or not. Some of the datasets are shared using common data repositories, e.g. Pangea, which are issuing DOIs to the data objects. Also other platforms like ResearchGate or FigShare are used. The use of PIDs is not common.

For the LTER Europe Data Node, as virtual node to share data with the community, also EUDAT services, e.g. B2SHARE will be used and the use of PIDs are implemented. In general for metadata a UUID is assigned.

iii. 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

In order to ensure reproducibility of the scientific results the reference to data fractions and queries would be desirable. In order to allow a reproducibility of the results the selected data entities (e.g. specific query of data, especially in terms of data services) should be citable. Currently, a system for dynamic this is not implemented for the different sites of LTER Europe.
In practise uploading of analysis data to data sharing platforms (e.g. ResearchGate, etc.) or data repositories (e.g. Pangea à DOI) are used by the scientific community. This should be replaced by a clear strategy of data provision and referencing by the data providers.
A separate provenance record needs to be provided. This is currently not supported by LTER Europe, but is in discussion in relation to the EUDAT2020 Project.

iv. 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?

A single PID (e.g. DOI) for a dynamic data series would be desirable. This means data which are continuously completed, e.g. automatic data flows of sensor based data which are sliced according to time (e.g. daily) for management reasons. In this case a single PID (e.g. DOI) should be attached.

For data collections PIDs should be applied, in order to describe the data collections (e.g. related measurements). The single attached data objects should be identified by a separate PID which is linked to the data collection.

Appropriate description of the data (metadata and annotations) following community based data metadata schemata (e.g. EML or ISO19115) and common vocabularies (e.g. EnvThes).

The research sites (see http://data.lter-europe.net/deims/) are currently referenced by a internal unique identifier (which will be changed to a UUID). PIDs (e.g. DataCite DOI) could be used to uniquely identify the locations of the observation. In this sense the research site is a kind of a “data collection” for different data objects/entities from this research site.

Persons could be referenced using a ORCID-ID as unique identifier for persons and organisations linked to data objects and research sites.

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

- Downloads/access
- Visualization at your own data portal
- Visualization at other data portals

- References in scientific literature
- References in non-scientific literature
- Scientific “impact”
  LTER Europe aims to collect information on two levels: a) the download and access of the data objects and b) the use of the data. The main aim is to show the added value of the data products provided and to have mechanisms to develop additional strategies for the planning of data collections based on user needs.

  The tracking of data download and access should be done via the data portal in order to assess the “usage” of the data. This will be an indicator of the importance. Usage tracking does not include the identification of the “user” as individual, but might include registration (e. g. by domain, see European Environmental Agency data download) for the download and access. The

  The use of DOIs or other PIDs should allow the use of the data in scientific and non-scientific literature in order to assess the scientific and societal impact of the data. Mechanisms to assess the usage of the data needs to be developed. This information is important for single LTER sites or research organisations in order to justify the public funding.

  Data will be visualised at the Data Integration Portal of LTER Europe. This portal is going to be developed in the current eLTER project. The data portal will track users in order to further develop the data portal.

vi. 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 credit goes to the principal investigator and the research site (e.g. LTER Site). The research infrastructure/network (e.g. LTER Europe) needs to be named in the metadata of the datasets or for the research site.

  This mechanisms needs to be further developed. In the long run it is planned to publish information on the data usage and the related results using the LTER Europe Data Integration Portal in order to have a single point of information for the research infrastructure.

  The LTER sites belong to the LTER Europe network. So referencing to data and receiving credit can be organised differently by the different national networks and even between the different research sites. A common strategy will be developed in the coming years.

Formalities (who & when)

Go-between	@Barbara Magagna
RI representative	@Johannes Peterseil
Period of requirements collection	201601
Status	collected