

Identification and citation in EISCAT-3D

Context of identification and citation in EISCAT-3D

Summary of EISCAT-3D requirements for identification and citation

Detailed requirements

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

The answers here apply to the existing EISCAT data. The EISCAT_3D data will be separated in a similar way.

The data are separated according to experiment type, EISCAT member country, date

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

Other: products related to campaign measurements are according to experiment scheduling

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

By measurement station and antenna scan pattern

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

The existing EISCAT systems use a combination. Raw data are stored as static files but the experiment directories are indexed in a MySQL database (by experiment sorting and absolute URLs).

Analysed data are stored in MADRIGAL. This is a system that stores results as files (will migrate to HDF5 in the foreseeable future) in a directory structure, but adds an index (plain text files; numbering changes as new experiments are added) and web services (both online and APIs) for search, retrieval and simple plotting.

These existing archive systems will not scale to raw or analyzed EISCAT_3D data. A standard data management system (TBD) will be used to the extent possible, together with a customized user portal.

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?

Older datasets are usually replaced by updates.

4) Is it important to your data users that:

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

EISCAT_3D will use a PID system (the exact system not decided yet) but there will probably be no need to have every object tagged, only the archived files. Other files are identified by station and timing, etc.

b) The metadata for data files contains checksum information for the objects?

Yes

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

Metadata are stored in databases and must be linked to data

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

The minimum requirement is to retrieve all relevant metadata when a user downloads data and attach the metadata to the data files.

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

a) "Raw" sensor data?

This will be decided at a later stage. Probably only timing and site information will be tagged.

b) Physical samples?

Not applicable.

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

Yes. This would apply to archived preliminary data before analysis of physical parameters.

d) Finalized “publishable” data?

Yes, there are plans to do so both for EISCAT_3D and at some point for migrated archives of existing EISCAT data.

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?

This will be decided at a later stage.

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)?

Not applicable.

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

Not applicable.

CITATION

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

They use the data products typically for comparisons with other measurements or model results.

10) Do your primary user community traditionally refer to datasets they use in publications:

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)?

The datasets are usually referred to by short descriptions in the text. Co-authorship may be required for extensive use of the EISCAT system.

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?

A method for data citation by PIDs should be taken into use when the PIDs are introduced.

d) By using other information?

Including the standard EISCAT acknowledgment is required.

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

This should be possible in the future and will have to be part of the EISCAT_3D portal functions

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

a) Date and time intervals

Yes

b) Geographic selection

Yes, such as radar beam directions or antenna scan patterns

c) Specific parameters or observables

Yes

d) Other

The data use is embargoed by EISCAT affiliate (country or institute) (According to the EISCAT statutes data use is restricted to the affiliate for a certain number of years after an experiment, then opened)

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, referring to all data from one measurement campaign as a single entity is desirable

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

a) Downloads/access requests

Currently EISCAT is only checking country of user's IP address. MADRIGAL registers email address and affiliation.

b) Visualization at your own data portal

The MADRIGAL system has some visualization functions

c) Visualization at other data portals

Data from the present EISCAT systems are accessible through ESPAS (<http://www.espas-fp7.eu>) and were also used in the ENVRI pilot project.

d) References in scientific literature

At the present EISCAT member countries gather lists of published articles by scientists connected to that country where EISCAT data have been used. EISCAT_3D should take a PID system into use

e) References in non-scientific literature

No strategy

f) Scientific “impact”

No strategy

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

a) The RI itself

Yes, using the standard EISCAT acknowledgment

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

Yes, the standard EISCAT acknowledgment lists the partners

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

Sometimes (in the acknowledgment or as co-authorship)

d) “Principal investigators” in charge of measurements or data processing (named individuals)

Yes, by authorship

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

Co-investigators or PIs responsible for the experiment are credited by authorship. The engineers operating the radars are not named.

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?

This is addressed by the ongoing EGI-CC EISCAT_3D portal project as well as an EUDAT pilot project.

Formalities (who & when)

Go-between	?? Questionnaire responses received by Maggie Hellström
RI representative	Ingemar Häggström and Carl-Fredrik Enell
Period of requirements collection	Oct 2015 - Dec 2015
Status	Information gathering completed, analysis yet to be done