Identification and citation requirements

Maggie Hellstrom and Alex Vermeulen with help from go betweens and others

The questions that were sent to the RIs are available here: 1 - Identification and citation questions.docx

The following RIs have contributed to identifying and describing ENVRIplus identification and citation requirements: ACTRIS, AnaEE, EISCAT-3D, EMBRC, EMSO, EPOS, Euro-ARGO, EuroGOOS, IAGOS, ICOS, IS-ENES2, LTER, SeaDataNet, and SIOS (click on individual RI names to see the respective responses).

Introduction

Identification of data (and associated metadata) throughout all stages of processing is really central in any RI. This can be ensured by allocating unique and persistent digital identifiers (PIDs) to data objects throughout the data processing life cycle. The PIDs allow unambiguous references be made to data during curation, cataloguing and support provenance tracking. They are also a necessary requirements for correct citation (and hence attribution) of the data by end users, as this is only possible when persistent identifiers exist and are applied in the attribution.

Environmental research infrastructures are often built on a large number of distributed observational or experimental sites, run by hundreds of scientists and technicians, financially supported and administrated by a large number of institutions. If this data is shared under an open access policy it becomes therefore very important to acknowledge the data sources and their providers. There is also a strong need for common data citation tracking systems that allow data providers to identify downstream usage of their data so as to prove their importance and show the impact to stakeholders and the public.

Overview and summary of identification and citation requirements

Identification

The survey found a large diversity between RIs regarding their practices. Most are applying file-based storage for their data, rather than data base technologies, which suggests that it should be relatively straightforward to assign PIDs to a majority of the RI data objects. A profound gap in knowledge about what persistent and unique identifiers are, what they can be used for, and best practices regarding their use, emerged. Most identifier systems used are based on handles (DOIs from DataCite most common, followed by ePIC PIDs), but some RIs rely on formalized file names. While a majority see a strong need for assigning PIDs to their “finalized” data (individual files and/or databases), few apply this to raw data, and even fewer to intermediate data - indicating PIDs are not used in workflow administration. Also, metadata objects are seldom assigned PIDs. Costs for maintaining PIDs are typically not treated explicitly.

Citation

Currently, users refer to data sets in publications using DOIs if available, and else provide information about producer, year, report number etc. either in the article text or in the References section. A majority of RIs feel it is absolutely necessary to allow unambiguous references to be made to subsets of data sets, preferably in the citation, while few find the ability to create and later cite collections of individual data sets is important. Ensuring that credit for producing (and to a lesser extent curating) scientific data sets is “properly assigned” is a common theme for all RIs - not the least because funding agencies and other stakeholders require such performance indicators, but also because individual PIs want and need recognition of their work. Connected to this, most RIs have strategies for collecting usage statistics for their data products, i.e. through bibliometric searches (quasi-automated or manual) of from scientific literature, but thus often rely on publishers indexing also data object DOIs.

Conclusion

The use of persistent and unique identifiers for both data and metadata objects throughout the entire data life cycle needs to be encouraged, e.g. by providing training and best-use cases. There is strong support for promoting “credit” to data collectors, through standards of data citation supporting adding specific sub-setting information to a basic (DOI-based) reference.

Research Infrastructures

The following RIs have contributed to identifying and describing ENVRIplus identification and citation requirements: ACTRIS, AnaEE, EISCAT-3D, EMBRC, EMSO, EPOS, Euro-ARGO, EuroGOOS, IAGOS, ICOS, IS-ENES2, LTER, SeaDataNet, and SIOS (click on individual RI names to see the respective responses).