Provenance requirements

Introduction

In order to correctly use and reuse and interpret data within a research infrastructure, and cross research infrastructures their evolutionary history must be known in detail.

This history covers all the steps of the pathway of data:

- data acquisition:detailed information about scientific question and investigation design, observation or measurement methods, measurement devices and so forth is needed,
- data curation: exact description of QA measurements (flagging and annotation of data),data replication
- data publication: which data were accessed, which data are not accessable (the selection of data can strongly influence any further results of data processing), which query was carried out and when
- data processing: which method was used for further processing (aggregation of data, transformation, modelling)
- data interpretation: scientific knowledge drawn out of data plus the theories behind.

It is important to point out, that knowing the evolutionary history of data is important for any use and reuse of data: use and reuse within institutes (reuse some years after the investigation was made, reuse by other persons within institutes), use and reuse within Research Infrastructure and cross Research Infrastructures.

Inter alia provenance can help to avoid undetected duplication of datasets.

In order to have information on those steps, their description has to be tracked in the so called “data provenance” and made available to data users.

The requirements questionnaire with focus on provenance intends to collect whether provenance was so far already considered in the RI's data lifecycle and if so which system is in use. If this was to date not implemented the next set of questions is grouped about the RI's possible interest in provenance tracking: which type of information should be tracked, which standard to rely on and finally which sort of support is expected by ENVRIplus.

Overview and summary of provenance requirements

Most RIs already consider provenance data as essential and are interested in using a provenance recording system. Among all of the nine RIs who gave feedback about provenance only two already had a data provenance recording system embedded in their data processing workflows. EPOS uses the dispel4py workflow engine in VERCE, which is based on and is able to export to PROV-O whereas in future it is planned to use the CERIF data model and ontology instead. IS-ENES2 instead does not specify which software solution is applied but mentions: the use of community tools to manage what has been collected from where, and what is the overall transfer status to generate provenance log files in workflows. Some, such as SeaDataNet and Euro-ARGO, interpret provenance as information gathered via metadata about the lineage data with tools like Geonetwork based on metadata standards like ISO19139, but the information gathered is not sufficient to reproduce the data as the steps of processing are not documented in enough detail. Other RIs, such as ICOS and LTER, are already providing some provenance information about observation and measurement methods used within the metadata files but are aware that a real tracking tool still needs to be implemented. IAGOS is using the versioning system GIT for code but not for the data itself. A versioning system can only be seen as a part of the provenance information sought.

On which information is considered to be important, the answers range from versioning of data to the generation of data and modification of the data as well as on who, how and why data is used. So there seems to be two interpretations about what provenance should comprise: should it enable the community to follow the data ‘back in time’ and see all the steps that happened from raw data collection, via quality control and aggregation to a useful product, or should it enable the data provider as a means of tracking the usage of the data, including information about users in order to understand the relevance of the data and how to improve their services? These two roles for metadata may be served by the same provenance collecting system. The provenance data is then interpreted via different tools or services.

Regarding the controlled vocabularies used for the descriptions of the steps for data provenance, some RIs already use research specific reference tables and thesauri like EnvThes and SeaDataNet common vocabularies.

There is a big interest among the RIs to get clear recommendations from ENVRIplus about the information range provenance should provide. This includes drawing an explicit line between metadata describing the ‘dataset’ and provenance information. Also it should be defined clearly whether usage tracking should be part of provenance.

It is considered as being very important to get support on automated tracking solutions and or provenance management APIs to be applied in the specific e-science environments. Although there are some thesauri already in use there is a demand for getting a good overview of the existing vocabularies and ontologies that are ready to use or that need to be slightly adapted for specific purposes.

There is a strong relationship between the task of *identification* of data and the *provenance* task as there must be a direct link between the data and its lineage that can be followed by the interested user. Provenance tracking is also an important feature for optimisation. The connections with curation and cataloguing is evident which also becomes clear in the IC_2 Provenance implementation case[1] which aims amongst others at defining a minimum information set that has to be tracked, finding a conceptual model for provenance which conforms to the needed information, maps existing models to the common model and finds a repository to store the provenance information.

---

[1] https://wiki.envri.eu/display/EC/Use+Cases

Research Infrastructures

The following RIs contributed to developing provenance requirements

<Delete from the following list any that were not able to contribute on this topic>

<Add an interest inducing sentence or two, to persuade readers to look at the contribution by a particular RI. e.g., What aspect of the summary of requirements, or the special cases, came from this RI. Check with RIs that they feel they are correctly presented.>

RI	Done	Comments
ACTRIS	Y	
AnaEE	N	no details about provenance so far
EISCAT-3D	N	
ELIXIR	(Y)	too complex not easy to assess for each interest group
EMBRC	N	
EMSO	N	
EPOS	Y	
Euro-ARGO	Y	
EUROFLEETS2	N	
ESONET	N	
EUROGOOS	Y	
FIXO3	N	
IAGOS	Y	
ICOS	Y	
INTERACT	N	no data therefore no provenance
IS-ENES2	Y	
JERICO	N	
LTER	Y	
SEADATANET	Y	
SIOS	N	