**CF meeting**
**9-11 June 2020**
**Zoom session link**

# Discussion topic notes

---

## Embedding provenance information - Metadata handling through processes (David Huard)

https://github.com/cf-convention/discuss/issues/33

When multiple netCDF files are aggregated to compute ensemble statistics, or when chains of algorithms are applied to netCDF files, recording operations and source metadata in the "history" attribute can become very messy. Recording such provenance information is necessary in applications where traceability and reproducibility are critical.

There are standards (e.g. PROV) to encode such provenance or data lineage information in machine readable formats. netCDF files could include a “provenance” attribute in which a machine-readable representation of the provenance would be stored. An example of this can be found at http://metaclip.org/, where provenance information is embedded in figure metadata.

---

### Literature review

Other examples of provenance information being used in the climate community:

- ESMValTool embeds provenance information into the figures it generates.
- Climate4R embeds provenance information into the figures it generates (see e.g. https://www.sciencedirect.com/science/article/abs/pii/S1364815218305036).

ESMValTool and Climate4R are used to generate some of the figures within the IPCC WGI report.

OGC related work:

- OGC netCDF-LD
- https://www.ogc.org/projects/groups/netcdfswg

NcML is an XML representation of netCDF metadata:
https://www.unidata.ucar.edu/software/netcdf-java/v4.5/ncml/

## Open questions for discussion

- Examples I've seen of provenance information are not human readable. Is this compatible with the CF-Convention principles? Would the CF principles suggest that provenance information should be considered a *private* attribute, e.g. "_provenance" ?
    - Consensus is to refer to an external document, not embed the information into the netCDF file.
- On what existing ontology should this provenance document be built ? PROV-O ?
    - Agreement on using an existing standard.
- Should the provenance attribute reproduce the information already contained in the standard metadata ? In other words, the content of the "source" field would be mapped to a "prov:wasGeneratedBy" item. But then there is a risk of redundancy and/or discrepancies between the information contained in standard CF attributes and the provenance attribute content. Doing this right implies a unique mapping between the CF-Convention and an existing provenance ontology.
    - Can possibly use NcML
- When aggregating datasets, how do we maintain global provenance from multiple sources? How do we distinguish between variable-level and global provenance (Note: global attributes and variable level attributes with the same name may not be handled as expected by some or most software)

## Brainstorming ideas

Let's imagine a process using two input files (A, B) to generate C:

A:netCDF
-> Process -> C:netCDF
B:netCDF

The provenance of C could be described by using references to the NcML representation of A, B and C.
A: NcML
B: NcML
C: NcML + PROV embedding A:NcML and B:NcML and using references to describe how C is connected to A and B.

(Joaquín Bedia, Universidad de Cantabria)

This is just a small sample of a RDF - JSON file containing the provenance information of a CMIP5 ensemble map of a bias-adjusted climate index based on maximum temperature. For simplicity, it is formed by only two GCMs.
It can for instance be dropped onto the www.metaclip.org interpreter to be explored, but its complexity, with just two models, is huge. With an ensemble of ~30 models it is not possible to go for the full detail provenance information visually, and somehow the information needs to be collapsed.
An advantage of such RDF provenance representation is that it allows to expose the provenance information at different levels of granularity. It also ensures interoperability. The RDF graph can be serialized onto many different formats, noy just JSON, but also for instance XML.
There are some examples of climate products available in www.metaclip.org with attached provenance information that are automatically displayed in the viewer. However, the provenance information can be also linked externally via a reference attribute, as recommended for netCDF/NcML files.

(Bryan Lawrence, NCAS)

- We still can't properly reference vector components from one file to another, so properly referencing an actual upstream file seems beyond us. Better to reference the properties of the upstream files.
    - Consider the ensemble case. How do we keep the reference to the other ensemble member "real"? Files move.
- I also think putting provenance in the metadata is a category error alongside puting the ISO19115 metadata in there, insofar as "information density" about provenance can change after the data is produced (and often does, a la CMIPx). That is, you only get machine generated information, and that's limited. Decorating it with human information is desirable.
    - A la the example by Joaquin Bedia
    - My point here is we never want to put _in_ a netcdf file information that would require the file to be rewritten.
- I am strongly in favour of pointing to externally managed provenance using a mechanism a la the _further_info_url mechanism. That should be machine readable and conform to PROV (or whatever).
    - But let's not pretend our provenance information should enable reproducible workflows. That's just far too brittle in terms of dependency on workflow context (e.g. https://www.bnlawrence.net/assets/images/2020-06-04-software-simple.png, your workflow cannot often be fully reproduced elsewhere.)
    - However, if what we are doing is describing the workflow, so the workflow can be *reconstructed* elsewhere, then that's what we need.
    - I guess what i am suggesting is a specific attribute which could be expected to dereference to some specific type of external information. (as just suggested by Miquel)

- I suppose a mix of internal and external provenance would address this point, and in fact that's what we do for CMIP … there is a lot of implied and explicit external provenance in this history and other global attributes.
    - So enough information about provenance to meet “self describing” should remain in the file … (Nan’s point).
- The volume and complexity of provenance information *can be* immensely complex. I second the notion of adding a reference to the provenance information in the netcdf file. I believe Blockchain is used to track entity provenance, so this is a complex challenge addressed by a number of external applications.

Sadie Bartholomew (NCAS@Reading): this is quite a general question/thought, but Niels comment about a ‘tree’ structure has made me think that there is perhaps an inheritance component that would be ideal to record. For example, if two files had at one point in time been the same and after that had taken different routes in how they were processed to diverge, you could tell that (and find the divergence point) from the provenance information. (This may be an obvious goal and/or too idealistic, but thought I’d note it).

Niels Drost (NLeSC, Netherlands eScience Center): I think starting “small” with only a URL to the provenance and letting the actual provenance format completely free would be the best way forward. We can then let the community come up with good implementations and formats before we try to tie ourself down to a single (or a set of) formats. Removing and/or replacing conventions is much harder than appending to it. It is very much possible to discover what formats are available at a certain url, so you could host both human and machine readable formats, even multiple formats if better ones come along. (Seth) The url would ideally be a doi, but we should not require it, only highly recommend.

Daniel Heydebreck (DKRZ, German Climate Computing Center): Creating our own netCDF-Provenance standard costs considerable time and resources. Different groups/communities work on representing provenance information (e.g. RDA Research Data Provenance Interest Group). Therefore, it seems to be more reasonable to point to external provenance information (URL to it) from within a netCDF file instead of adding detailed information to netCDF files themselves. I would not add a second attribute describing the type of content to be expected. The provided URL could lead to a web service that allows doing HTTP content negotiation in order to get the format of choice of the provenance data. (comment from Nan Galbraith, WHOI: I agree that using an external standard is going to be more efficient and more interoperable than starting from scratch. Whether this could be incorporated into a netCDF file itself is another question.)

(Bryan) On coupling files to provenance, we really want to decouple any service layer from the content layer. So, we want to have

Provenance file which consists of

- Unique provenance file identifier
- Provenance stuff includes
- data tracking-ids.

Data files have

- Unique tracking identifier
- Attribute which includes the identifier of the provenance file.
    - This could be a URL, but if so, it has to include an identifiable piece which is the unique provenance id.
- Data stuff

Doing this means that we can always build new infrastructure rebuilding links that are lost by changing the server layers. We can query provenance and we can query data and find the relevant information.

Example of a provenance xml file generated by ESMValTool:
https://github.com/cf-convention/discuss/files/4759676/MultiModelMean_Amon_ta_2000-2001_mean_provenance.xml.txt

## Synthesis

General agreement on

- A variable/global attribute (e.g. “has_provenance”, “has_provenance_query_service”) whose content is a URL pointing to provenance information.
- The provenance file is not meant to replace human-readable information within netCDF files.
- Unique identifiers (persistent ids, checksums) to make sure there is an unambiguous relation between files and their provenance file.
- Support many-to-one model (multiple files to single provenance URL).
    - Provenance format should be able to describe aggregations.
- Recommend to use a URL with long survival time, such as DOIs.
- Recommend agreeing on a provenance convention, as it would help with user/developer training and focus software development.
- This is not meant to enable reproducibility out of the box.

Concerns were expressed about

- Provenance file growing larger than netCDF data
- Large projects could end up storing millions of files if each individual file has its own provenance information

- Having provenance at the coordinate level (different provenance information for the same variable but different vertical levels for example)
- Unclear what level of duplication between CF content and provenance file is useful. Probably depends on the processing workflows.

Use cases

- Detecting bugs in processes by inspecting the provenance file
- Identifying which version of which tool was used to generate netCDF data