Equipping OPeNDAP with data citation functionality

Niklas Griessbaum
griessbaum@bren.ucsb.edu
Earth Research Institute
University of California Santa Barbara

James Frew
frew@ucsb.edu
Bren School of Environmental Science & Management
University of California Santa Barbara

Why data citation?

Data citations are persistent and technology-agnostic interfaces, serving multiple purposes:

1. Credit

Preparing and publishing data is an intellectual activity that, in many cases, deserves its own credit.

2. Provenance

Use of well-known data helps establish credibility.

3. Access

Citations should not just *describe* data, but also facilitate its (re-)use.

Properly-implemented data citations should promote data sharing, increase transparency, enable reproducibility, and allow researchers to extend each other's work.

Problems with data citations

- There is no standard way to cite data. Most current "data citations" are really just citations to papers or websites that **describe** a dataset.
- Data **structures** are often richer than the document hierarchies supported by conventional citation. This is particularly an issue for geospatial data.
- Data is often published as a **dynamic service**, as opposed to a static dataset.

More generally, scientific data are often retrieved as complex subsets of larger datasets that may change over time, and a citation to the data should include information about the subset request [4]. For convenience and accuracy, such citations should be generated automatically [1]. Our proposed solution extends the API of the Open-source project for a Network Data Access Protocol (OPeNDAP) to generate citations precisely matching the requested data.

OPeNDAP at a glance

OPeNDAP facilitates remote data access by abstracting the local data storage format and by allowing subsetting.

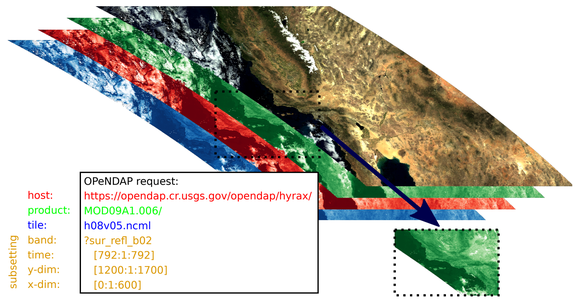

Figure: Subsetting of MODIS data with DAP.

Handlers exists for a variety of data stores, such as SQL databases or CSV and netCDF files, allowing conversion of these stores into the DAP data model. A DAP server may then convert the data from the DAP data model into the format of the client's choice through the use of responses [3].

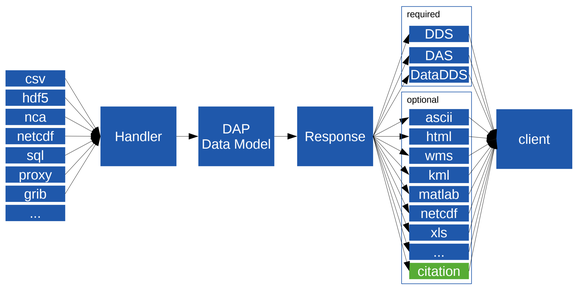

Figure: DAP handlers and DAP responses

Similar to the handlers, a variety of responses such as netCDF, KML, and WMS are supported. We extend this service with a 'citation response'.

Citation generation

The citation response has been implemented in the *pydap* server, and as a DAP-citation proxy.

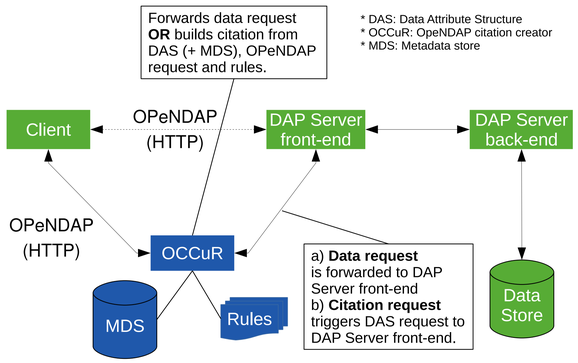

Figure: DAP-citation proxy schema

Metadata

Citations are generated with metadata stored in the DAP data attribute structure (DAS). A citation will contain at least the DataCite mandatory properties [2]. If the mandatory properties are not available from the DAS, the citation generator will fall back to retrieving metadata from a metadata store (MDS).

Cited data retrieval

Cited static DAP datasets can easily be accessed through e.g. a permanent link. However, if a DAP dataset is evolving over time, a DAP resource needs to be able to resolve "as-of" specifiers to avoid differences between the cited data and the re-retrieved cited data. This could possibly be implemented by:

1. ETags

Comparison of resource ETag (e.g. last modified time) with the dataset ETag in the citation.

2. Fingerprints

Comparison of fingerprint (e.g., cryptographic checksum) of cited subset with fingerprint of re-requested subset.

3. Versioning

Comparison of the cited data's version with the currently available version. (We use *version* here as a catch-all for assigned fixity, as opposed to fixity inferred from data properties.)

References

[1] Peter Buneman, Susan Davidson, and James Frew. Why data citation is a computational problem. *Communications of the ACM*, 59(9): 50–57, 2016. ISSN 0001-0782. doi: 10.1145/2893181.

[2] DataCite Metadata Working Group. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data Version 4.0. DataCite e.V. 2016. doi: 10.5438/0012.

[3] James Gallagher, Nathan Potter, Tom Sgouros, Steve Hankin, and Glenn Flierl. The Data Access Protocol - DAP 2.0. 2007. URL `https://www.opendap.org/pdf/ESE-RFC-004v1.2.pdf`.

[4] Stefan Proll and Andreas Rauber. Scalable data citation in dynamic, large databases: Model and reference implementation. pages 307–312, 2013. doi: 10.1109/bigdata.2013.6691588.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1302236, administered by the Earth Research Institute, University of California, Santa Barbara.