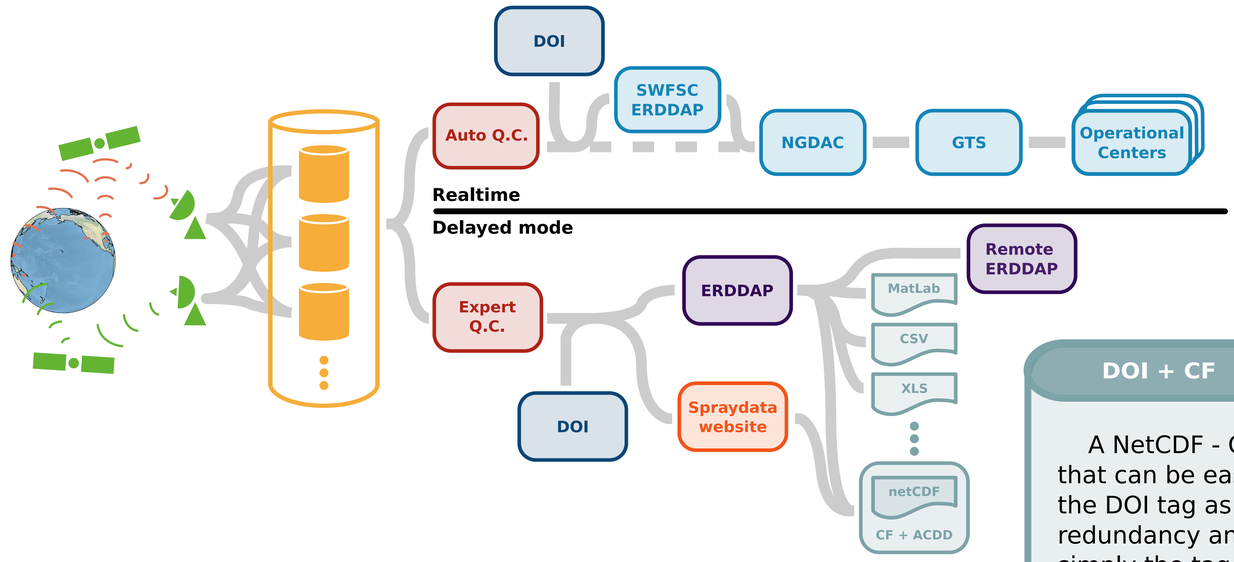
Spray Data Management with DOIs

Guilherme Castelão and Daniel Rudnick

Instrument Development Group - Scripps Institution of Oceanography

Introduction

Digital Object Identifiers (DOIs) allow easily tracking the scientific impact of data and software on the exact same fashion that scientific papers are tracked. One of the many advantages of DOIs is that anyone involved in the acquisition and production of data can be formally credited, including funding agencies. Since the past few years an increasing number of scientific journals have started to require DOIs for datasets used in publications. Although several groups already include DOIs as an attribute in NetCDF-CF datasets, a lack of standardization makes it difficult to automate



Digital Object Identifier

Each project dataset receives a Digital Object Identifier (DOI) allowing the users to accurately reference the origin and quality of the data while providing a means of tracking the impact of our products among the scientific community.

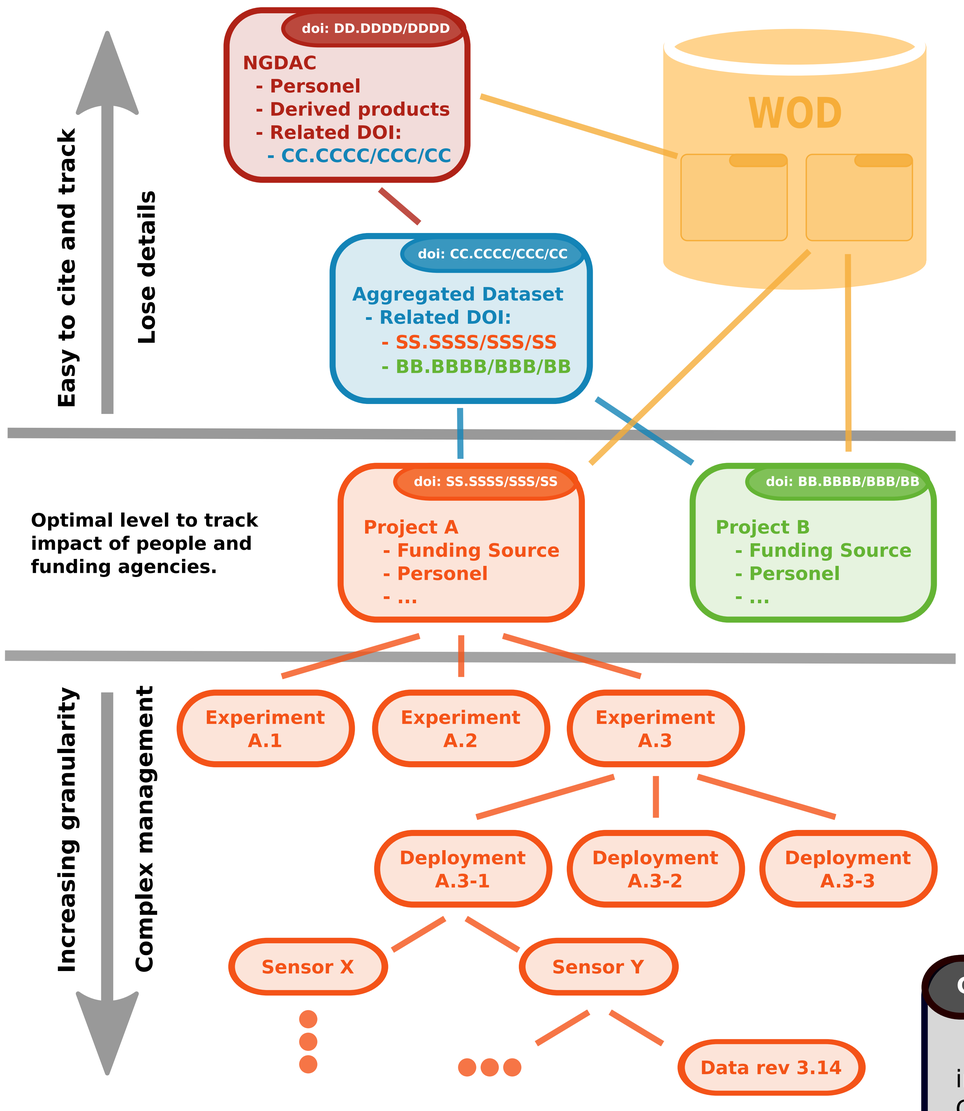
DOI - Metadata

A DOI is a unique identifier with aggregated metadata that can be used to refer external databases like ORCID. When correctly used, those links connect all the way from the data to scientific publications, allowing to track the impact of data providers and funding sources on the same fashion that scientific articles are currently tracked.

Below is an example of possible metadata:

DOI: 10.21238/S8SPRAY1618

- URL: https://spraydata.ucsd.edu
- Title: California Underwater Glider Network
- Creator:
  - Daniel: ORCID=000...
- Contributor:
  - Jeff (team leader): ORCID= 000 ...
  - Evan (glider operations): ORCID= 000...
  - Derek (glider operations): ORCID= 000 ..
  - Ben (pilot): ORCID= 000 ...
  - Kyle (pilot): ORCID= 000 ...
  - Gui (data): ORCID= 000 ...
  - Research Group: Inst. Dev. Group
  - ...
- Funding Reference:
  - Funder Name: NOAA Climate Observations Division, Southern California Ocean Observing System
  - Award Number:
  - Award Title:
- Related Identifier:
  - Is Complied By:
  - Is Derived From:
  - Is Variant Form Of:
  - ...



DOI + CF

A NetCDF - CF is a self-explaining data container that can be easily integrated with DOIs by including the DOI tag as an attribute. There is no need for redundancy and repeat all the DOI metadata, but simply the tag is enough. In the following example the only change is at the bottom.

The proposition #160 on the CF-Conventions recommends the standard shown below as a general CF standard.

```
netcdf CUGN_line_90 {
 dimensions:
  depth = 50 ;
  profile = 39758 ;
  name_strlen = 8 ;
  trajectory = 51 ;
 variables:
    int profile(profile) ;
       profile:cf_role = "profile_id" ;
    ...
    double depth(depth) ;
    ...
    double time(profile) ;
    ...
    double temperature(depth, profile) ;
       temperature:_FillValue = NaN ;
    ...
 // global attributes:
    :Conventions = "CF-1.7, ACDD-1.3" ;
    :title = "California Underwater Glider Network" ;
    :summary = "The overarching goal of ...";
    :featureType = "trajectoryProfile" ;
    :id = "CUGN_line90" ;
    :naming_authority = "edu.ucsd.spray" ;
    :references = "Rudnick, D. L. (2016). Ocean ..."
    ...
    :doi = "10.21238/S8SPRAY1618" ;
}
```

Optimal level for DOIs

A DOI can identify any kind of object, physical or virtual, thus it can tag a project, a glider, a sensor, or even a specific subset of measurements. With increasing details comes complexity to manage this hierarchical database and it also gets harder for users to decide what to cite. For Spray we assign a DOI at the project level grouping the data from all deployemnts from that project.

Acknowledgments

Acknowledgments: NOAA - Ocean Observing and Monitoring Division, Southern California Coastal Ocean Observing System, and Integrated Ocean Observing System; Office of Naval Research; National Science Foundation. Bob Simons created ERDDAP and maintain SWFSC ERDDAP; Spray operations are maintained by Jeffrey Sherman, Evan Randall Goodwin, Derek Vana, and Benjamin Reineman.

Citing a Dataset

A dataset should be cited like any other publication and included in the bibliography. For example, if using the California Underwater Glider Network dataset, the citation would look like:

Rudnick, D. L., K. D. Zaba, R. E. Todd, and R. E. Davis, 2017: A climatology using data from the California Underwater Glider Network - Dataset [Downloaded at 2019-05-21]. Scripps Institution of Oceanography, Instrument Development Group. doi: 10.21238/S8SPRAY7292