Data publication of open time series, Argo data and DOIs

Justin Buck (juck@bodc.ac.uk), BODC, UK
Thierry Carval, Thomas Loubrieu, Ifremer, France
Kenneth Casey, NODC, USA

Introduction

Open time series

Current Argo data DOI situation

Moving toward a single DOI

Introduction

Motivation

Push from publishers and scientists for data citation:

- Publishers want to link journal articles to the data
- Scientists want credit for data set creation and usage

Adopted approach presented here is:

Digital Object Identifiers (DOI)

In particular DataCite DOIs as defined in:

http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.0.pdf

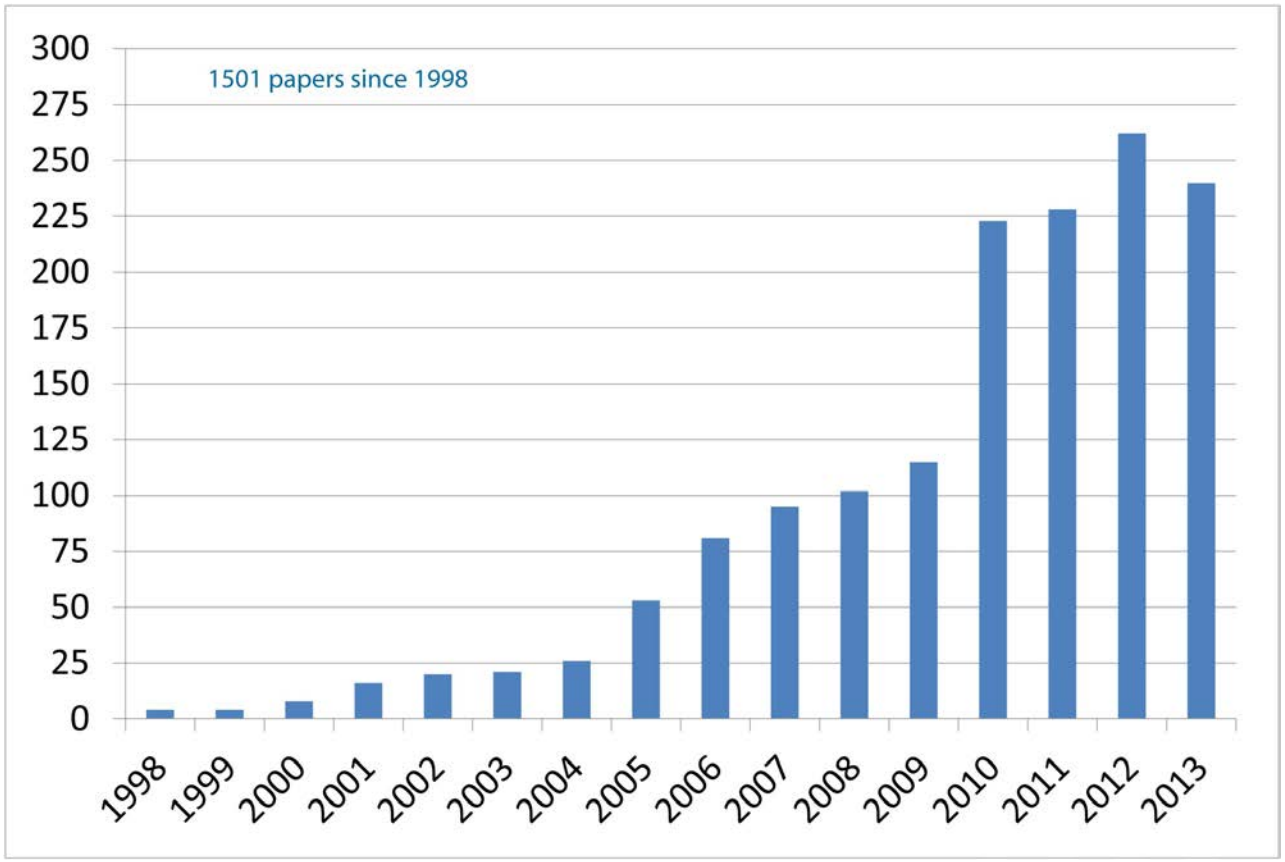
Reproducibility versus attribution of credit

Enabling reproducible research implies attribution of credit is possible but not vice versa.

Therefore to meet both needs or goal should be to enable:

Reproducible research

Argo – 200+ publications per year

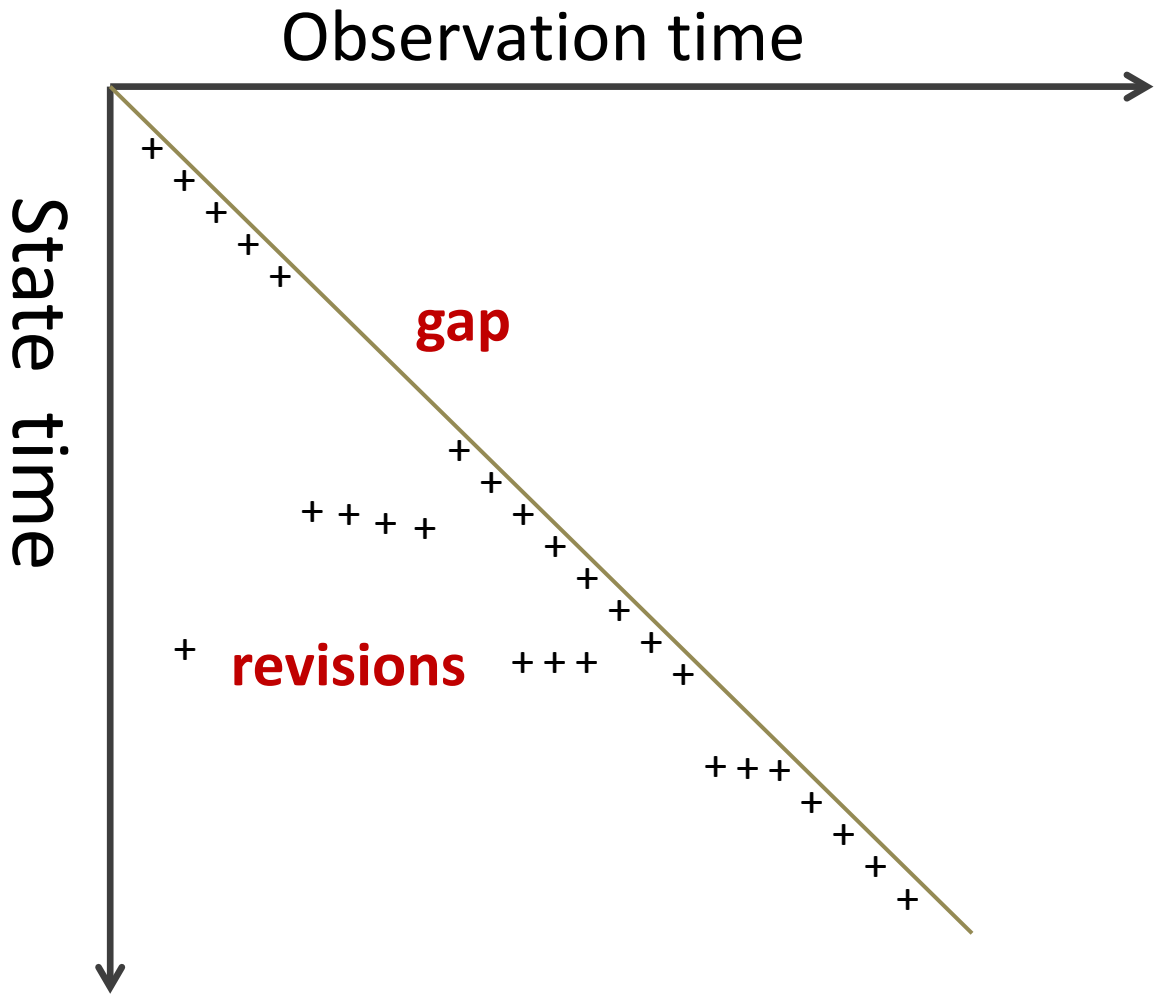


How to cite Argo data at a given point in time?

Can we do this with a single DOI?

Open time series

Growing and updating time series



EUDAT

DataCite – Dynamic data policy

For citation, three approaches are possible:

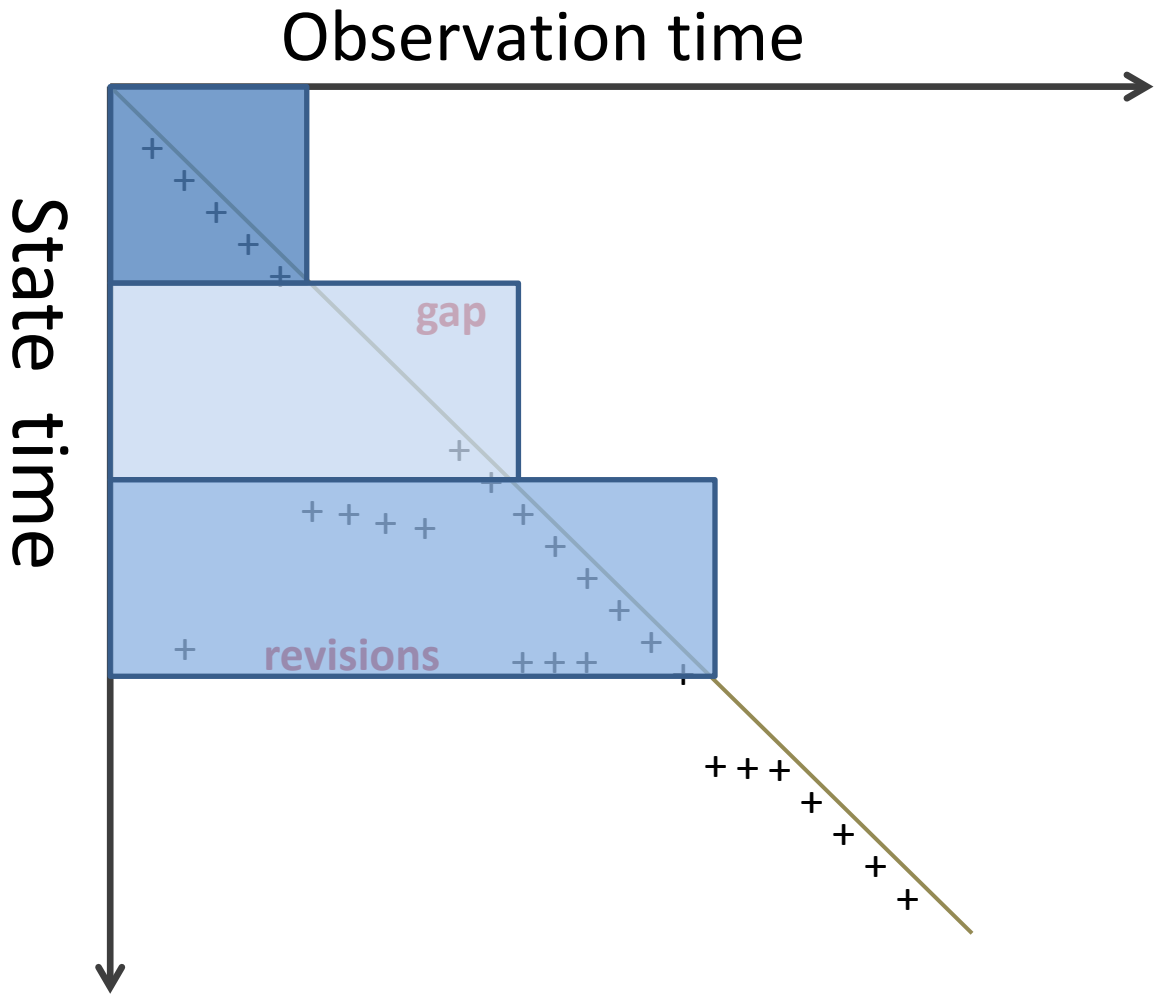
a) Cite a specific time slice (the set of updates to the dataset made during a particular period of time);

b) Cite a specific snap shot (a copy of the entire dataset made at a specific time);

c) Cite the continuously updated dataset, but add an Access Date and Time to the citation.

Copied from "DataCite Metadata Schema for the Publication and Citation of Research Data", version 3.0, July 2013

Time slices



DataCite – Dynamic data policy

For citation, three approaches are possible:

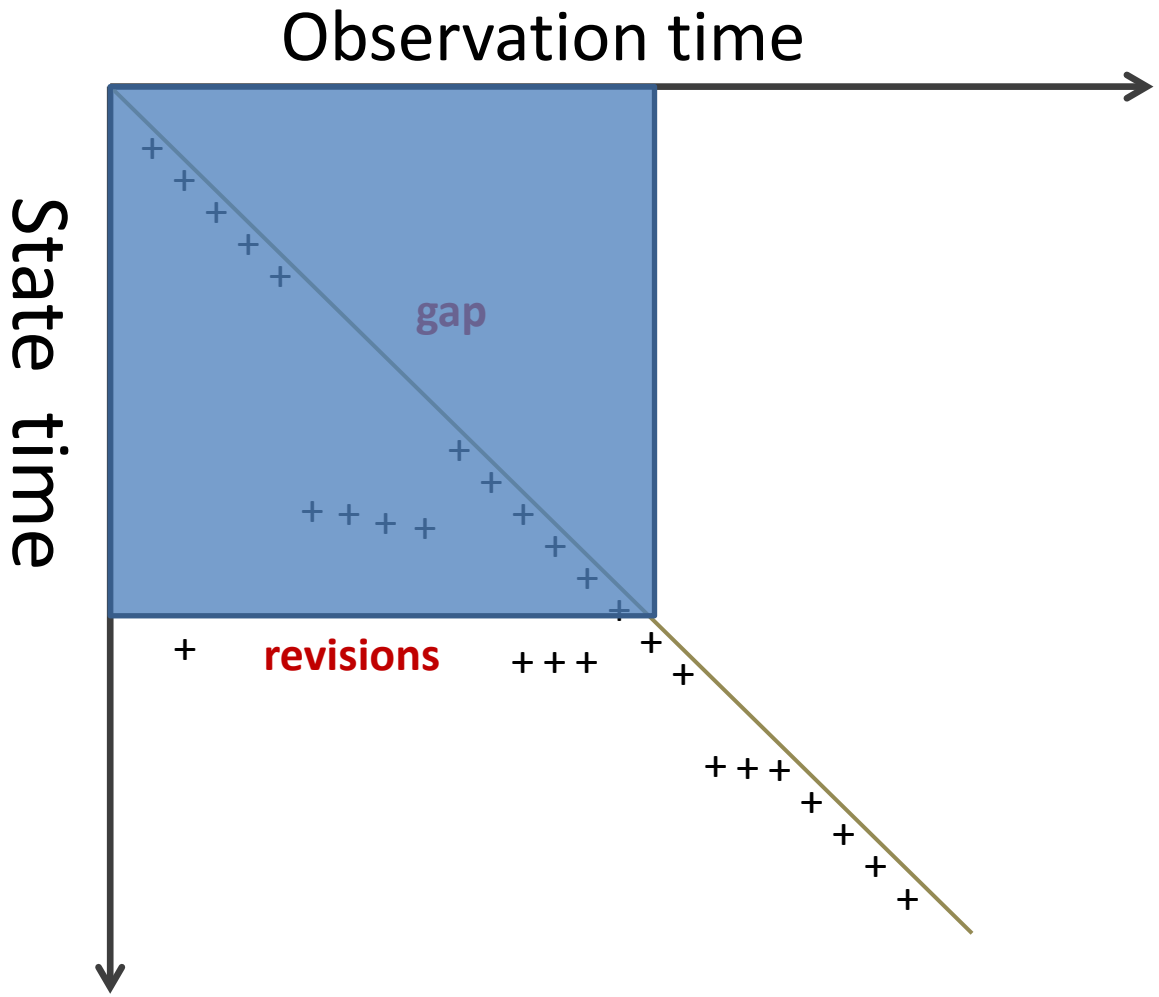
a) Cite a specific time slice (the set of updates to the dataset made during a particular period of time);

b) Cite a specific snap shot (a copy of the entire dataset made at a specific time);

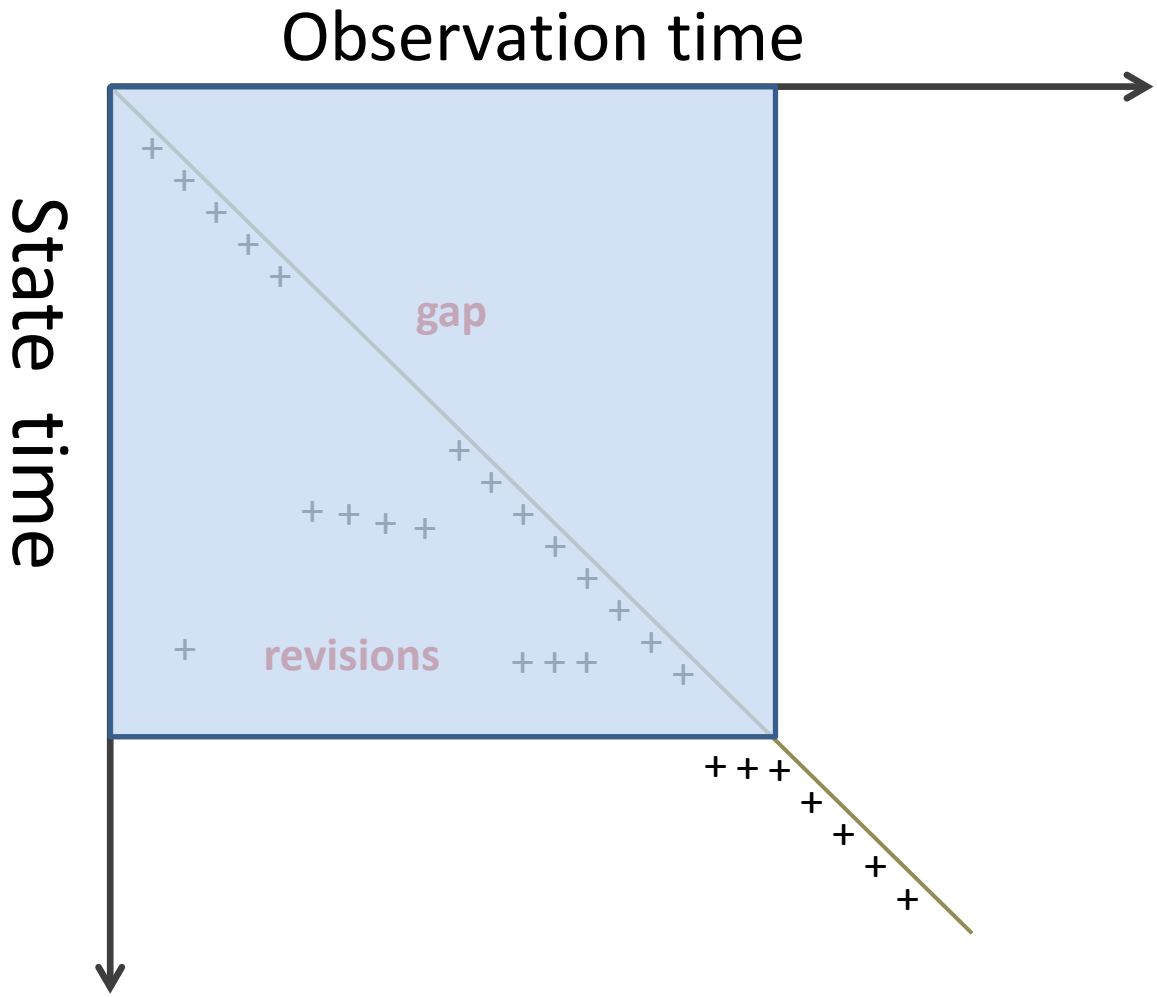
c) Cite the continuously updated dataset, but add an Access Date and Time to the citation.

Copied from "DataCite Metadata Schema for the Publication and Citation of Research Data", version 3.0, July 2013

Snapshot



Snapshot



DataCite – Dynamic data policy

For citation, three approaches are possible:

a) Cite a specific time slice (the set of updates to the dataset made during a particular period of time);

b) Cite a specific snap shot (a copy of the entire dataset made at a specific time);

c) Cite the continuously updated dataset, but add an Access Date and Time to the citation.

Copied from "DataCite Metadata Schema for the Publication and Citation of Research Data", version 3.0, July 2013

However ...

There are a couple of caveats:

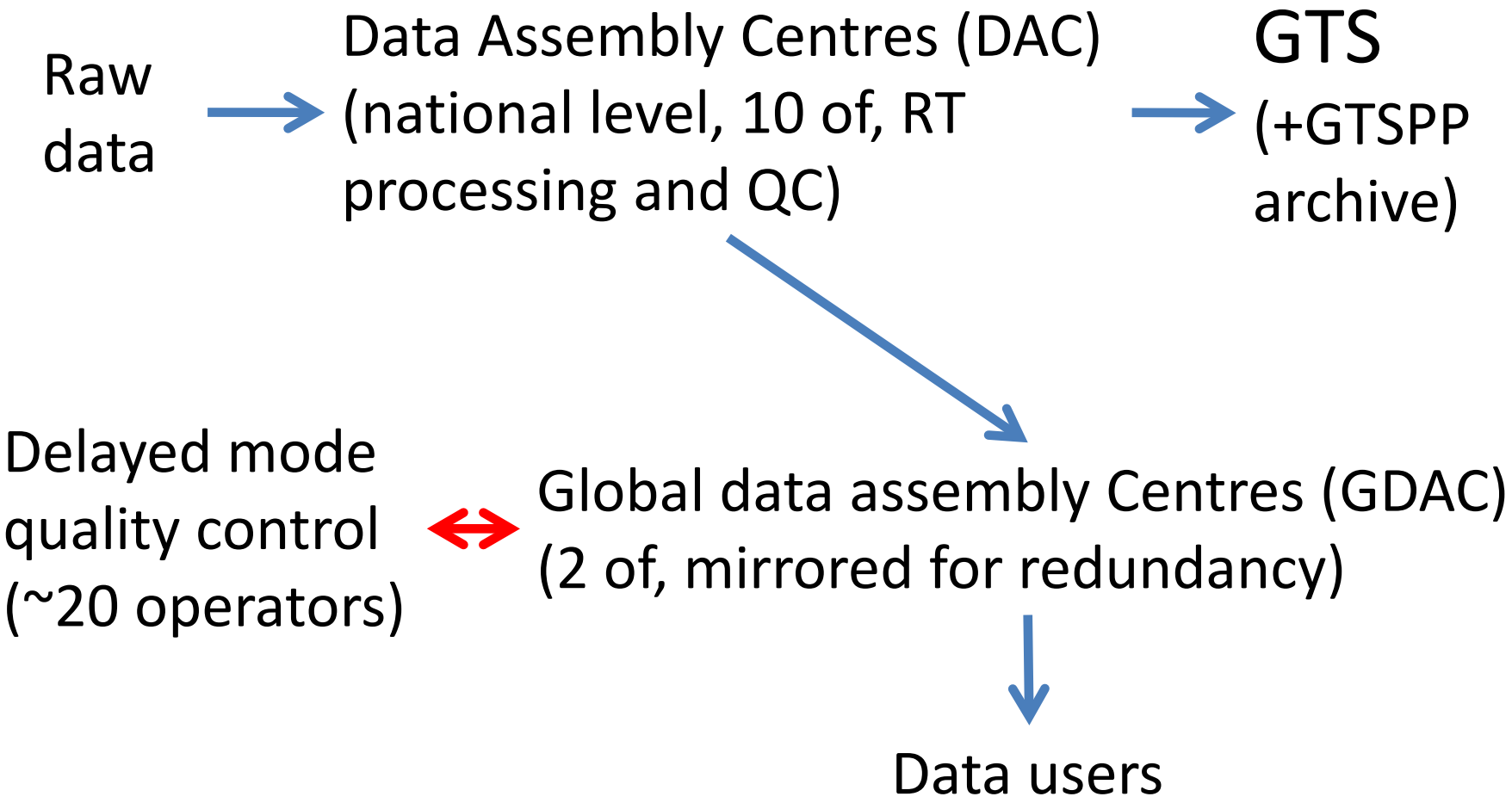
- Note that a “time slice” and “snap shot” are versions of the dataset and require unique identifiers.
- The third option is controversial, because it means that following the citation does not necessarily result in observation of the resource as cited.

Copied from “DataCite Metadata Schema for the Publication and Citation of Research Data”, version 3.0, July 2013

Argo and DOIs, current situation

(For current DOIs see: http://www.argodatamgt.org/Access-to-data/Argo-DOI-Digital-Object-Identifier)

Argo data system (simplified)



For the real time data stream

Ifremer minted a Digital Object Identifier (DOI) for the GDAC as a whole:

ARGO (2000): Argo floats data and metadata from Global Data Assembly Centre (Argo GDAC). IFREMER. Dataset.

http://dx.doi.org/10.12770/1282383d-9b35-4eaa-a9d6-4b0c24c0cfc9

and the recent version of the user manual:

Argo data management (2013). Argo user's manual.

http://dx.doi.org/10.13155/26387

These are sufficient for Argo if long term reproducibility of the data is not required by the user.

Argo snapshots enable reproducibility

To occur at GDAC level initially.

Monthly granularity of snapshots.

As per DataCite schema each snapshot will have a new DOI.

An example minted by Ifremer:

Argo floats data and metadata from Global Data Assembly Centre (Argo GDAC) - Snapshot of Argo GDAC as of March, 8th 2014

http://dx.doi.org/10.12770/e5e22416-cb4c-4c57-9dcb-f42d823dce33

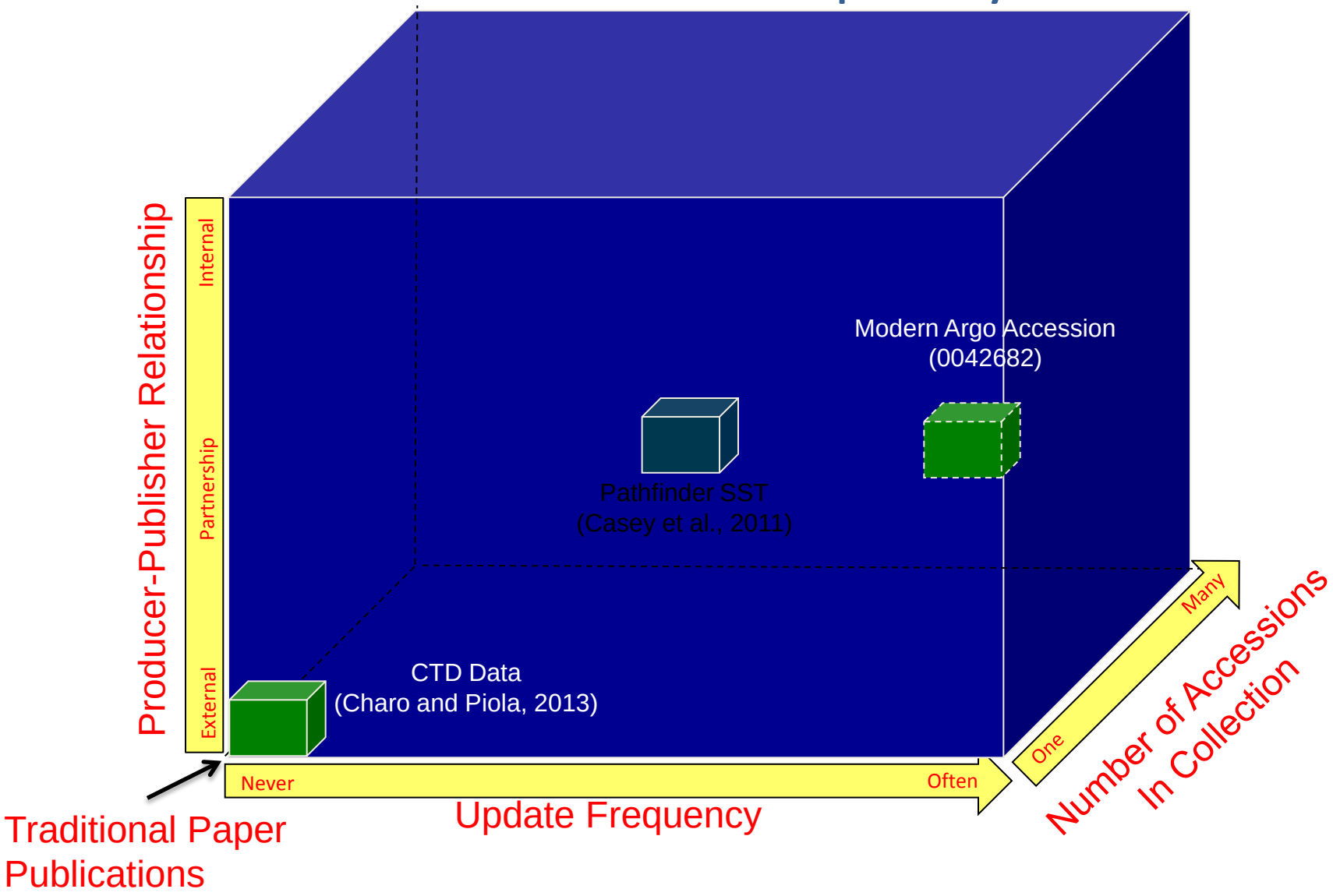
Argo and DOIs, aiming for the single DOI

NODC also working on an Argo DOIs

Current NODC plan:

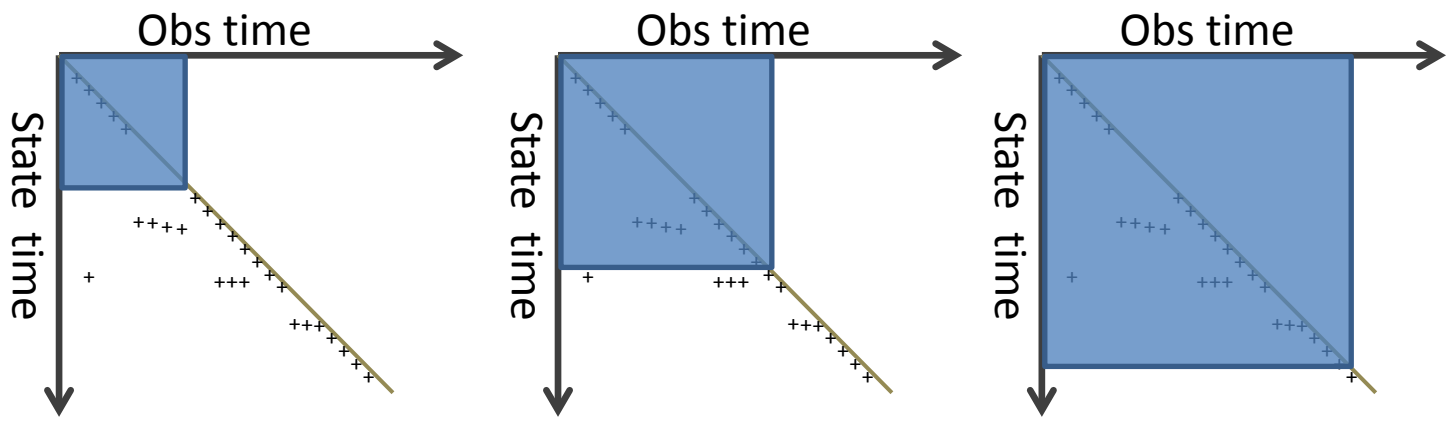
- Develop good ISO record for Argo accession 0042682
- Mint a single DOI
- Expose on the Argo DOI landing page the previous versions and dates of their publication
  - http://www.nodc.noaa.gov/archive/arc0022/0042682/
- Still need to figure out how to deal with the pre-1995 Argo accessions archived

NODC Dimensions of DOI Complexity



How to cite NODC Argo Accession (0042682)?

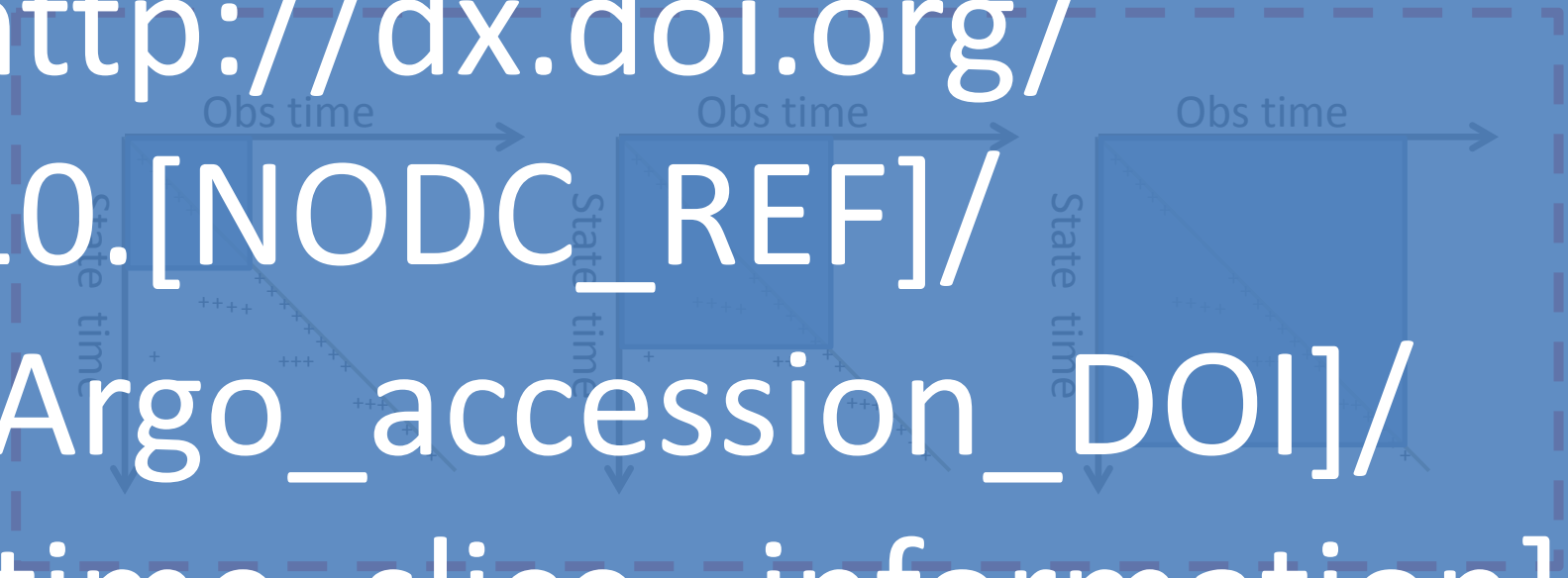
NODC Archive (collection of snapshots/granules)



To cite a particular snapshot one can potentially cite a time slice of the NODC archive i.e. the snapshot at a given point in time.

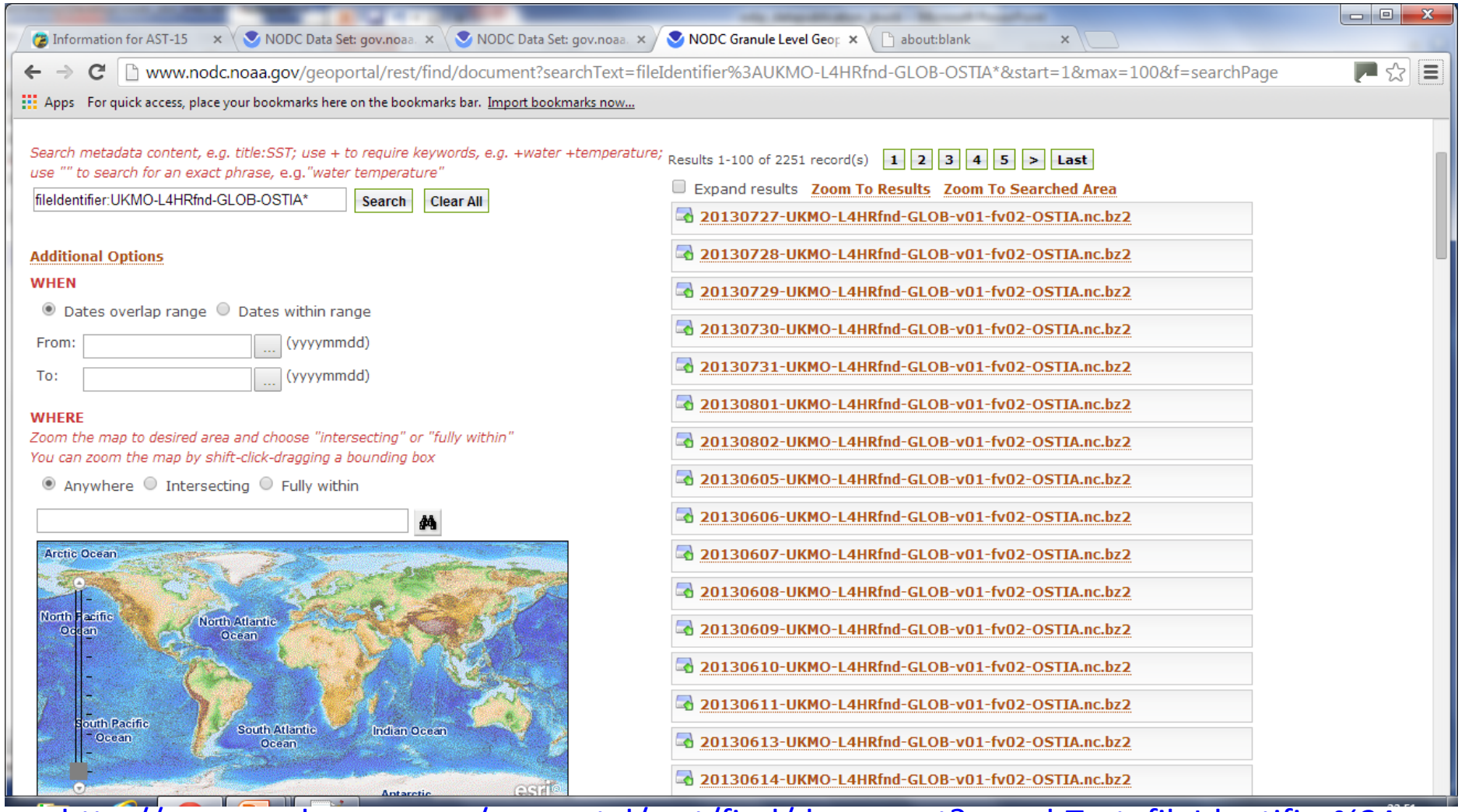
How to cite NODC Argo Accession (0042682)?

http://dx.doi.org/
10.[NODC_REF]/
[Argo_accession_DOI]/
[time_slice _information]



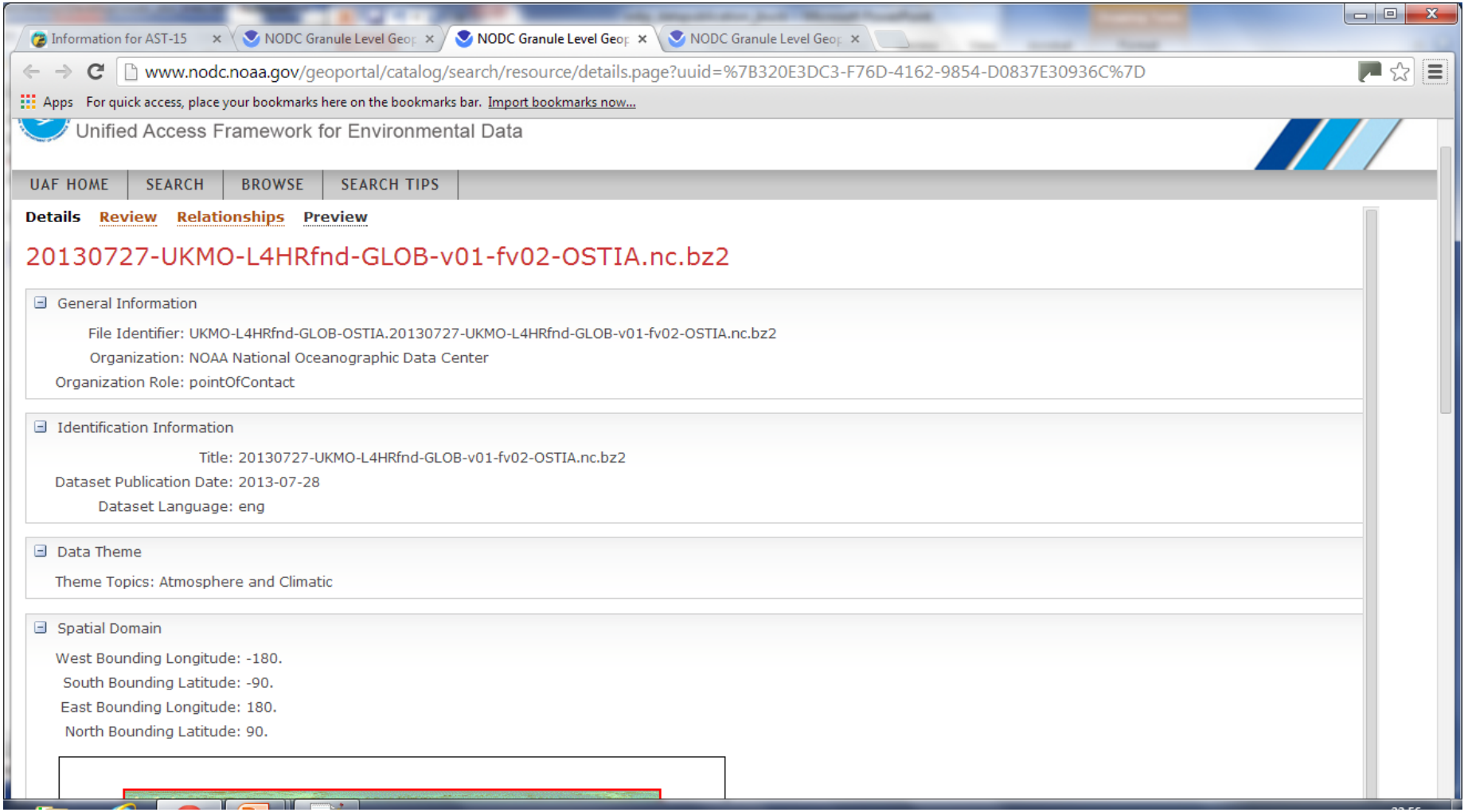
To cite a particular snapshot one can potentially cite a time slice of the NODC archive i.e. the snapshot at a given point in time.

Citing individual granules within single DOI (akin to GLOB OSTIA dataset access at NODC)



http://www.nodc.noaa.gov/geoportal/rest/find/document?searchText=fileIdentifier%3AUKMO-L4HRfnd-GLOB-OSTIA*&start=1&max=100&f=searchPage

Citing individual granules within single DOI (akin to GLOB OSTIA dataset access at NODC)



http://www.nodc.noaa.gov/geoportal/catalog/search/resource/details.page?uuid=%7B320E3DC3-F76D-4162-9854-D0837E30936C%7D

Expanding the granularity

Are types of granularity beyond the 'state time' are required for Argo?

- Spatial criteria
- Temporal criteria
- Individual floats
- Others?

Once this is known, the granules that need landing pages could be produced.

Conclusions

- Argo should aim to enable reproducible research in its data citation and archive infrastructure.
- Work by DataCite and EUDAT has now precisely defined needs and approaches for citation of dynamic data.
- The first approach is operational at Ifremer (via multiple DOIs).
- NODC has a proposed approach which brings us closer to the single DOI.
- The exact granularity within the dataset required by Argo needs to be defined.

Thank you for your attention

Questions?