Report on RDA Dynamic Data Citation Working Group Workshop

1st & 2nd July 2014

Organisers: RDA Data Citation WG, Centre for Ecology and Hydrology (CEH), British Library (BL)

Attendees: John Watkins (CEH), Anita Weatherby (CEH), Rick Stuart (CEH), Kate Harrison (CEH), Harry Dixon (CEH), Sarah Callaghan (STFC), Andi Rauber (RDA), Stefan Pröll (RDA), Justin Buck (BODC), Rod Bowie (BGS), Elizabeth Newbold (BL/DataCite), Sergio Ruiz (DataCite), Louise Corti (UKDA), Sue Rennie (CEH/ECN), David Leaver (CEH), Jane Hunter (ITEE), Jonathan Tedds (UKEOF/RDA).

Aims:

- To present RDA WG conceptual model addressing citation of dynamic data to a group of data curation practitioners
- To assess goodness of fit of the model for the requirements of users, curators, publishers, authors
- To extend and/or improve the model to meet the widest range of data users
- To plan test implementations of the citation model with various dynamic data curated by the group
- To provide input to WG reporting at RDA P4

Summary

Member of the Research Data Alliance (RDA) working group on dynamic data citation held a 2 day workshop at the British Library on the 1st and 2nd July 2014 in order to explore a proposed citation model against a number of research community use cases. The participants mainly represented the UK Natural Environment Research Council data centres, the UK Data Archive of the Economic and Social Research Council, the British Library and DataCite. Through a number of facilitated sessions, the participants explored the issues around the proposed model and possible improvements or adaptations for their own user communities. A number of currently used pragmatic solutions were presented and explored. Finally, possible steps forward were proposed for a few of the use cases presented that most addressed the issues and range of use cases presented. The participants would make best efforts to progress these steps forward with the immediate resources available to them and share outcomes with the RDA WG. The possibility of joint proposals for funding of these developments was discussed as a possible route to address resourcing issues.

Record of the meeting.

Day 1

1. Introduction

JW welcomed attendees and introduced the meeting aims. AR gave an introduction to the RDA and the aims of the Scalable Dynamic Data Citation Working Group. Attendees each introduced themselves and their interest in this meeting. AR gave a presentation on the Working Group proposal for a model for enabling data citation using timestamped, versioned datasets and assignment of persistent identifiers to user-defined queries.

2. User perspectives

Attendees took part in an exercise to identify where the proposed model worked well or did not work so well in meeting the needs of four perspectives: data user, data depositor, publisher and repository. This exercise followed a "carousel" approach, for which each of four flip-charts was labelled with one of the perspectives. Attendees split into four groups, each started at one flip-chart to consider how the model met the needs from one perspective. After 10 minutes each group moved round to the next flip-chart to review input from the previous group(s) and add their own. This was repeated until each group had considered each perspective. Attendees were each given five stickers and asked to vote by sticking these next to the points they felt were most important to address. The points with the most votes were as follows (* indicates number of stickers):

Pros and Cons of the Data Citation Model from 4 user perspectives:

Data publishers:

☺

- Increase data linkage and reuse **

☹

- How to resolve peer review (makes it more complicated) **
- Publishers need to develop / link to tools to manage dynamic data **
- Lots of broken PIDs potentially *
- Potentially too granular for publishers ***

Data repositories:

☺

- Establishes best practice / standards *
- Promotion of use of data centre services / metrics ****
- Provides reproducibility of historic data *

☹

- Do I need new infrastructure / new landing pages *
- Resources / costs *****
- Does not work for file based archives ***
- Does this work for small data centres **

Data Depositors:

☺

- Recognition ******
- Access to full history of different versions ****

☹

- Data depositor needs to maintain database. *
- Will require overhead of knowledge & support *
- Need a way to flag issues /points of contention esp. if corrections ****

Data Users:

☺

- Enables persistent access route to specific referenced data ***
- Allows awareness of related datasets (versions etc)**

☹

- Lack of info around extent of change/similarity between subsets/timestamps***
- Not clear what citation string would be /how generated***

The full list of issues identified is given in Annex A.

3. Community perspectives

Four of the attendees presented case studies illustrating their current approaches to scalable dynamic data citation. Each presenter was asked to identify any key points on how the Working Group model met their needs well or would not meet their needs and these were captured on post-it notes. The presentations and issues raised are listed below.

UK Butterfly Monitoring Scheme – John Watkins

☺

- Provides a start for dynamic citation standards based
- Reproducibility for future for metrics

- Give accreditation for UKBMS

☹

- How does it work with DataCite syntax – currently uses DOIs
- Infrastructure issues –, how to version, maintain and improve flat file infrastructure

Argo Buoy Network – Justin Buck

☺

- Could apply model to snapshots – if agreed referencing (DOI to series – subset record time slice. Could have database on top to manage master listing of files
- Priority is time slices, then address regional subsets of data
- Model is good, need to address applying to legacy data

☹

- No file versioning
- Resource not available to implement

National River Flow Archive – Harry Dixon

☺

- Could use model to snapshot subsets for data centre to publish – but not user defined PIDs

☹

- Want citation for whole dataset not subsets
- If assign PIDS to every download is big change to data centre policy – presently we try to encourage users to examine details metadata for each site (e.g. 10 PIDs for set of sites) before downloading an aggregated data set

UK Data Archive – Louise Corti

☺

- UKDA – Already have systems of query-based ID for sub-set (no time deltas) + autogenerated citations

☹

- UKDA issue – also need more informal IDs (not DataCite) for quantitative subsets

After all the presentations had been given, the following summary of issues from the use cases was constructed:

- Reproducibility
- Links to DataCite
- Accreditation – how DOI is issued to give credit to originator
- Infrastructure / resourcing
- Legacy – before DOIs – versioning
- Granulation – beyond time slicing
- Cultural issues – data centre defined & user defined query IDs
- UKDA implementation as a reference – link to DOIs – wider use
- When are DOIs needed? Other TR metrics

Day 2

4. Introduction to DataCite

EB gave a presentation introducing DataCite

5. Generating ideas

JW recapped the issues identified on day 1 and the group discussed how best to work to address them. Attendees each identified ideas for the priorities they would like to address, these could be specific solutions for issues raised or more general approaches. Each attendee wrote their ideas on post-it notes, then stuck these on a board. JW worked with the group to identify groupings of ideas. The groups and ideas identified are listed below:

Policy/Recommendations:

- Recommendations for dealing with legacy issues
- Define minimum expectations for dynamic data citations, reproducibility etc
- Guidance & policy from DataCite when solutions found
- Citation metrics on data sets -> links to RDA bibliometrics WG etc
- Is it possible or desirable to assign a DOI to every use defined query on a database (if they need to be formally minted)
- Need case studies to show how different approaches work to help others make decisions
- Provide 1 or 2 examples of a dynamic data citation workflow to feed into new RDA workflow

Versioning:

- How do we deal with collections?
- How do we obtain full history of versions?
- Recommendations for applying versioning on an existing database
- Standard approach to versioning complete datasets not subsets
- Time stamping of RDF database linked data
- How to present the changes which have been introduced between versions?

Citation generation:

- How to extend APA model of paragraph citation to other subsets
- Integrating citation text generation with dynamic data set identification
- Citation text for complex queries – could be impractically long if many authors and institutes
- What / how are we going to create citation strings
- What will they look like
- Can we give credit for version / snapshots
- Semantics in identifiers vs opaque identifiers
- Data citation for derived aggregated data sets
- Citatation to individual researchers eg cit sci bird data

Subset PIDs:

- 2 types of DOI – DOI for specific data set – DOI / fragment ID for specific subset
- I want to cite a series (collection) and the fragments (datasets) that make up the series & be able to collect metrics on use of both
- Change to syntax of DOI? Series + fragment like journal + vol / pageno?
- Data Centres mint DOIs for series but user defined fragment DOIs which include the series part?
- Requirement to cite the whole dataset / collection with DOI (even if it is changing or growing but also ability to cite subset
- Fragment identifiers
- Extend DOIs for parts/fragments
- Don't need to assign DOI to everything – use GUIDs / URIs to identify sections /subsets within dataset
- Subsetting of data inside files – need service with records of parameters in files (see climate models)
- File based system – low tech version of a 'manifest' simply a list of files in a particular dataset – can use as list to query to retrieve them
-

6. Exploring solutions

The group revisited the list of use cases described on day 1 and selected three to explore further. Participants split into three groups to work on these:

- Use case 1 – NRFA
- Use Case 2 - UKBMS
- Use Case 3 - ARGO

For each use case, each group identified practical steps that could be taken forward. The outputs were reported back to the main group as follows.

Use case 1 - NRFA

- Implement solution on a dataset (of which the NRFA has two primary ones) basis NOT for the whole NRFA database. The below describes a possible solution for the 'Gauged Daily Flow' dataset.

- Users currently access the main 'WORKING' Gauged Daily Flow tables on ORACLE. The proposal would build a second 'PUBLIC' copy of these tables with the same structure. Once the system is running user access would be switched from the 'WORKING' tables to these 'PUBLIC' tables.

- The implementation steps are therefore:
  1) Build 'PUBLIC' copy of dataset's 'WORKING' time-series tables and their related history (known as 'AUDIT') tables. These copies would have an identical structure.
  2) Develop and build a PID storage table to contain the PID, timestamp of issue and hash key on complete dataset.
  3) Decide on and implement an update timetable where by the data in the 'WORKING' tables are copied to the 'PUBLIC' tables (perhaps daily, weekly, etc.).
  4) When data in the 'PUBLIC' tables is updated then a hash key is calculated across the whole dataset and compared to the PID store. If the hash key doesn't exist then a new PID is issued and its information recorded in the PID store.
  5) Create one landing page for all dataset PIDs which contains basic version information (e.g. key changes from previous versions).
  6) User access moved to the new 'PUBLIC' tables.

- At a later point the following extension would be added:
  7) Develop tools (internally and externally) to allow users to access previous versions. This would probably involve some form of user defined subset filter to avoid performance issues with recreating the whole dataset when a users only requires a certain part.

Use case 2 – UKBMS

Option 1 - A single DOI for dynamically changing dataset (without PID etc). The landing page would only provide access to the current version of the whole dataset i.e. no way of accessing the dataset or subset at the point it was referenced. This is not currently possible as a solution for DataCite DOIs.

Option 2 – obtain a DOI for a 'series' or collection of static copies of the 'trends' datasets deposited at EIDC Hub. The landing page would not itself provide access to the current version of the whole dataset but list the related DOIs for the related datasets (subsets or versions at the point it was referenced).

Option 3 - UKBMS implement model. DOI given for raw dynamic data which can only be used with a PID based on a user-defined query.

a. UKMBS need:

i. An interface would be needed for generating queries on dynamic dataset (need to decide if querying and issue of PID is run in-house or by public users)
ii. mechanism to generate PID for specific subset of data tied to query tool
iii. to store query and timestamp.
iv. to create hash file for resultant data.
v. Create a database table of updates to raw data in order to run stored queries on data and check against hash file.
vi. Guarantee continued resource against this overhead

Use case 3 - ARGO

Fit to proposed RDA model

The approach provides the reproducibility and single DOI for Argo that the Steering Team desires. The monthly snapshots are effectively data centre defined subsets of the Argo data. The subsets can potentially be extended to a finer granularity (e.g. by ocean region) if there is a science need.

In the long term a database based on the RDA model can potentially sit on top of the dataset. Limited resources mean it is not possible in the next couple of years. The NODC single archive is a step towards following the common approach presented by the RDA.

Next steps

Data citation and publication community approval for the proposed approach of minting a single DOI at NODC is needed. Resource at NODC to build a prototype system is currently being confirmed. Once this is approved and the prototype is built, a data publication in a recognised data journal (e.g. Scientific Data) describing the data is necessary.

7. Next steps

The following next steps were agreed:

- A report of the workshop would be produced and circulated to the participants and then to the RDA Dynamic Data Citation WG.
- This would be used together with follow-up work and discussion to improve the citation model as it stands
- Participants from the workshop would make best efforts under their current resource to implement steps toward dynamic data citation either within the three use cases explore or within similar use cases aligned to the technical and cultural issues identified
- The UKDA tools for dynamic citation would be used as a reference model especially for sub setting of text and other qualitative data
- For additional resourcing of dynamic data citation developments, those wanting to developing new methods should consider grant proposals (such as through H2020) that

would align with the RDA WG model and/or with others trying to do the same. This could be coordinated through the RDA WG

- The report of the workshop and any immediate follow-ups will be reported at the RDA 4th Plenary in Amsterdam in September.

Annex A – Full record of issues produced by sessions in the meeting

2. **Consideration of the citation model from 4 user perspectives:**

DOES THE MODEL MEET THE NEEDS OF PUBLISHERS

☺ Attempt to address real issue – saves replication

They know its stable long term

Increase data linkage and reuse **

Prefer static data slices

Could support disciplinary difference in citation

☹ How will this look needs human readable form

How much context in human citation

Complex for indexes e.g. Thomson Reuters for metrics

How to resolve peer review (makes it more complicated) **

Publishers need to develop / link to tools to manage dynamic data **

Potentially a large lengthy, complicated citation not good for journals

Lots of broken PIDs potentially *

Potentially too granular for publishers ***

Managing submissions that include datasets

Manage errata in journals

Potential problem with dealing with large numbers of citations

DOES THE MODEL MEET REPOSITORY NEEDS?

☺ Good for new data

Establishes best practice / standards *

Easy re use for data access

Promotion of use of data centre services / metrics ****

Reduce duplication of data / snap shots

Provides reproducibility of historic data *

Links to journal requirements

Could work for static and dynamic data

☹

Difficult for old data / new versioning required

Do I need new infrastructure / new landing pages *

Where do I get expertise / help with implementation

Resources / costs *****

Lots of PIDs to manage

If 2 papers use most of the same data, there are still different PIDs **

Do we PID different queries that give the same data results? *

How does one PID relate to a range of PIDs (i.e. show me related PIDs)

More complex for data centre management

Does not work for file based archives ***

Does this work for small data centres **

Does it work for multi format data sets

Attribution for data centre information / accreditation

Subsets from multiple data sources?

Resilience of model?

Assumes time stamping can be implemented on data set

Security issues in automatic PID generation (DOS attack)

DOES THIS MEET DATA DEPOSITORS NEEDS

☹ data depositor needs to maintain database. *

Will require overhead of knowledge & support *

☺recognition ******

☹ No attribution

☹ Concerns about loss of control because machine readable

Not human ☺

☹ Might need to re-write depositor database to fit in with system

☹ if different data centres exchange data make it more complicated to transfer

☺ Could correct errors found later

☺ Access to full history of different versions ****

☹ Potential confusion when observations added

Not want to perpetuate spread of errors that have been corrected later

> 😐 Not a major concern so long as human (need this as well) mediated overview information
> **** Need a way to flag issues /points of contention esp. if corrections

☹ Need to support greater contextual metadata e.g. prepossessing, QA, QC

How specify if team produced dataset for attribution (cultural issue?)

☺ Increased impact of data because more people use it

☺ Give metrics for data usage -> justify future funding

☹ Practicable increase no of DOIs

☺ Loss limit now on citation easier citation - data paper make it easier

☹ if contributed data set that has multiple IDs -> lack of recognition, complicated

☺ By being able to cite additions, changes with reference to original data set makes versioning genuinely possible + resolve confusion

☺ Helps provide credit for secondary re-use

☺ Will help incentivise people to produce better documented datasets

DOES THE MODEL MEET DATA USER NEEDS?

☺ enables persistent access route to specific referenced data ***

☺ allows awareness of related datasets (versions etc)**

☺ user acknowledgment / credit in defining query

☺ credit for re-use of data

☺ efficiency and ease of query re-run

☺ ease of handling (smaller packages of data)

☹ lack of info around extent of change/similarity between subsets/timestamps***

☹ lack of human readable context info

☹ not clear what citation string would be /how generated***

☹ potentially large numbers of DOIs may be generated over time

☹ need to be sure of which subset

☹ need citation analysis to be aware of data re-use awareness of feedback

USE CASE ISSUES

☺

UKDA – Already have systems of query-based ID for sub-set (no time deltas) + autogenerated citations

☹ UKDA issue – also need more informal IDs (not datacite) for quantitative subsets

UKBMS

☺ Provides a start for dynamic citation standards based

☺ Reproducibility for future for metrics

☹ How does it work with datacite syntax – now use DOIs

☺ Give accreditation for UKBMS

☹ Infrastructure issues – flat files, how to version, maintain and improve infrastructure

ARGO

☹ No file versioning

☹ resource not available to implement

☺ Could apply model to snapshots – if agreed referencing (DOI to series – subset record timeslice. Could have database on top (master listing of files) of archive

☺ Priority is timeslices, then address regional datasets

☺ model is good, need to address legacy issue

NRFA

☹ Want citation for whole dataset not subsets

If assign PIDS to every download is big change – now try to encourage users to examine details metadata for each site (e.g. 10 PIDs for set of sites)

☺ Could use model to snapshot subset for data centre to publish – not user defined

SUMMARY OF ISSUES FROM USE CASES

Reproducibility

Links to DataCite

Accreditation – how DOI is issued to give credit to originator

Infrastructure / resourcing

Legacy – before DOIs – versioning

Granulation – beyond time slicing

Cultural issues – data centre defined & user defined query

UKDA implementation – link to DOIs – wider use

When are DOIs needed? Other TR metrics

CLUSTERS

POLICY / RECOMMENDATIONS

Recommendations for dealing with legacy issues

Define minimum expectations for dynamic data citations, reproducibility etc

Guidance & policy from DataCite when solutions found

Citation metrics on data sets -> links to RDA bibliometrics WG etc

Is it possible or desirable to assign a DOI to every use defined query on a database (if they need to nee formally minted

Need case studies to show how different approaches work to help others make decisions

Provide 1 or 2 examples of a dynamic data citation workflow to feed into new RDA workflow RDA

VERSIONING

How do we deal with collections

How do we obtain full history of versions

Recommendations for applying versioning on an existing database

Standard approach to versioning complete datasets not subsets

Time stamping of RDF database linked data

How to present the change which have been introduced between versions

CITATION GENERATION

How to extend APA model of paragraph citation to other subsets

Integratin citation text generation with dynamic data set identification

Citation text for complex queries – could be impractically long if many authors and institutes

What / how are we going to create citation strings

What will they look like

Can we give credit for version / snapshots

Sematics in identifiers vs opaque identifiers

Data citation for derived aggregated data sets

Citatation to individual researchers eg cit sci bird data

SUBSET PIDs

2 types of DOI – DOI for specific data set – DOI / fragment ID for specific subset

I want to cite a series (collection) and the fragments (datasets) that make up the series & be able to collect metrics on use of both

Change to syntax of DOI? Series + fragment like journal + vol / pageno?

Data Centres mint DOIs for series but user defined fragments DOIs which include the series part?

Requirement to cite the whole dataset / collection with DOI (even if it is changing or growing but also ability to cite subset

Fragment identifiers

Extend DOIs for parts/fragments

Don't need to assign DOI to everything – use GUIDs / URIs to identify sections /subsets within dataset

Subsetting of data inside files – need service with records of parameters in files (see climate models)

File based system – low tech version of a 'manifest' simply a list of files in a particular dataset – can use as list to query to retrieve them

FACTS

If its to complex researchers will go elsewhere

We can cite data without using DOIs!

USE CASES CONSIDERED

UKDA Qualidata – as candidate implementation for subsetting of text documents to disseminate to RDA WG (esp use of APA)

UKBMS & Geology Lexicon – DB & CSV currently non versioned

NRFA – specific versioning of RBDMS required

ARGO & Chilbolton – Flat files, legacy and versioning issues – should fit to Chilbolton

USE CASE 2: ARGO [CHILBOLTON]

- Fit database on top of file system – infrastructure doesn't exist – need approval from governance committee
- OSTIA data – landing page – accession containing all data – lists all datasets
- Reproducibility
- Single citation for all ARGO
- Data centre pre-defined subsets – user defined subsets harder to implement

Community approval and accepted approach to PIDs/DOIs/Citation

Annex B – Use Case Write-ups

1) NRFA

RDA WG Data Citation – Use Case Details

Use Case Name:

UK National River Flow Archive

Institution:

Centre for Ecology and Hydrology of the NERC in the UK (http://www.ceh.ac.uk/data/nrfa)

Scenario:

Domain:

Data Characteristics:

This dataset is maintained by CEH on behalf of UK and Devolved Government. The NRFA collates, quality controls, and archives hydrometric data from gauging station networks across the UK including the extensive networks operated by the Environment Agency (England), Natural Resources Wales, the Scottish Environment Protection Agency and the Rivers Agency (Northern Ireland). The NRFA data underpin much of the hydrological research and water resources management activity in the UK, with a wide user base across the research, regulatory, commercial, education and policy sectors.

The primary NRFA datasets consist of time series data (predominantly daily resolution) of river flow from ~1500 points across the UK. The database currently holds around 50 000 station-years of flow data, with an average record length of 34 years per station.

Type:

Daily and irregular time series of river discharge at defined geographical points.

Storage:

Data are stored on an ORACLE RDMS in a structure which logs all changes to the time series in audit tables.

Access:

Data are distributed through a web portal on the CEH website which allows data discovery, view and download (via .csv files, > 10,000 downloads p.a.). Additionally data are provided via a manual enquires service.

Current citation approach:

No current suggested citation method, although the T&C under which the data are distributed do place a requirement on users to acknowledge the data source.

Ideal way of citing:

Key requirements from the data centre's perspective are:

a) the ability to continuously update data (on a near daily basis) so that the most recent data is always available to users;

b) the ability to store changes to the data not snapshots of the complete database.

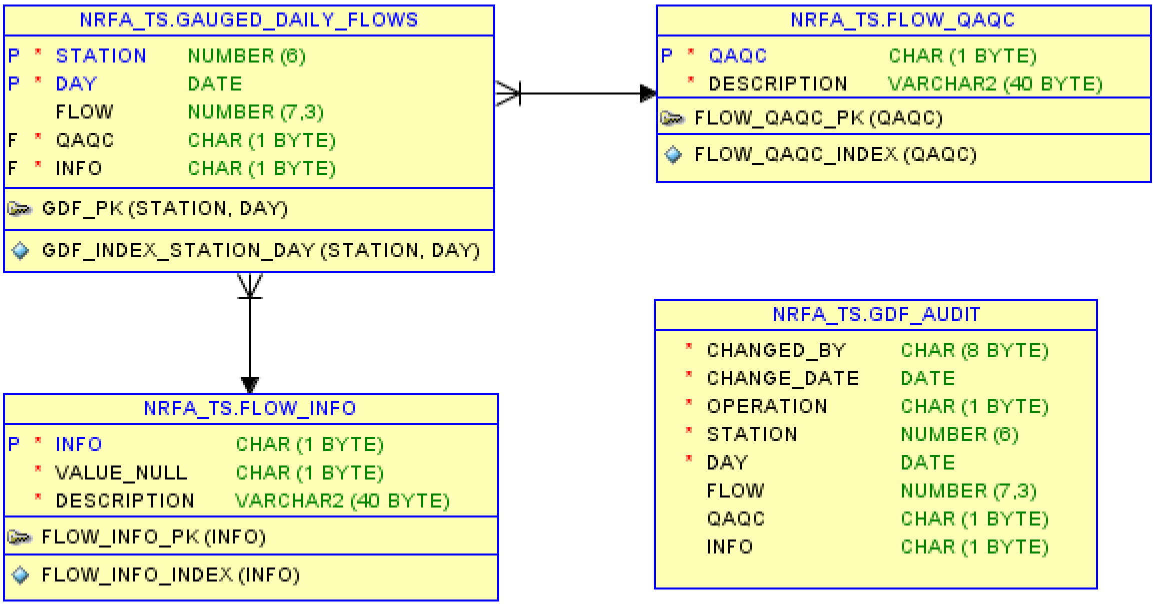
Key requirement from a user's perspective are:
a) the ability to cite data in a consistent way showing when the data was accessed. For example, the data is used by consultants in flood risk assessments for new developments and they must be able to say that the assessment was completed using data version X;
b) the ability to discover whether the currently available version of data differs from their downloaded version and what those differences are. For example, regulatory authorities need to be able to rerun flood risk assessments and understand if (and what) differences exist between the most recently available data and that used by the consultant.

Other aspects:

Contact info:
Harry Dixon (harr at ceh dot ac dot uk)

Details:

The data model:
e.g DB schema, structure of the data
Data are currently sorted in part of the NRFA ORACLE database. For example, daily flow data are stored in NRFA_TS.GAUGED_DAILY_FLOWS with triggers to enter details of changes into NRFA_TS.GDF_AUDIT.



```
NRFA_TS.GAUGED_DAILY_FLOWS
P * STATION   NUMBER (6)
P * DAY       DATE
    FLOW      NUMBER (7,3)
F * QAQC      CHAR (1 BYTE)
F * INFO      CHAR (1 BYTE)
GDF_PK (STATION, DAY)
GDF_INDEX_STATION_DAY (STATION, DAY)
```

```
NRFA_TS.FLOW_QAQC
P * QAQC          CHAR (1 BYTE)
  * DESCRIPTION   VARCHAR2 (40 BYTE)
FLOW_QAQC_PK (QAQC)
FLOW_QAQC_INDEX (QAQC)
```

```
NRFA_TS.FLOW_INFO
P * INFO          CHAR (1 BYTE)
  * VALUE_NULL    CHAR (1 BYTE)
  * DESCRIPTION   VARCHAR2 (40 BYTE)
FLOW_INFO_PK (INFO)
FLOW_INFO_INDEX (INFO)
```

```
NRFA_TS.GDF_AUDIT
  * CHANGED_BY    CHAR (8 BYTE)
  * CHANGE_DATE   DATE
  * OPERATION     CHAR (1 BYTE)
  * STATION       NUMBER (6)
  * DAY           DATE
    FLOW          NUMBER (7,3)
    QAQC          CHAR (1 BYTE)
    INFO          CHAR (1 BYTE)
```

Versioning/timestamping:
Current or planned
The most recent data is provided to users. No versioning is currently applies but details of changes to the dataset are stored internally within the database.

Dynamics:

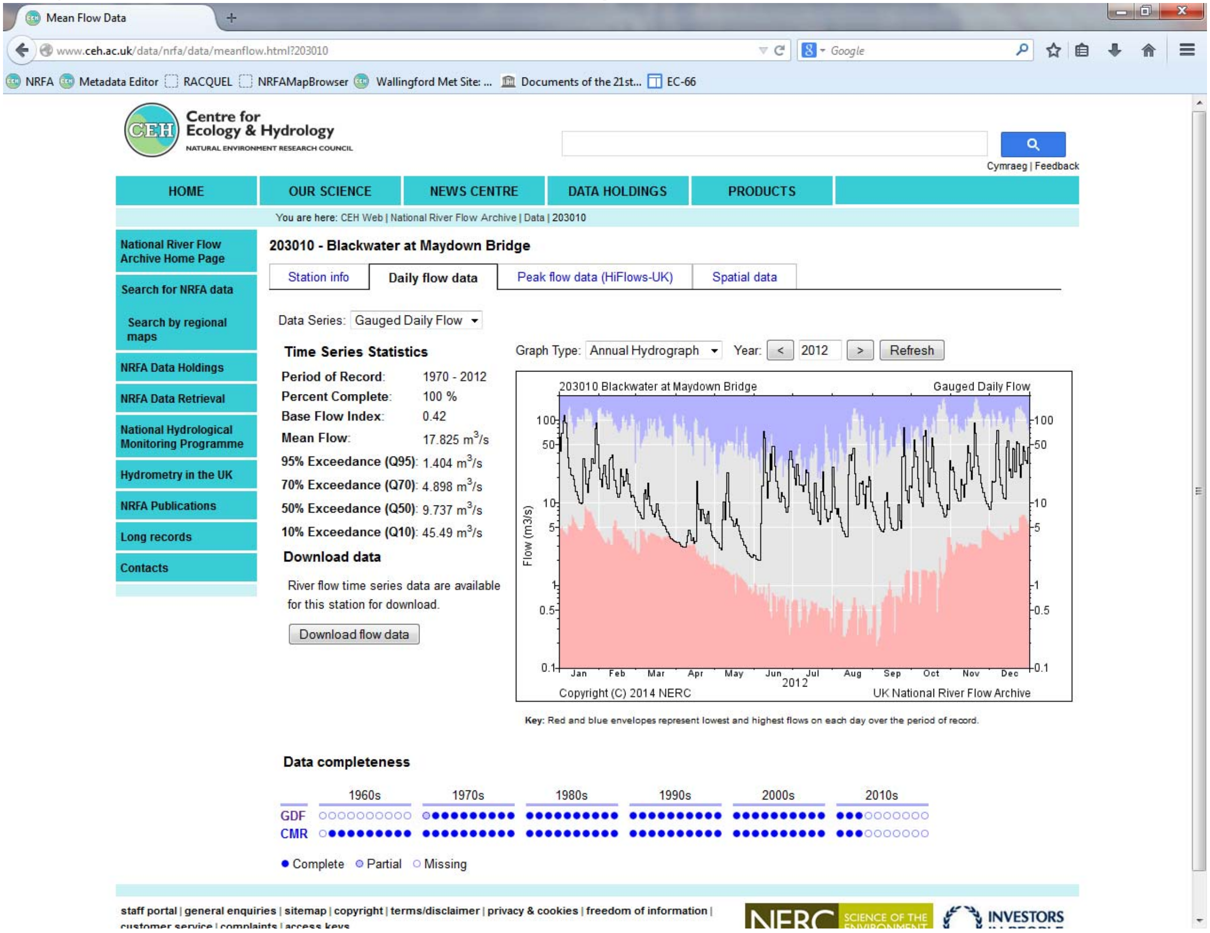
e.g. how much data, how much added in which time intervals, any corrections/updates or just additions

The daily river flow dataset is currently the largest time series dataset managed as part of the NRFA and contains 20,700,000 rows.

Major updates (addition and changes) to the data are made annually with smaller, sporadic updates processed throughout the year as and when corrections/improvements to data are required. 500,000 changes have been made to this dataset in the last 12 months.

Screenshots:

Interface/workbench that researchers are using to create subsets


Mean Flow Data
www.ceh.ac.uk/data/nrfa/data/meanflow.html?203010
Google
NRFA
Metadata Editor
RACQUEL
NRFAMapBrowser
Wallingford Met Site: ...
Documents of the 21st...
EC-66
Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL
Cymraeg | Feedback
HOME
OUR SCIENCE
NEWS CENTRE
DATA HOLDINGS
PRODUCTS
You are here: CEH Web | National River Flow Archive | Data | 203010
National River Flow Archive Home Page
Search for NRFA data
Search by regional maps
NRFA Data Holdings
NRFA Data Retrieval
National Hydrological Monitoring Programme
Hydrometry in the UK
NRFA Publications
Long records
Contacts
Station info
Daily flow data
Peak flow data (HiFlows-UK)
Spatial data
<
>
Refresh
Download flow data
203010 Blackwater at Maydown Bridge
Gauged Daily Flow
Flow (m3/s)
100
50
10
5
1
0.5
0.1
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2012
Copyright (C) 2014 NERC
UK National River Flow Archive
staff portal | general enquiries | sitemap | copyright | terms/disclaimer | privacy & cookies | freedom of information |
customer service | complaints | access keys
NERC SCIENCE OF THE ENVIRONMENT
INVESTORS IN PEOPLE

203010 - Blackwater at Maydown Bridge

Data Series: Gauged Daily Flow

Graph Type: Annual Hydrograph Year: 2012

Time Series Statistics

Period of Record:	1970 - 2012
Percent Complete:	100 %
Base Flow Index:	0.42
Mean Flow:	17.825 m^3/s
95% Exceedance (Q95):	1.404 m^3/s
70% Exceedance (Q70):	4.898 m^3/s
50% Exceedance (Q50):	9.737 m^3/s
10% Exceedance (Q10):	45.49 m^3/s

Download data

River flow time series data are available for this station for download.

Key: Red and blue envelopes represent lowest and highest flows on each day over the period of record.

Data completeness

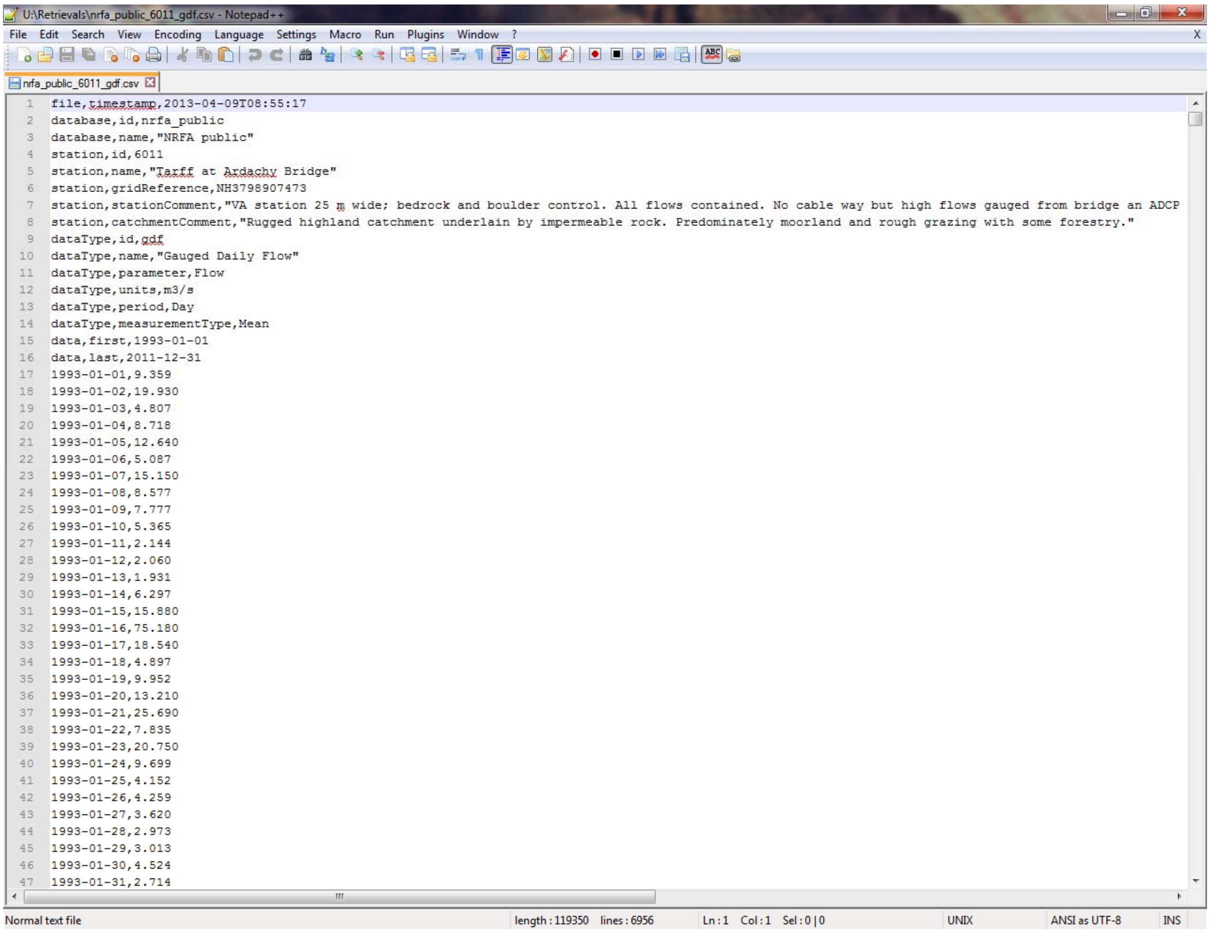
	1960s	1970s	1980s	1990s	2000s	2010s
GDF	○○○○○○○○○○	◉●●●●●●●●●	●●●●●●●●●●	●●●●●●●●●●	●●●●●●●●●●	●●●○○○○○○○
CMR	○●●●●●●●●●	●●●●●●●●●●	●●●●●●●●●●	●●●●●●●●●●	●●●●●●●●●●	●●●○○○○○○○

● Complete ◉ Partial ○ Missing

Example of subsets:

How they were created, what they look like, to get a feeling of what/how researchers would like to use the data and cite it.

Example single site time series in our current standard user export format:



```
file,timestamp,2013-04-09T08:55:17
database,id,nrfa_public
database,name,"NRFA public"
station,id,6011
station,name,"Tarff at Ardachy Bridge"
station,gridReference,NH3798907473
station,stationComment,"VA station 25 m wide; bedrock and boulder control. All flows contained. No cable way but high flows gauged from bridge an ADCP
station,catchmentComment,"Rugged highland catchment underlain by impermeable rock. Predominately moorland and rough grazing with some forestry."
dataType,id,gdf
dataType,name,"Gauged Daily Flow"
dataType,parameter,Flow
dataType,units,m3/s
dataType,period,Day
dataType,measurementType,Mean
data,first,1993-01-01
data,last,2011-12-31
1993-01-01,9.359
1993-01-02,19.930
1993-01-03,4.807
1993-01-04,8.718
1993-01-05,12.640
1993-01-06,5.087
1993-01-07,15.150
1993-01-08,8.577
1993-01-09,7.777
1993-01-10,5.365
1993-01-11,2.144
1993-01-12,2.060
1993-01-13,1.931
1993-01-14,6.297
1993-01-15,15.880
1993-01-16,75.180
1993-01-17,18.540
1993-01-18,4.897
1993-01-19,9.952
1993-01-20,13.210
1993-01-21,25.690
1993-01-22,7.835
1993-01-23,20.750
1993-01-24,9.699
1993-01-25,4.152
1993-01-26,4.259
1993-01-27,3.620
1993-01-28,2.973
1993-01-29,3.013
1993-01-30,4.524
1993-01-31,2.714
```

Proposed Possible Solution

- Implement solution on a dataset (of which the NRFA has two primary ones) basis NOT for the whole NRFA database. The below describes a possible solution for the 'Gauged Daily Flow' dataset.

- Users currently access the main 'WORKING' Gauged Daily Flow tables on ORACLE. The proposal would build a second 'PUBLIC' copy of these tables with the same structure. Once the system is running user access would be switched from the 'WORKING' tables to these 'PUBLIC' tables.

- The implementation steps are therefore:
  8) Build 'PUBLIC' copy of dataset's 'WORKING' time-series tables and their related history (known as 'AUDIT') tables. These copies would have an identical structure.
  9) Develop and build a PID storage table to contain the PID, timestamp of issue and hash key on complete dataset.
  10) Decide on and implement an update timetable where by the data in the 'WORKING' tables are copied to the 'PUBLIC' tables (perhaps daily, weekly, etc.).
  11) When data in the 'PUBLIC' tables is updated then a hash key is calculated across the whole dataset and compared to the PID store. If the hash key doesn't exist then a new PID is issued and its information recorded in the PID store.
  12) Create one landing page for all dataset PIDs which contains basic version information (e.g. key changes from previous versions).
  13) User access moved to the new 'PUBLIC' tables.

- At a later point the following extension would be added:
  14) Develop tools (internally and externally) to allow users to access previous versions. This would probably involve some form of user defined subset filter to avoid performance issues with recreating the whole dataset when a users only requires a certain part.

2) UK Butterfly Monitoring Scheme (UKBMS)

Background summary

UKBMS is both a research group and a component repository of the federated Environmental Information Data Centre hosted at CEH:

Currently UKBMS continuously collate raw data from many contributor surveyors and load into a dynamic ORACLE schema. Analyses are run on the raw data annually and 'Trends' dataset are output.

These 'trends' dataset are deposited to EIDCHub (who issue DOIs on behalf of the federated EIDC. Datasets are stored as static copies and a new DOI allocated each time. This process is iterated each year when trends data are updated. A 'series' discovery metadata record highlights users to the existence of all of the related datasets.

Requirement

UKBMS do not want a series of DOIs but instead want a single DOI in order to reference the entire dynamic database for citation / credit purposes.

Solution

Option 1 - A single DOI for dynamically changing dataset (without PID etc). The landing page would only provide access to the current version of the whole dataset i.e. no way of accessing the dataset or subset at the point it was referenced. This is not currently possible as a solution for DataCite DOIs.

Option 2 – obtain a DOI for a 'series' or collection of static copies of the 'trends' datasets deposited at EIDC Hub. The landing page would not itself provide access to the current version of the whole dataset but list the related DOIs for the related datasets (subsets or versions at the point it was referenced).

Option 3 - UKBMS implement model. DOI given for raw dynamic data which can only be used with a PID based on a user-defined query.

a. UKMBS need:
    i. An interface would be needed for generating queries on dynamic dataset (need to decide if querying and issue of PID is run in-house or by public users)
    ii. mechanism to generate PID for specific subset of data tied to query tool
    iii. to store query and timestamp.
    iv. to create hash file for resultant data.
    v. Create a database table of updates to raw data in order to run stored queries on data and check against hash file.
    vi. Guarantee continued resource against this overhead

3) RDA WG Data Citation – Use Case Details – International Argo Programme

Use Case Name:

International Argo Programme data

Institution:

Various, 26 countries participate, these contributions include ten national level data assembly centres (DAC) and two mirrored Global Data Assembly Centres (GDAC) that aggregate data from the DACs. Full details are available from http://www.argo.ucsd.edu/ and http://www.argodatamgt.org/ .

DOI activity is based at Ifremer (France) and the US National Oceanographic Data Centre (NODC), with advice from the British Oceanographic Data Centre.

Scenario:

Domain:

Oceanography, meteorology

Data Characteristics:

The dataset contains over one million vertical profiles of temperature and salinity from autonomous freely drifting profiling robots along with trajectory, meta and technical data. Data have been collected over 15 years and collated from 10 national level Data Assembly Centres (DAC) at two mirrored Global Data Assembly Centres (GDAC). The data setup provides global coverage of the deep ocean at a 3 by 3 degree resolution.

Type:

Data are geophysical measurements of the ocean. Data are appended to constantly as new profiles arrive. Data are updated or revised as quality control and calibration are performed.

Storage:

A collection of NetCDF format files (multiple files per float and profile) on two mirrored GDACs. There is no version control of files at the GDACs. Resources to introduce file versioning to the GDAC infrastructure is not available in the near term. Long term archive of the data is done by the US National Oceanographic Data Centre (NODC). The long term archive is in the form of snapshots of the full dataset on a weekly frequency. NODC snapshots of Argo data exist for the last decade.

Access:

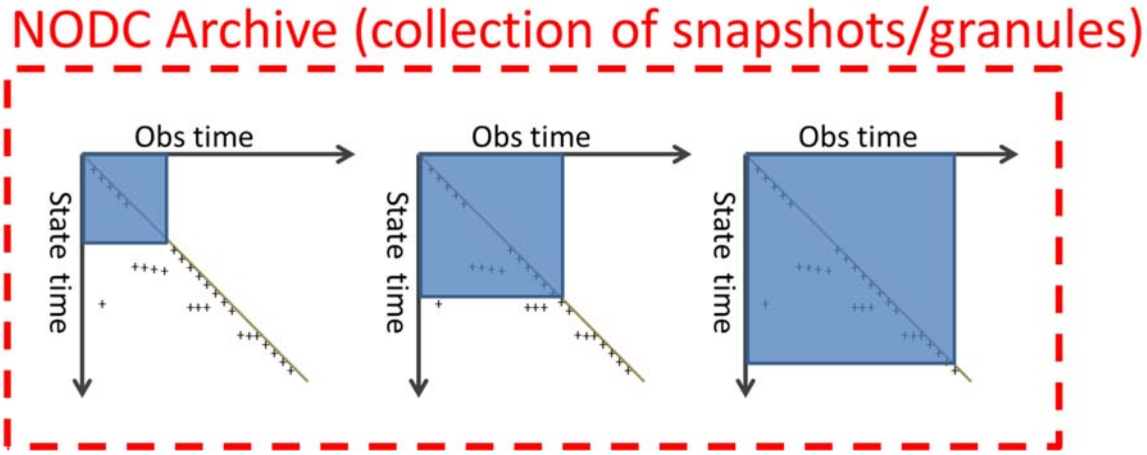
Data are freely available via ftp at both GADCs and from NODC. Only the current version of the data is currently available directly. Earlier snapshots of the Argo dataset are available from NODC on request.

Current citation approach:

Full copy of the GDACs archived every month by the Ifremer GDAC and DOI assigned to each version (snapshots in DataCite terminology).

Snapshots available at:
http://www.argodatamgt.org/Access-to-data/Argo-DOI-Digital-Object-Identifier
Documentation also has DOIs to preserve data descriptions.

Ideal way of citing:
A single DOI to reference the entire Argo dataset with the ability to append a granule reference at the end is strongly desired by the Argo Steering Team. This enables data reproducibility and easy identification of data usage in scientific literature.
The proposed approach for this is to mint a DOI for the NODC accession containing monthly snapshots of the Argo GDAC. In DataCite terminology a citation is effectively citing a time slice within the NODC accession where each time slice is a snapshot of the GDACs.



NODC store and deliver OSTIA data in an approach that is analogous (although the OSTIA accession does not have a DOI) to the proposed single DOI for Argo.

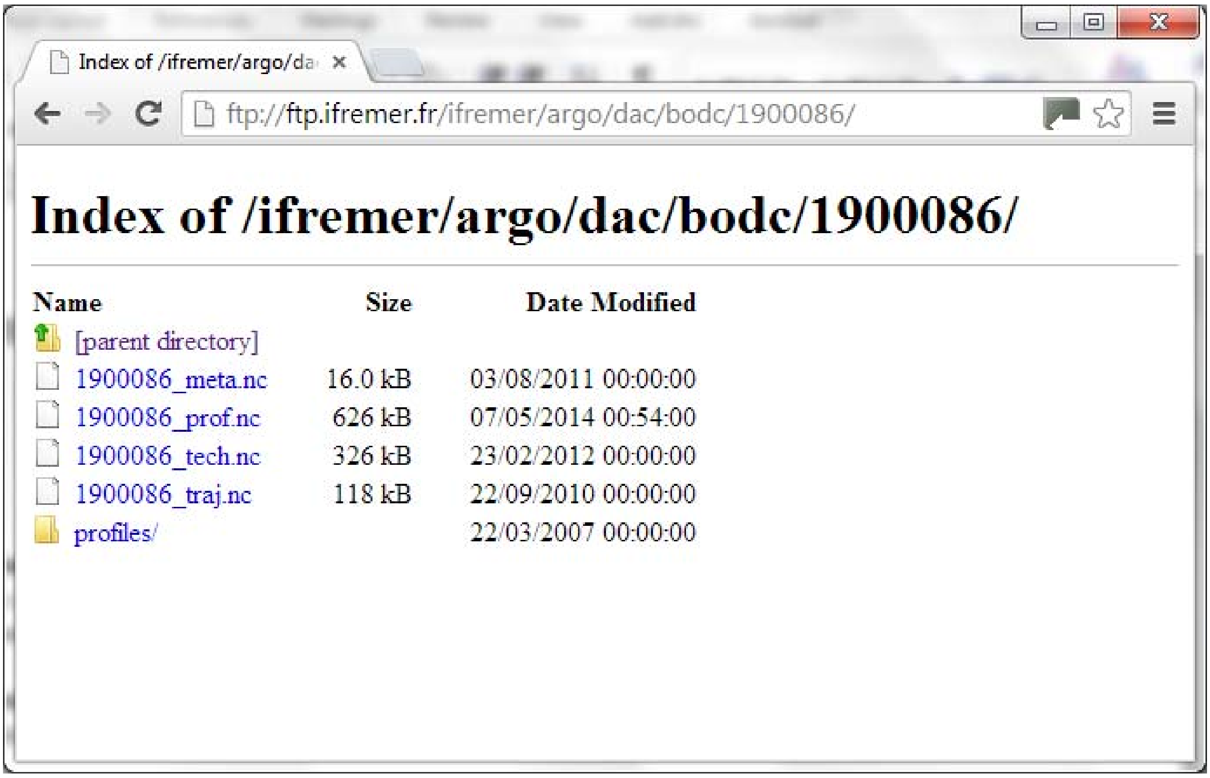
Other aspects:
The complexity of measurements is increasing with new sensors and inclusion of new oceanographic communities in the programme.

Contact info:
Justin Buck (juck at bodc dot ac dot uk)

Details:

The data model:
Data is currently in a file stored as a collection of NetCDF files organised by data assembly centre then float. Then each float has 4 types of file associated with it to fully describe the data:

- [Meta]data file
- [Traj]ectory data file
- [Tech]nical data file
- [Prof]ile data file (in a single merged file and individual profiles in a profiles subdirectory



ftp://ftp.ifremer.fr/ifremer/argo/dac/bodc/1900086/

Index of /ifremer/argo/dac/bodc/1900086/

Name	Size	Date Modified
[parent directory]		
1900086_meta.nc	16.0 kB	03/08/2011 00:00:00
1900086_prof.nc	626 kB	07/05/2014 00:54:00
1900086_tech.nc	326 kB	23/02/2012 00:00:00
1900086_traj.nc	118 kB	22/09/2010 00:00:00
profiles/		22/03/2007 00:00:00

Versioning/timestamping:

Current or planned

The GDACs do not support file versioning and only contain the latest version of the data.
The metadata within the files includes the file creation date.
Every month a snapshot is taken of the GDACs and a DOI assigned.
The proposed approach to reduce the number of identifiers is to place the snap shots in a single accession.

Dynamics:

e.g. how much data, how much added in which time intervals, any corrections/updates or just additions

Each profile is currently a few 100 kByte and this is increasing to a few Mbytes for the most recent floats. Data additions and updates are made almost constantly.

Screenshots:

Interface/workbench that researchers are using to create subsets

At present there are no plans to create subsets of the snap shots. The ability to cite a subset e.g. by float(s), spatial criteria, or temporal range would be beneficial.

Example of subsets:

How they were created, what they look like, to get a feeling of what/how researchers would like to use the data and cite it.

Subsets could be a combination of the following:

- Particular floats/platforms
- Spatial ranges

- Temporal ranges
- Quality control state

Fit to proposed RDA model

The approach provides the reproducibility and single DOI for Argo that the Steering Team desires. The monthly snapshots are effectively data centre defined subsets of the Argo data. The subsets can potentially be extended to a finer granularity (e.g. by ocean region) if there is a science need.

In the long term a database based on the RDA model can potentially sit on top on the dataset. Limited resources mean it is not possible in the next couple of years. The NODC single archive is a step towards following the common approach presented by the RDA.

Next steps

Data citation and publication community approval for the proposed approach of minting a single DOI at NODC is needed. Resource at NODC to build a prototype system is currently being confirmed. Once this is approved and the prototype is built, a data publication in a recognised data journal (e.g. Scientific Data) describing the data is necessary.