Dynamic data citation: summarised findings from initial meetings

Gerry Ryder, 21/7/16

During April and May 2016, staff from ANDS held a series of meetings with representatives from a number of research data intensive organisations identified as actively addressing challenges associated with the citation of dynamic data. The meetings were held on a one-to-one basis with each organisation and were intended to be an information gathering opportunity for ANDS. The notes provided here summarise the findings from those meetings. Additional information has been sourced to provide context as required. These notes are augmented by diagrams that visually represent the range of use cases and associated approaches to citation of dynamic data the meetings revealed.

Table of Contents

Defining dynamic data

Acknowledging there is no commonly agreed definition of 'dynamic data', Nick Car (Geoscience Australia) has initiated a controlled vocabulary to describe dynamic data. It is hoped that members of the DDC IG will contribute to its further development.

The types of data that may be labelled 'dynamic' include:

- where new data are regularly and systematically appended to an existing data set over time, e.g., with outputs from a satellite or sensor, and no changes are made to the existing dataset.
- pre-existing data in a large data set is modified or updated where:
  - errors are found in pre-existing data
  - new analytical and or processing techniques are applied to a select number of attributes/components of the existing data set
  - models and derivative products being revised with new data
  - the data itself is revised as processing methods are improved
- where data is accessed via Service where either or both of the query and the data being queried are dynamic

This last point was raised at a number of our meetings as a problem space. Participants mentioned that when considering approaches to the citation of dynamic data, whether the data to be accessed is large or small in size and whether access is via a repository (download) or via a portal (web service) are key considerations. In some cases, the same data may be accessed using either or both methods eg. a dataset may be downloaded via the AEKOS portal, or output via a service at the Biodiversity and Climate Change Virtual Laboratory (BCCVL).

As Wyborn et al (2016) have noted, large datasets accessed via Services present specific challenges for citation and long term access. It is becoming increasingly common for researchers to use Services to query and manipulate large scale data. In such cases, researchers can log in and dynamically create user-defined subsets for specific research projects. They may also mix and match data from multiple collections, each of which can have a complex history. Where the resulting dataset is itself large-scale, preserving a citable copy may not be feasible.

What are the drivers for enabling the citation of dynamic data?

4 key drivers for enabling dynamic data citation were identified by participants:

1. acknowledgement and impact assessment for data centre and/or data creator
2. enable reproducibility of published research
3. enhance data discovery
4. compliance with publishers' data availability policies

What are the issues associated with making dynamic data citable?

4 key issues identified by participants:

1. Form of the citation statement when citing dynamic data
2. Assigning a DOI or other identifier to dynamic data
3. Providing ongoing access to a subset or time-slice of data that is cited (eg in a publication)
4. Version management

DDC IG participant stories

Terrestrial Ecosystem Research Network (TERN)

TERN host and publish data in several different repositories based on the types of data being published. Hence, data management practices vary between repositories. Following is information about some of the different repositories used for data publication.

TERN Advanced Ecological Knowledge and Observation System (ÆKOS)

Summary

The Advanced Ecological Knowledge and Observation System (ÆKOS) is an online repository for Australia's ecological data. It is managed by the Eco-informatics Facility of TERN, which has partnered with a range of institutional data custodians to make ecological "plot" data (including quadrants, transects, pitfall traplines, cage trap arrays, and other systematic collection methods) more widely available to researchers. It also enables individual researchers to store and publish their data through the SHaRED data submission service.

- The AEKOS portal enables discovery and access (via download) to primary ecological data often provided by government agencies. The AEKOS portal supports various options for searching with selected results added to a download cart. Data selected for download are extracted using an offline batch process. The batch process produces a zipped file with the data and download instructions that is sent by email. The zipped file contains two sets of files to import data into either Postgres or MySQL databases, a Readme pdf giving instructions for installing and uploading the extracted data into the databases and a data description file.
- The SHaRED data submission service enables self-submission of any type of ecological dataset to be archived and published via AEKOS. This includes data derived from source data in the AEKOS portal. The service supports publishers' data availability requirements.

Key drivers for TERN Eco-informatics to enable citation of dynamic data are:

- acknowledgement and impact assessment for data centre and/or data creator
- compliance with publishers' data availability policies

Form of citation statement

Instructions are available for citing data obtained via AEKOS.

The format suggested is:

Original data author(s). (Year of Publication). Title, Version. DOI/Persistent hyperlink (if available). Obtained via ÆKOS Data Portal (http://www.aekos.org.au/), made available by University of Adelaide (http://www.adelaide.edu.au). Accessed [dd mmm yyyy, e.g. 01 Apr 2013].

A challenge identified by TERN Eco-informatics is ensuring consistent citation of data where it may also be consumed by, and output by, a Service – particularly where that Service does not support citation.

Assign a DOI or other identifier

AEKOS: DOIs are not assigned by TERN Eco-informatics to source data available in the AEKOS portal.

SHaRED: TERN Eco-informatics offers a DOI service for data published via SHaRED.
In cases where data and corresponding metadata is updated, the repository supports version control via the DOI. New versions are linked to older versions via metadata using the old DOI. In the case of annual and periodic data, new data is appended to previous published version (e.g., 2000-2015, add 2016 to give 2000-2016) and published as a full dataset with a new DOI to maintain the integrity of the data collected. This means all versions of a particular dataset are accessible.

Access to a subset or time-slice of data that is cited (eg in a publication)

ShaRED: no use case encountered to date where a dataset submitted to SHaRED was too large or complex to publish in its entirety.

AEKOS: Potential exists for all site-based data values to be updated (additions, modifications, deletions including sites and within sites). Datasets may be updated monthly, annually or periodically.

Version management

TERN Eco-informatics reports that they support incremental versioning and differentiate between updates supplied by the custodian and those based on changes to their model (includes updates to classification systems such as taxonomy).

Some data providers do not provide any curation information, so although a change may be detected, the reason for the change is unknown. Because the system supports site level and sometimes finer persistent identifiers, AEKOS staff need to be mindful of deletions as users may attempt to use a previously legitimate identifier for a data product that no longer exists.

Legacy versions are stored offline, but could be recreated if absolutely necessary (e.g., a legal case). There has been no requirement for this to date.

TERN OzFlux

Summary

The OzFlux network consists of nearly 30 flux towers in Australia and New Zealand, many of which are also members of the Australian Supersite Network (ASN). OzFlux is also a member of the global FluxNet community.

The Flux data is half-hourly time-series of various micro-meteorology parameters that are published at different quality control level (L1 - L6). Data from L3 and beyond are reusable for different applications. In the current data publication workflow, the flux data is processed annually and published as a separate netCDF file.

Data are available from the OzFlux data portal. The data are organised into collections with each collection representing at least one site. Metadata is written at instrument level (Flux tower), and yearly data files are appended to the same metadata. A handle is assigned at metadata level, and no persistent identifier is available at data-level. The datasets change due to improvement in the processing algorithm, identification of bugs, etc. Datasets are frequently appended with new data.

Form of citation statement

Instructions for citing data are provided in each metadata record.

Eg. *If you make use of this collection in your research, please cite:*

Jason Beringer (2013) Adelaide River OzFlux tower site OzFlux: Australian and New Zealand Flux Research and Monitoring hdl: 102.100.100/14228

Assign a DOI or other identifier

A handle is assigned at metadata level, and no persistent identifier is available at data-level.
OzFlux does not currently assign DOIs to data it publishes. Resource implications are prohibitive and no demand currently exists.

TERN Long term Monitoring and Supersite Network

The site-level long-term monitoring and supersites data is published yearly based on predefined themes. The data may change after it is published due to various reasons including errors due to bias in the observation and amendments to the published data after further analysis. All the data is hosted from KNB Metacat. Currently, the data and corresponding metadata are updated as and when required and the repository supports version control i.e., all the changes made to the metadata and data are readily accessible. As a default, the latest version will be shown on the repository. However, changing the URL to the exact version can access any previous version.

Assign a DOI or other identifier

The minting of DOIs is supported in KNB Metacat. But, the DOI is linked to the latest version of the metadata record instead of an actual version. However, this can be changed to capture the dynamicity of data.

Integrated Marine Observation System

Summary

The Integrated Marine Observation System (IMOS) is designed to be a fully-integrated, national system, observing at ocean-basin and regional scales, and covering physical, chemical and biological variables. IMOS Facilities, operated by eight different institutions within the National Innovation System, are funded to deploy equipment and deliver data streams for use by the entire Australian marine and climate science community and its international collaborators. Datasets range in size from a few 1000 rows in a database to 20tb of satellite data.

The Australian Ocean Data Network (AODN) Portal provides access to all available Australian marine and climate science data and provides the primary access to IMOS data including access to the IMOS metadata. A large proportion of data is stored as netCDF on a THREDDS server. From there it is extracted to relational databases. Users may access data directly from the THREDDS server using OPeNDAP or via Amazon using programs such as S3Browser. Access via the AODN portal is based on a 3-stage process:

1. Select a Data Collection: search collections by parameter, organisation, platform and/or date and location.
2. Create a Subset of chosen data collections,
3. Download the selected data collection subsets.

The key driver for IMOS to enable citation of dynamic data in the AODN is:

- compliance with publishers' data availability policies

Form of citation statement

Instructions are provided.

For IMOS data, the following advice is provided in each metadata record:

The citation in a list of references is: "IMOS [year-of-data-download], [Title], [data-access-URL], accessed [date-of-access]."

Any users of IMOS data are required to clearly acknowledge the source of the material derived from IMOS in the format: "Data was sourced from the Integrated Marine Observing System (IMOS) - IMOS is a national collaborative research infrastructure, supported by the Australian Government." If relevant, also credit other organisations involved in collection of this particular datastream (as listed in 'credit' in the metadata record).

For non-IMOS data:

All data that is provided to the AODN is owned by the originator organisation. The AODN has no rights to the data and is only aiming to make others aware of the data and provide direct access to the underlying data. For this reason, the AODN prefers that data is provided under an open access licence – and recommend Creative Commons Attribution 4.0. The AODN specifies that data should be cited and an appropriate acknowledgment included in any publication as requested by the data originator (specified in the metadata record).

For example, metadata from IMAS has the following suggested citation alongside their licence:

Barrett, NS (2015), Macquarie Harbour WHA oxygen logger dataset. Institute for Marine and Antarctic Studies. Data accessed at http://metadata.imas.utas.edu.au:/geonetwork/srv/en/metadata.show?uuid=20b07936-3bfb-4a72-805d-0b24f1fd4d3f

Citation statement provided for the Australian Phytoplankton Database 2016 snapshot:

This is the bibliographic reference for the dataset and this metadata record that describes it:

Davies CH, Coughlan A, Hallegraeff G, Ajani P, Armbrecht L, Atkins N, Bonham P, Brett S, Brinkman R, Burford M, Clementson L, Coad P, Coman F, Davies D, Dela-cruz J, Devlin M, Edgar S, Eriksen R, Furnas M, Hassler C, Hill D, Holmes M, Ingleton T, Jameson I, Leterme SC, Lonborg C, McLaughlin J, McEnnulty F, McKinnon AD, Miller M, Murray S, Nayar S, Patten R, Pritchard T, Proctor R, Purcell-Meyerink D, Raes E, Rissik D, Ruszczyk J, Slotwinski A, Swadling K, Tattersall K, Thompson P, Thomson P, Tonks M, Trull TW, Uribe-Palomino J, Waite AM, Yauwenas R, Zammit A, Richardson AJ (2016), The Australian Phytoplankton Database (1844 - 2016) - abundance and biovolume. Australian Ocean Data Network - DOI: 10.4225/69/56454b2ba2f79 (http://dx.doi.org/10.4225/69/56454b2ba2f79)

Assign a DOI or other identifier

DOIs are not currently assigned to IMOS datasets published via the AODN, due to the complex nature of the bulk of it. Several requests for DOIs for AODN datasets (datasets that are hosted by the AODN, due to the DOI request) have already been made. The Australian Phytoplankton Database (APD) is the first use case, driven by requirements of Nature Scientific Data journal which requires a stable version of the data to be deposited in a public repository. As the APD is regularly updated, to satisfy publisher requirements, a snapshot of the Database was taken and published with a DOI. This meets the immediate requirement, but does not enable the most current version of the database to be cited with a DOI.

Access to a subset or time-slice of data that is cited (eg in a publication)

Currently, for either the APD, or any other IMOS data - queries are not stored and the AODN is unable to recreate a version of the respective database for any given time.

Version management

Data is quality checked before being sent to IMOS but may be corrected or reprocessed several times. When new versions are published, the previous version is archived. Therefore, it is possible that a mismatch may occur between data cited and data that is readily available.

CSIRO

Summary

The CSIRO Data Access Portal provides access to research data, software and other digital assets published by CSIRO across a range of disciplines. The DAP enables CSIRO researchers to publish data with a DOI and license, or share data internally with all CSIRO staff, or a selected group of colleagues. An approval workflow is invoked where data is marked for public release. The DAP supports a variety of general and domain specific search options. Users may choose to download specific files or small collections via a browser. Large collections may be accessed via webDav or SFTP.

The DAP is currently optimised for publication of stable datasets stored within the DAP infrastructure. While the DAP is domain agnostic, it offers advanced functionality for specific data types, including astronomy data. A recent foray into dynamic data was driven by the CASDA project described in brief below. It is acknowledged that the approach taken for the CASDA project will not readily transfer to other domains. While the focus here is on CASDA, use cases from the geoscience domain are well described by Klump et al (2016)

Overview of the CASDA use case

CSIRO Australian Square Kilometre Array Pathfinder Science Data Archive (CASDA) forms a core component in the Australian Square Kilometre Array Pathfinder (ASKAP) system. CASDA will essentially become the primary point for storing, managing, sharing and using processed ASKAP data products.

ASKAP will produce astronomy data at an unprecedented rate and CASDA will be home to a subset of this, pushing the envelope on the 'big data' paradigm with data ingest rates expected to reach ~15TB per day.

Form of citation statement

In the DAP, a citation statement is provided as part of the metadata record eg.

Crosbie, Russell; Broadhurst, Linda; Doerr, Veronica (2016): Murray NRM Water Synthesis data. v1. CSIRO. Data Collection. http://doi.org/10.4225/08/5726ED34D9CD1

Harvey-Smith, Lisa; Chapman, Jessica; Lenc, Emil; McConnell, David; Edwards, Philip; Phillips, Chris; Sault, Bob; Reynolds, John; Heywood, Ian; Serra, Paolo; Chippendale, Aaron; Popping, Attila; Allison, James Richard; Indermuehle, Balthasar; Bell, Martin; Bannister, Keith; Bunton, John David; Kimball, Amy; Procopio, Pietro; Marquarding, Malte (2015): ASKAP Data Products for Project AS031 (BETA Science Observations): catalogues. v1. CSIRO. Data Collection. doi

Assign a DOI or other identifier

DOIs are routinely (system) assigned to data collections publicly released via DAP. To date, CSIRO has assigned DOIs to ~1400 published datasets.

In the case of CASDA data, each data collection is created with a 'parent' record that is assigned a DOI. At a specified interval (say, 30 days), a 'snapshot' of the accreting collection is taken to create a new 'child' collection which is assigned a DOI. The DOI for children resolve to the parent so the user is always taken to the most recent and complete version of the collection. Data previously published does not change, so each new 'snapshot' incorporates additional data, but no changes to existing data are made.

Access to a subset or time-slice of data that is cited (eg in a publication)

The CASDA approach is the only model implemented in the DAP to date. However, when the DAP starts to reference CSIRO data held outside DAP storage infrastructure (eg NCI) the issue will need to be addressed.

Version management

In the DAP, a new version is created whenever a dataset or its metadata record are changed (apart from minor typos etc). A new landing page is created, and a DOI assigned, to each new version created. The version number is included in the citation statement.

National Computational Infrastructure

Summary

The National Computational Infrastructure (NCI) is Australia's national research computing facility. The NCI co-locates a vast, publicly funded databank with petascale computing facilities. The NCI has a focus in the environment, and in climate and earth system science in particular. Users can query the entire 10 PB NCI National Research Data Collection at once to retrieve data subsets in a compliant format for their research task. NCI currently hosts most datasets through THREDDS. They also support other data servers, including Hyrax, ERDDAP and Geoserver, as required. Data held on the NCI is dynamically changing and being accessed by queries that are dynamic.

Data can be accessed online through analysis tools or downloaded directly through the HTTP service. Alternatively users can subset the data for download in ASCII or NetCDF formats using OPeNDAP, or extract geospatial areas of data prior to download using the NetCDF Subset Service.

NCI supports Open Geospatial Consortium (OGC) standards including Web Map Service (WMS) and Web Coverage Service (WCS).

The key driver for NCI to enable citation of data held at the NCI is:

- compliance with publishers' data availability policies, however preserving a copy of data extracted from the NCI is generally not feasible due to data scale.

Assign a DOI or other identifier

The NCI are not routinely minting DOIs, but have the capability to do so if requested by data providers. As of May 2016, the NCI had minted 9 production DOIs.

Access to a subset or time-slice of data that is cited (eg in a publication)

This is a key problem space for the NCI. With large volume data arrays that are over a petabyte in volume, storing multiple time stamped snap-shots is not feasible primarily due to cost of the infrastructure to store them.

Therefore, the NCI is moving forward with a provenance (aka recipe) approach to making dynamic data citable. Where publishers require data availability and cited data cannot practically be deposited in a repository, a provenance record would be submitted instead.

In the approach being taken by the NCI, the provenance record is automatically captured as part of the research workflow and assigned a PID. The NCI has developed a standardised provenance model to capture agent activities and processing workflows. This concept has been tested within NCI, and is soon to be tested by Edward King (CSIRO). This testing will also look at the cost of implementing the provenance model.

The model being implemented by the NCI:

- requires source datasets to go through a controlled release process, similar to software and the exact changes to the data set are documented, so that if required a data set can be recreated at a particular time
- utilises the provenance data model – PROV-DM
- requires a Linked Data platform

To be successful in practice, the NCI approach requires:

- new scholarly publication procedures accept the persistent identifier of the provenance workflow (recipe) that created the data extract
- the provenance workflow to link to a series of persistent identifiers that:
    - at a minimum, provide complete dataset production transparency
    - if required, would facilitate reconstruction of the dataset (cf. reproducibility)
- Requires strict adherence to patterns for provenance representation (PROV-DM)
- Requires data repositories to develop provenance workflows on all datasets that they expose, and any changes to those datasets throughout their history

Version management

This is an issue the NCI are keen to address. While guidelines such as the RD-A *Recommendations of the Working Group on Data Citation* discuss data versioning, there is currently no agreed definition of, or approach to, data versioning. Well documented versions are key to the provenance model NCI are implementing, however, each organisation with data hosted on the NCI has a different approach to versioning.

Geoscience Australia

Summary

Geoscience Australia (GA) is Australia's pre-eminent public sector geoscience organization being the nation's trusted advisor on the geology and geography of Australia. A key strategic priority for GA is:

Maintain an enduring and accessible knowledge base and capability to enable evidence-based policy and decision-making by government, industry and the community. Maximise the value of public sector information by creating opportunities for innovative use and reuse of data.

A key role for GA is to manage and provide access to Australia's geoscience data. GA manages a substantial catalogue of data, publications, maps, online tools and web services. Currently, the primary access point is the Data & Publications portal on the GA website.

While much data may be directly downloaded from the portal, some data is only available via mediated access (contact name, or a fee may apply). Geoscience Australia also provides web services for public use that allow access to selected data without having to store datasets locally. Geoscience Australia supports a variety of web service protocols, including Open Geospatial Consortium (OGC) services and ESRI mapping and image services.

Perhaps the largest and most complex example of dynamic data managed by GA is the EarthCube. Currently, there is no mechanism to ensure the specific 'cut' of EarthCube data used by a researcher is citable.

While there is currently no immediate business driver for enabling dynamic data citation as such, there is a drive towards enabling the replication of data. This could be regarded as a component of enabling citation of data. Implementation of the new eCat data product catalogue may be a catalyst for enabling dynamic data citation - to allow for more granular citation of datasets underlying a data product.

Form of citation statement

A form of citation is not suggested in the ISO-19115 compliant metadata records that describe GA data. When the metadata is transformed to RIF-CS for display in Research Data Australia, a DataCite compliant citation statement is created.

Assign a DOI or other identifier

GA routinely assigns DOIs to, and commits to retention of, datasets referenced in reports. While this is currently feasible, GA foresees when the scale of data will make this prohibitive. As of June 2016, GA had minted >4500 production DOIs

Access to a subset or time-slice of data that is cited (eg in a publication)

Like the NCI, GA is moving forward with a provenance approach to enabling reproducibility of data. There is management support (and funding) for two provenance projects in 2016/17 intended to enable the identification (hence citation and reuse) of data via its provenance.

The 'provenance' data identification model has been developed to enable the identification of a dataset that was constructed and used, but not stored. In such a case, the recipe is intended to enable a rerun of the 'recipe' to construct a dataset identical to that previously constructed.

An example is the persistent storage and identification of a web service along with its query parameters, both identified in relation to an additional provenance chain that identifies the (not-stored) source data and how it came to be. At its finest, such information would allow the actual reconstruction of an identical dataset; less fine, it allows the understanding of what processes, agents and data inputs were used toward its creation. The 'recipe' would take the form of a provenance 'chain', documented according to standards.

This provenance approach being taken by GA is very similar to that being taken by the NCI.

Version management

Like the NCI, version management is a key issue for GA. An issue in common for TERN Eco-informatics, NCI and GA is knowing when and why the data provided to them changes, and hence, when a new version should be created. Current versioning practice for GA is to retain the rawest form of data (Level 1). However, only the version previous to the current version of processed data is retained.

References

Wyborn, L., Car, N., Evans, B., Klump, J. (2016) How do you assign persistent identifiers to extracts from large, complex, dynamic data sets that underpin scholarly publications? Geophysical Research Abstracts, Vol. 18, EGU2016-11639-1, 2016

Klump, J., Huber, R., Diependbroek, M. (2016) DOI for geoscience data – how early practices shape present perceptions. Earth Science Informatics, vol.9, pp.123-136. DOI 10.1007/s12145-015-0231-5