Dynamic Data Citation [1]

Submitted by ramdeen on Mon, 2014-09-15 12:10  
Thursday, January 8, 2015 - 13:30 to 15:00  
Thursday, January 8, 2015 - 15:30 to 17:00  

Session Type: Breakout [3]  
Expertise Level: Beginner [4]  
Collaboration Area: Data Preservation [5]  
Preservation and Stewardship [6]

Abstract/Agenda:  
The primary goal of this workshop will be to explore the feasibility of implementing a data citation model developed by the Research Data Alliance Dynamic Citation Working Group against a number of use cases contributed by members of the ESIP community. These cases focus on dynamic data. The workshop will be run by Andreas Rauber of Vienna University of Technology. Data sets will be provided by ESIP members such as NCAR, Biological and Chemical Oceanography Data Management Office (BCO-DMO), and NSIDC. In addition, a use case about Vector Borne Disease Network EMOD simulation data is being provided by the Hesburgh Libraries of the University of Notre Dame.

In this two session workshop, an overview of the data citation model will be given; followed by brief descriptions of the use case data sets to be examined during the workshop. The remainder of the workshop will focus on assessing the feasibility of applying the data citation model to these datasets, and exploring any issues with the proposed model. The workshop will close with a discussion to determine how to move forward with these steps in combination with the current ESIP Data Citation Guidelines.

A variation of this workshop was previously held in the UK. The attendees included representatives from 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 of that workshop 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, and possible steps forward were proposed for a few of the use cases presented. The report from this workshop can be found on the RDA DCWG [7] website.

Simplified summary of the RDA DCWG model goals.

***Note the existing ESIP guidelines cover citing subsets but only in a human readable way. This model looks at citing through machine understanding.***

1) Citing or making subsets of data citable. For example, if a researcher applies a set of filters, identifying some data which they use in a study. Your organization wants them to cite that specific subset. How do you give it a persistent identifier without having to store a duplicate copy of the
data?

2) When you have continuously created data (dynamic or static), or data which is updated, how do you ensure that the citation resolves to the previous version of the data?

Notes:
Overview: We started to have data citation recommendations; however, there are different “modes” of data. For example, static data would have different characteristics versus dynamic data. In addition, the citation should be machine actionable.

Purpose: We would like to discuss the proposed solution by RDA, introduce the selected datasets, and review how well the proposed solution could be applied to the selected dataset.

Speaker 1: Andreas Rauber
Title: RDA WG Data Citation

Notes:

- 2 areas of focus:
  - Citing arbitrary subsets of data
  - Citing data that is dynamic
- Not in focus:
- Metadata for citing data, landing page design
- Which PID solution to adopt
- Citable datasets so far have mostly been static; however, data is dynamic.
- Also, what about granularity of data to be cited? In other words, how to cite the subset of data? Also, even if the dataset is static, we would like to be able to cite an arbitrary subsets of the data.
- The solution proposed needs to be stable, machine/human-actionable, and scalable.
- This is the part of the evaluation that we would like to go through today with the proposed solution.
- A question from the audience: Are we asking too much at the onset of data citation?
- Advantage is that the user will get an option to retrieve one of the following: original data, current version, or changes → this allows the tracking of provenance as well.
- Some issues that might be discussed later: timestamp assignment, importance of unique sort/query rewrite complexity; hash-key computation, subset-to-full data relationship; technological migration, and distributed data sources.
- A comment from the audience: the purpose we cite data might create different mechanisms to cite data.
- Dynamic Data Citation and Deployment: subset of data will be identified; once selected, PID will resolve to a landing page where the PID is associated with a data citation.
- Key items to discuss today: Q/A on RDA recommendation; needs in pilots present; feasibility of approach for pilots, and effort for/impact of/interest in moving forward.

Questions/Comments:
1. Solr has been trying to implement similar mechanism. However, even though the history of query can be recorded, the older data might not always be kept due to storage limitation.
   1. Ruth - getting at the biggest issue. Where is the line what you put in methods versus in citation?
      1. Mark Parsons - It depends on the use case. There may be needed citations that are not in a paper.
      2. Andreas - Citation is applicable beyond “paper” type.
      1. Mark - Need to separate credit concern from provenance concern. See poster from Parsons and Fox: [http://www.slideshare.net/MarkParsons/parsons-citation-agu2014](http://www.slideshare.net/MarkParsons/parsons-citation-agu2014)
   2. Ruth suggested that we use Joe Hourcle’s example as a use case.
   3. Ruth provided a simple diagram example to the discussion of the comment: "dynamic" citation is the return of the data set, the query, and the timestamp.
   4. Greg - One use of citation is tracking uses and impact. If refs are being seen "hits" will this get the hits on the data sets as well as the query?

   Answer: yes, that relationship needs to be maintained.

5. Generally, different PID system might have different strengths or weaknesses for different purposes.
6. People receive data from repositories might not have a consistent way to refer to these data.
7. All the comments and questions demonstrated that it is important and helpful to walk through the proposed method with the use cases to explore all these issues further.

Speaker 2: Ruth Duerr (NSIDC)

Title: MODIS/Terra Snow Cover 5-Min L2 Swath 500m, Version 5 (MOD10_L2)

Notes:

- This is the first of the proposed use cases.
- MOD10_L2 Characteristics:
  - Key: granule required every 5 min, every data file has an accompanied metadata.
  - Access methods: currently, data pool (FTP) and Reverb (central metadata) - subsetting is available.
  - The options of different access methods → how does it affect DOI?
  - DOI will be available per version of the dataset, in general.
  - In addition to insert times, there are also production times.

Speaker 3: Shannon Rauch

Title: “Dynamic” Data at BCO DMO

Notes:

- These are the second and third of the proposed use cases.
- Time series from cruises:
  - Cruises are monthly, and data are received nearly yearly.
  - Only 1 data version is available at a time, but it represents an overall dataset with new data appended. Changes are also updated.
- Aerosols:
  - New data files will be submitted → the new data files will have additional information, and the
newest data file is made available publicly.
  - The older data files are archived, but not readily available to the public.
- In general, for both use cases, only newest data file is available online; metadata changes are time-stamped in Drupal, but row-level changes are not time stamped.
- Datasets are small enough in size that users usually download the entire dataset → no subsetting by BCO DMO is required.
- When DOIs are used, what happens to the DOI when new files are made available?
- In some cases, new DOIs are assigned, but not sure.
  - Feedback from the audience - it would be good to retain the original DOI.
- Adding rows and updates, such as new parameters, might be treated as 2 different scenarios.
- Comment from Anne: Regarding data versions, my criteria for distinguishing a new dataset and issuing a new DOI is whether the algorithm has changed. If so, it's a new dataset and should have its own DOI.
- In general, a DOI should be assigned to an overall intellectual object; if any additional DOIs are assigned, the links should be established between these other DOIs to the overall intellectual object's DOI.
- This is one of the reasons that be “dynamic” or “subset” might not be 2 different concepts.
- Comparison to journal’s DOIs → an analogy to help us consider how to assign DOIs to datasets and its subset?
- Mark - we should step away from discussing DOI and focus on the identifying/locating the subsets.

Speaker 4: Anne Wilson
Title: LASP Interactive Solar Irradiance Datacenter
Notes:

- This is the fourth example of the use cases (link to the dataset: [http://lasp.colorado.edu/lisird/tss.html](http://lasp.colorado.edu/lisird/tss.html) [9])
- LISIRD uses LaTiS (LASP Time Series Server): data access middleware and an API that allows dynamic selection of dataset, parameters within a dataset, subsetting on time and spectral range, running filters, and output format. The API is publicly available, anyone can use it.
- Mark’s comment - LATIS seems a prime candidate for adopting the dynamic citation model (Andreas agrees).

Speaker 5: Joe Hourcle
Title: FRBR Aggregates as Collections
Notes:

- Joe volunteered his use case during the session.
- Mainly, the collection generates images of the sun.
  - The primary image does not have colors; colors are added afterward.
  - The images are also available in various sizes.
  - Further, the images can be viewed at different positions because a full image of a sun is composed of 14 days of data.
- Question asked in the presentation: What image is “next”? - Fully processed is seconds later, but “section” of data is minutes later. Also, there are 4 different telescopes that are producing the data for the overall image.
- Typically, people use dates as their primary filter for data selection.
- Data are organized more in a “documented oriented” database.
• The data are also publicly accessible, so non-scientists could either learn the scientists’ query syntax or obtain an image using general characteristics.
• People can also search using catalogs and request data after identifying the desired data.

Continuation from the Session at 1:30pm.

Speaker 6: Natalie Meyers (VecNet)

Notes:

• Use case is based on software simulators.
• Rate of change is irregular because it depends on the time frame that people are doing simulations.
• Some of the data are file based, but there are other forms possible.
• Citing simulations and runs: metadata has been implemented on the run level.
• There is not a clear distinction between “run” and “simulation”.
  ○ This affects how the output files are generated, both in terms of format and timestamps.
• Is simulation analogous to a sum of several runs?
  • Essentially, yes.
• The input files could also be geospatial.
• Preferably, the input files, the metadata, and the simulations results would all be bundled together, so that it would be possible for someone to repeat the simulation and reproduce the same results using their own systems.
• It is desirable for the IDs to be assigned to the input data files as well, so that they are also machine-actionable.
• The input data files are essentially “recreated” from the “original” version. As a result, the provenance is another issue that VecNet will need to consider as well.
• This use case is proposed here because the input for the simulations are mainly from Earth sciences (climate data, such as weather forecast).
• VecNet would like to receive feedback on how the processes and results can be reproducible.
• The input data is obtained from some other sources → these files are stored in VecNet’s digital library for direct access or for allowing users to determine the original sources.
  ○ The preprocessing techniques and information would be of interest to others who would like to reproduce the results.

3 groups will be formed, and the dataset selected are:

• LaTIS
• MODIS
• BCO DMO

Other comments from Anne:

• This paper explores some of the ways in which scientific data is an unruly and poorly bounded object, and goes on to propose that in order for datasets to fulfill the roles expected for them, the following identity functions are essential for scholarly publications: (i) the dataset is constructed as a semantically and logically concrete object, (ii) the identity of the
dataset is embedded, inherent and/or inseparable, (iii) the identity embodies a framework of authorship, rights and limitations, and (iv) the identity translates into an actionable mechanism for retrieval or reference."

Attachments/Presentations:

- BCO-DMO_dynamic_data.pptx [12]
- FRBR-aggregates-code4lib.ppt [13]
- MOD10_L2.pdf [14]
- NKMData Citation Recommendations for VecNet Transmission Simulator.pptx [15]

Session Leads:

Name: Ruth Duerr [16]
Organization(s): The Ronin Institute [17]

Name: Sarah Ramdeen [18]
Organization(s): School of Information and Library Science UNC-CH [19]
Email: ramdeen@email.unc.edu [20]

Name: Andreas Rauber [21]
Organization(s): Vienna University of Technology [22]

Participants:

On webex: Anne Wilson, Bruce V, Heather Brown, Mark Parsons, Steve Aulenbach, Ge Peng, and Philip Jones.

Creative Common License: Creative Commons Attribution 3.0 License

Teaser: This workshop will explore the feasibility of implementing a data citation model developed by the Research Data Alliance for dynamic data.

Accepted:
Keywords: data citation [23]
dynamic data [24]

Source URL: http://commons.esipfed.org/node/2677

Links:
[18] http://commons.esipfed.org/node/557
[20] mailto:ramdeen@email.unc.edu
[22] http://commons.esipfed.org/taxonomy/term/1843