Data Stewardship throughout the Data Life Cycle [1]

Submitted by superadmin on Fri, 2012-11-30 18:42 Thursday, January 10, 2013 - 10:30 to 12:00
Event: Winter Meeting 2013 [2]
Session Type: Breakout [3]
Collaboration Area: Data Management Training [4]
Data Preservation [5]
Abstract/Agenda:

The Biological and Chemical Oceanography Data Management Office (BCO-DMO) was created in late 2006, by combining the formerly independent data management offices for the U.S. GLOBEC and U.S. JGOFS programs. BCO-DMO staff members work with investigators to publish data from research projects funded by the Biological and Chemical Oceanography Sections and the Office of Polar Programs Antarctic Organisms & Ecosystems Program (OPP ANT) at the U.S. National Science Foundation. Since 2006, researchers have been contributing data to the BCO-DMO data system, and it has developed into a rich repository of data from ocean, coastal and Great Lakes research programs. Data management services are provided at no additional cost to investigators funded by those offices. The main goals of BCO-DMO are to ensure preservation of NSF funded project data and to provide open access to those data.

BCO-DMO has developed an end-to-end data stewardship process that includes all phases of the data life cycle: (1) working with investigators at the proposal stage to write their two-page NSF data management plan; (2) registering their funded project at BCO-DMO; (3) adding data and supporting documentation to the BCO-DMO data repository; (4) providing geospatial and text-based data access systems that support data discovery, access, display, assessment, integration, and export of data resources; (5) exploring mechanisms for exchange of data with complementary repositories; (6) publication of data sets to provide publishers of the peer-reviewed literature with citable references (Digital Object Identifiers) and to encourage proper citation and attribution of data sets in the future and (7) submission of final data sets for preservation in the appropriate long-term data archive.

Recent efforts by BCO-DMO staff members have focused on identifying globally unique, persistent identifiers to unambiguously identify resources of interest that are curated by and available from BCO-DMO. The process involves several essential components: (1) identifying a trusted authoritative source of complementary content and the appropriate contact; (2) determining the globally unique, persistent identifier for resources of interest and (3) negotiating the requisite syntactic and semantic exchange systems. Recent enhancements will be highlighted including: (1) use of controlled vocabularies and ontologies; (2) Linked Open Data; and (3) use of Digital Object Identifiers (DOIs). In addition, we are investigating ways to use persistent resource identifiers to improve the accuracy of the BCO-DMO data collection and to facilitate exchange of information with complementary ocean data repositories.

This session will begin with a brief introductory presentation to clarify the eight phases of the research data life cycle followed by a moderated group discussion. The purpose of the discussion is to identify commonalities and differences in the way different communities meet the challenges of data stewardship throughout the full data life cycle and also determine any gaps that currently exist.

Notes:
participants will be invited to complete a matrix that represents the way they currently address requirements in each phase of the Data Life Cycle

see Google doc:
https://docs.google.com/spreadsheet/ccc?key=0AswVs_NyGD3ddC1rY1JwWlNjbm5... [6]

document title is: datalifecyclematrix

NEW LINK:

https://docs.google.com/spreadsheet/ccc?key=0AtskHfF4KIE8dEptaWpQRC1WSlA... [7]

ESIP Student Fellow Comment: Note will be taken in the following google document during the meeting. Feel free to edit as needed during the meeting: https://docs.google.com/document/d/1flX3fQqjGPEXZSnrQPQXLtQD6qah7FOL4J9G... [8]

Actions:

Data Stewardship throughout the Data Life Cycle

Participants:
Ken Casey, NOAA NODC (remotely)
Steve Aulenbach (remotely)
Don Collins, NOAA NODC (remotely)

Meeting notes:

Introductions by Cyndy

Presentation using slides (to be posted later).

Overview of her work - how many of these facets relate in what she is calling a Data Life Cycle. There is a change in expectations in your data - it is not just for your own research, but there is a demand for data to be released to the community with metadata. It is being driven at the moment by program managers but it is also very real value for science researchers who need access to a broader range of data and need information about this data due to an unfamiliarity with these different fields.

Cydny mostly works for NSF but also works with large legacy projects that are interagency and have been carried forward.

Data consumers, data managers etc have different perspectives, but there are a lot of similarities of what we think about as data from the beginning to the final step of preservation. What documentation will you need to begin acquisition etc. What she hopes to get out of this session is more input on this model she has demonstrated.

Your role could change in the life cycle - from creator to consumer. It is important to understand the stewardship approach at each of these stages. AS individuals or as partners and collaborators along each of these phases.

It is important to know your role in this cycle and what has been done before your part. To talk to each other around this process and at each phase. The last step being archiving these materials.

Cydny would like us to think about these issues, will finish her presentation and then will open this up to discussion.

The way she currently thinks about this (in BCODMO) Their data scientists do not go out on the ships anymore. They have a project called Rolling Deck to Repository. NSF recognized that the data from the ships were not making it to the archives. If that occurs, the rest of the cycle falls apart. They found funding to work on this and it has been very succesful. It was seen to be a very effective idea and has been copied. They have outsourced a GIS facet to provide access. They are not an archive but they do send it out to a separate group to work with. They are recognizing the importance of this need/function.

There are a lot of ESIP collaboration groups who are focusing on these areas. Slides list a number of communities that are focused on these issues.

Proposal - would like to hear differences in opinions.

IN each phase there are different roles that take part in the activity. It changes depending on the role, but she makes the researchers aware of the expectations, benefits and changes. Thinking about data management from the start of the project. NSF requires a two page data management plan.

During the data acquisition phase there are such activities as event log sampling. The data managers encourage the producers to use controlled terms.

In the analysis and synthesis phase - Cyndy’s group’s role is not as large, but at this point it falls to the data producers and the archivists etc to create backups and move data off of researcher’s laptops.

Putting the data into a repository - not every domain has someone to help shepherd the data to the data repository. identify for your community though what the repository is, this is an important step to know where the data is headed at the start of the project.

Data discovery and access - all about metadata!

Use and Re-use of the data. This has become very important and is the main science driver. It has increased the importance of the efforts of how important the data can be in ways we have not thought of yet. It is very valuable for things like developing a time series.

Publication phase - the ability to cite a data set. The recent changes that allow the author of a data set to receive citation credit for citation metrics and advancement. This is important for the community to understand as incentive. Another step is publishing the data and assigning a DOI and citing these in the publications on the science.

Preservation - 100 year OIAS archive. To support discovery and re use by colleagues.

During the next step - Cydny would like to do an exercise. She has put up a google doc with a spreadsheet of this data life cycle. She would like to see how this is represented at other organizations and what roles might be included that she has not thought about.

Mike pointed out Quality control has not been addressed in this model. Cydny said it could go in to a few different sections but this is really important for us to share with her.

Here is a link to the shared document
https://docs.google.com/spreadsheet/ccc?key=0AswVs_NyGD3ddC1rY1JwWlNjbm5QVEY3Tk5jQnBPWEE#gid=0 [6] STATIC DO NOT USE

Use this version:
https://docs.google.com/spreadsheet/ccc?key=0AtskHfF4KIE8dEptaWpQRC1WSlAwcjc3Mk1jVGMza3c
[7]

Instructions are in Cydny’s slides - fill this out from your own personal perspective.
The columns are the different phases and the rows can be used to identify partnered organizations who address these needs. And include some of the tasks they do for these phases.

For example Cydny partners with the library for DOIs so she submits it to the library and they put up the landing page and assign a DOI.

John asked if it could be a shared document.
https://docs.google.com/spreadsheet/ccc?key=0AtskHfF4KIE8dEptaWpQRC1WSlAwcjc3Mk1jVGMza3c
[7]

Nancy suggested a different life cycle model from DataOne http://www.dataone.org/best-practices
[9]

- which focus on research data. As opposed to the DCC model which does not http://www.dcc.ac.uk/resources/curation-lifecycle-model [10]

Cydny asked people to share others if they are aware of them.

Ken asked a question - can Cydny explain what she wants in each column?
Cydny posted a slide with more instructions, but suggested that we all enter a row or more and enter your roles, activities and partners in each stage.

From a paper by Parsons and Fox (2012)
https://dl.dropbox.com/u/546900/parsons_fox_metaphor_dsj_revised_submitt... [11]
" As we examine different worldviews, we need a fuller development and understanding of all the roles in the entire data stewardship enterprise. Lawrence et al. (2011) lay out a series of defined roles from a Data Publication perspective. Baker and Bowker (2007) do the same from an ecological infrastructuring perspective. Baker and Yarmey (2009) further examine the specific roles of data curation. Schopf (2012) has yet another examination from a production software perspective. They all emphasize different roles with different terms, and even seem to define the term “role” differently. A deeper comparison of these roles and how data managers and all the players in the enterprise perceive them is warranted. Are the different actors using the same frames and metaphors and in the same way? Is there a difference across disciplinary cultures? How do the worldviews and metaphors of data creators and data users align? Do the metaphors and frames of data scientists help or hinder that alignment? A particularly critical set of roles falls in the category of what Baker and Bowker (2007) call “in between” work. These roles of the intermediary and “middleware” connecting computer science and domain science are central to informatics and data science (Fox, 2011), yet they are also often hidden from view. Similarly, the role of a curator is critical, but as Fleischer and Jannaschk (2011) illustrate, curation can also introduce a bottleneck in data archiving and release processes. They suggest a closer examination of the role of the data manager or curator and automated curation services. In such an examination, we must consider not justthe science domain but also the culture from which curators emerge. The culture of an academic library or archive is vastly different from that found in operational weather center, for example. Finally, in the examination of roles, we should use different worldviews to tease out what important roles we have missed. For example, the roles of the financial sponsor or the unintended non-specialist in the overall data ecosystem have not beenexamined in depth. "

Hello everyone on the webex - the system crashed and we are restarting. I will let you know when we are back.

Gap in note taking due to computer problems.

Currently Brent and Cydny are discussion data preservation. Brent provided an example of graduate students posting data online directly and the life cycle makes sense for big data but not for ad hoc or smaller organizations.

Cydny agrees and thinks that the data publication requirements. Working with the repositories but also an educational point of teaching everyone about metadata and getting them to agree that these things are important and that includes methodology.

Mark - it is important that you do not think of this just as cycle but a collection of cycles. You might do some things in eddies along the way.

There is currently a discussion on how this is a specific view on the data lifecycle and it is also simplified. She thought about creating transparent overlays to show the other steps. It is fascinating when you start making links to other groups. You do not have to go around teh diagram clockwise.

John - a note or instructions that any phase could work or support from any other phase might be useful to add to this.

EVERYONE WHO WAS ON THE WEBEX - can you rejoin the meeting? It looks like it is listed as the afternoon session instead of the morning.

Discussion still on going about providing access to research by Brent.

John S. asked if she has a dedicated person who keeps track or creates metadata for each of these stages?

Cydny - we have 4 full time people who are activity doing this for the NSF community. Cydny also mentioned that outreach and communication are important parts of her role. To raise awareness. The big role is to gather the metadata and recast that in an ISO or other standard compliant way.

Stephan said it would be interesting to find out what metadata people use and what stage of the process it is collected at. He would be willing to bet there are large differences at each phase.

Cydny mentioned it being important to also capture metadata during these phases as well.

Contributor - the point where a scientists exports the materials. This is different than the acquisition stage which is when it is shaped in to the data set. It is later given to an organization where others can get access to it and that is where the formal metadata process occurs.

John had two comments - in GHRSST (http://www.ghrsst.org [12]) metadata standards are defined ahead of time and they are distributed. [Note: GHRSST also defines the data format standard (CF-compliant netCDF) and the data content standard (what must go in the netCDF files), known as the GHRSST Data Specification.] Also he was talking about ways to provide incentives for the researchers to consider more of this work. For example the metadata work is done at the contribution phase, but the publishability of the data set - that feeds back into the citation credit process that the research community understands immediately.

Cydny agreed with this point and has seen it take off well. She has also been using things to protect the researchers as well. It is a time stamp formal publication of when the data was deposited and assigned a DOI so if they were misquoted - contact information and such is available and there is a way to protect the PI.

John, importantly they have also documented this data as well. With methods.

Mark - where does rights negotiation fit into the lifecycle, including publication. This looks like it assumes open access to the data. He has three different projects that he would like to create something similar for, but each one is so different it does not fit well into this model. He also has a big hangup with the term data publication. Not all investigators are fans of data citations - feel like it is pulling away from citations to publications.

Cydny agrees, and tells people to cite both.

Laura asked about the DOIS

Cydny mentioned the Woods Hole library who helped create DOIs for the different datasets. They had a workshop 5-6 years ago that helped developed this process. They developed a process around the publishers so that users would not be charged for access. They started this 3 years ago.

Laura asked who maintains the data vs the landing page.

Cydny said that they still need to edit the data as it changes. But need to provide backbone data for a published manuscript. Here is a paper and here is the data it is based on. Leaves it to the user to know that there might be more data beyond what was just published. They extract the data from the database as a CSV file, and the essential metadata records - 20 Dublin Core fields. And generate a PDF file that goes in as a supplemental document. It gets submitted as a METS run through a checksum. It is sent off and unpacked and processed by the library. They put up the landing page and they maintain it. It also has a link back to the office who created it and you can track back to see if there is an updated version of this. They have recently gotten CrossRef to help them create versioning on the DOIs. They were worried about the publishers having copyright on the data and that was a step in the wrong direction, particularly if the data has not yet been published.

This was an international project that included developing countries. It seems to be working, Dublin Core is something that libraries are used to.

Denise - we deal a lot with Oil and Gas data in my state agency which is proprietary. We are required by law to keep the physical samples but we are not provided funding to manage the information about the materials - so this lifecycle does not really fit. It can be forced, but it is a very different thing. They also do not have control over some of the processes or documentation either at some point.

Cydny said I don’t have control but I do feel responsibility.

Denise - we can not publish the data - because it is proprietary.

Sarah mentioned dark archives - and that some of these things are being done while still maintaining restrictions to access to these materials.

Denise asked about thinking backward - what about existing data?

Cydny - it is hard to get funding for that and attention to it. But there is a shift from the public and NSF to the importance to access to these materials. So there is no funding now, but it could change. The rolling deck project had been proposed twice before before it finally was approved. It took NSF hearing that it was not a full effective use of research dollars!

Denise asked about private research groups? How do we get their buy-in and that other data would be useful in the future. Private buy in to preserve legacy data.

Mike - Climate community this is becoming very important. Digitization and other project process are being done and there are funding opportunities too.

Mark - funding does not just come in at the proposal phase. You might get a grant to help sustain your archive. These steps prompt a need for funding.

Cydny - the point that you speak up and voice your concern about the need for funding. And it takes time but she was able to get funding for her projects through NSF and it will continue to make big changes. Embedded in the research community through organizations like this.

Sarah will post a static copy of the spreadsheet to the commons but will keep this open and hopes people include more notes as time goes along.

Cydny would like this to become a white paper. Also this might be an interesting student research project. She will synthesize the results and find some way to share that back with the group. Asked about interest from the group?

Mark - discussion on the term curation and how complex that is.

Attachments/Presentations: datalifecyclematrix_ESIP_1_10_2013.pdf [13]

Session Leads:

Name: Cynthia Chandler [14]
Organization(s): Woods Hole Oceanographic Institution [15]
Email: cchandler@whoi.edu [16]

Notes takers:

Name: Sarah Ramdeen [17]
Organization(s): School of Information and Library Science UNC-CH [18]

Email: ramdeen@email.unc.edu [19]

Teaser: Introductory presentation to clarify the eight phases of the research data life cycle followed by a moderated group discussion #ESIPFed

Accepted:
Keywords: data management [20]
data preservation [21]
BCO-DMO [22]

Source URL: https://commons.esipfed.org/node/760

Links
[1] https://commons.esipfed.org/node/760
[2] https://commons.esipfed.org/taxonomy/term/464
[3] https://commons.esipfed.org/session-type/breakout
[4] https://commons.esipfed.org/collaboration-area/data-management-training
[5] https://commons.esipfed.org/collaboration-area/data-preservation
[6] https://docs.google.com/spreadsheet/ccc?key=0AswVs_NyGD3ddC1rY1JwWlNjbm5QVEY3Tk5jQnBPWEE#gid=0
[7] https://docs.google.com/spreadsheet/ccc?key=0AtskHfF4KIE8dEptaWpQRC1WSlAwcjc3Mk1jVGMza3c
[8] https://docs.google.com/document/d/1flX3fQqjGPEXZSnrQPQXLtQD6qah7FOL4J9G5_BO2K0/edit
[9] http://www.dataone.org/best-practices
[10] http://www.dcc.ac.uk/resources/curation-lifecycle-model
[11] https://dl.dropbox.com/u/546900/parsons_fox_metaphor_dsj_revised_submitted.pdf
[12] http://www.ghrsst.org/
[13] https://commons.esipfed.org/sites/default/files/datalifecyclematrix_ESIP_1_10_2013.pdf
[14] https://commons.esipfed.org/node/741
[15] https://commons.esipfed.org/taxonomy/term/391
[16] mailto:cchandler@whoi.edu
[17] https://commons.esipfed.org/node/557
[18] https://commons.esipfed.org/taxonomy/term/373
[19] mailto:ramdeen@email.unc.edu
[20] https://commons.esipfed.org/taxonomy/term/379
[21] https://commons.esipfed.org/taxonomy/term/795
[22] https://commons.esipfed.org/taxonomy/term/1202