Describing the repository landscape for data curators [1]

Submitted by annieburgess on Mon, 2016-04-04 10:38  Wednesday, July 20, 2016 - 14:00

Event: 2016 ESIP Summer Meeting [2]
Session Type: Breakout [3]
Expertise Level: Beginner [4]
Collaboration Area: Sustainable Data Management [5]

Abstract/Agenda:
We have many repositories in scientific domains for natural and social science research data, and an increasing expectation that primary research data will be deposited there. Data centers and repositories offer a variety services to researchers for this purpose, and a growing community of data managers and curators act as liaisons between primary researchers and repositories. To understand how to work with repositories, data managers need to know their basic features. Several groups have embarked on a discussion of the landscape of repositories and their services, eg, the Research Data Alliance (RDA), the Council of Data Facilities (CDF), and at a recent workshop focused on planning collaborative efforts among repositories in Tempe, AZ, that culminated in an ESIP cluster (Sustainable Data Management). This session will continue the discussion. We will become familiar with the existing and planned material describing repositories (e.g., from RDA, CDF, re3data.org [6]), and assemble questions asked by curators and researchers when deciding which repositories to contribute to, and how to work with them.

Notes:

- Presentations associated with the session are attached.
- This session is based on the work that is being done by the ESIP Sustainability Data Management Cluster.
  - Wiki page for this session (additional information and notes will also be made available at this page, including summaries of Recommendations for Registries, Recommendations for Repositories, Recommendations for Liaisons, and Next Steps for Landscape Analysis): http://wiki.esipfed.org/index.php/Sustainable_Data_Management/20160720_E... [8]
- Understanding the landscape of the current repositories helps inform the decisions for future repositories, including the strategies for return on investment (ROI).
- Presentation #1: "Coalition on Publishing Data in the Earth and Space Science (COPDESS)" by Kerstin Lehnert of LDEO/Columbia University
  - Foster consensus and consistency among publishers, editors, funders, and data repositories.
  - COPDESS has generated a Statement of Commitment signed by >40 publishers and data facilities and released guidelines for journal data policies.
  - Examples of specific commitments: online community directory, domain standards, education of researchers, Joint Declaration of Data Citation Principles, data location and availability, workflows and use of persistent identifiers.
  - Link to COPDESS: https://copdessdirectory.osf.io/ [9]
  - COPDESS used re3data.org metadata schema to describe data repositories.
  - Collaboration with re3data.org will enable further integration of existing efforts.
  - Examples of metadata fields that might describe a repository:
    - Type of Certification
    - Endorsement journals/publishers
    - Use of identifiers
  - Activities such as joint initiative with political/social science data repositories will also facilitate additional education of editors and authors and training trusted leaders in the domains for further advocacy and education.
  - COPDESS currently does not include information regarding the APIs a repository uses.
However, it is an information that Council of Data Facilities (EarthCube) might be able to help in providing.

- **Presentation #2:** "Interoperable Data Federation" by Matt Jones of NCEAS/University of California - Santa Barbara
  - There are many factors that caused barriers to interoperability of repositories:
    - Ex: momentum of built infrastructure, allure of shiny new technology, and institutionalized funding models.
  - DataONE is creating an interoperable federation.
    - Key principles:
      - API-centric: not only having an API, but also share the API.
      - Loose coupling: Specific components of the system can be reused more easily.
      - Diverse components: Allowing different components by be developed by different members of the community encourages collaboration.
      - Tiered deployment: Could be more time and resource efficient as not everything needs to be deployed and functional all at once.
    - Examples of DataONE repository's features are discussed, including DataONE repository members' profiles, and ORCID support.
    - Ultimately, DataONE aims to help build community.
    - Comments: Even though it might be difficult to manage human resources to allow and support the sustainability of data repositories, it is an area that requires continued effort.

- **Presentation #3:** POV: Data contributor by Margaret O'Brien of Long Term Ecological Research (LTER) Network Environmental Data Initiative (EDI)
  - Areas to consider from a scientist’s point of view:
    - Where is the best place for “my data”?
    - How do I get it there?
    - Can those who want it find it?
    - Who needs to find the data?
  - The "knowledge spectrum" (access vs. computing expertise) affect how scientists interact with data repositories.
    - Specific knowledge needs/support might include: technical specs, templates, definitions, web forms, communication, vetting systems, documentation, code libraries, etc.
    - These could also be the features that a repository could help in providing.
  - Comment: Even when someone has knowledge of a repository, it could still be daunting to navigate through the entire repository landscape. As a result, consolidating the repository could help with the ease and efficiency of working with the repositories.
  - Comment: There are also other elements of a repository, such as storage and use of identifier, that are important even though they have not been included in the "knowledge spectrum" discuss above.

- **Presentation #4:** Scientific/Research Data Ecosystem by Shelley Stall of AGU
  - Historical view versus the current view.
  - Current view includes the following roles:
    - Funder
    - Publisher
    - Researcher/Scientists
    - Repository/Long Term Archive
    - Data Manager/Data Steward
  - Shelley is also working on the different products and outcomes that are being generated by scientific research.
  - Suggestion: Librarians and commercial/business sectors should also be added to Shelley's ecosystem.
  - Suggestion: Study of motivation could also be very informative to understand further the elements that work well and those that cause tension point (including the overlaps of the activities by different roles).
Describing the repository landscape for data curators
Published on Commons (http://commons.esipfed.org)

Attachments/Presentations:

- OBrien_ESIP_2016.pptx [10]
- Jones_2016_ESIP_repo_landscape_v02.pdf [12]
- Lehnert_COPDESS_ESIPSummer2016.pptx [13]

Presenters:

- Name: Kerstin Lehnert [14]
  Organization(s): IEDA [15]
  Email: lehnert@ldeo.columbia.edu [16]

- Name: Matthew Jones [17]
  Organization(s): University of California at Santa Barbara [18]

- Name: Margaret O'Brien [19]
  Organization(s): LTER [20]

- Name: Shelley Stall [21]
  Organization(s): American Geophysical Union [22]
  Email: sstall@agu.org [23]

Notes takers:

- Name: Sophie Hou [24]
  Organization(s): UCAR/NCAR [25]

Participants:

In person (based on imhere.esipfed.org): Sophie Hou, Stevan Earl, Justine Woo, Chris Lenhardt, Shelley Stall, Kathleen Carr, Leslie Hsu, Nancy A Ritchey, Corinna Gries, Susan Haberer, edesfandiari, William Michener, Donna Scott - NSIDC, Karen Stocks, donhenshaw, Ward Fleri, Charles Schmitt, Viv Hutchison, Cyndy Chandler, Nate James, Karl Benedict, Rebecca Koskela, Andrea Thomer

Remote: Donald Collins, Paul Lemieux, Robert Partee, Joseph Koch, Jonathan Flaishans, Patrick Rae

Creative Common License: Creative Commons Attribution 3.0 License

Source URL: http://commons.esipfed.org/node/9139

Links:
[16] mailto:lehnert@ldeo.columbia.edu
Describing the repository landscape for data curators
Published on Commons (http://commons.esipfed.org)

[23] mailto:ssall@agu.org