Scholars Portal Dataverse Contacts Meeting

Meeting Notes, 2021-05-27

Attendees (please enter your name and institution): Alex Guindon (Concordia) Chris Hurst (Brandon University), Sarah Giesbrecht (UNBC), Jasmine Hoover (CBU), John Huck (UAlberta), Chris Shoniker (Carleton), Chantal Ripp (uOttawa), Liz Hill (Western), Will Roy (Queen's), Paul Lesack (UBC), Isaac Pratt (McMaster), Minglu Wang (York U), Yayo Umetsubo (Uof T Mississauga), Dana McFarland (VIU), Ève Paquette-Bigras (Université de Montréal), Jolaine Marcotte-Naud (Université de Montréal), Sandra Keys (Waterloo), Monique Grenier (UVic), Alex Cooper (Queen's), Alison Farrell (MUN), Doug Brigham (UBC), Louise Gillis (Dalhousie), Moira Davidson (Lakehead), Émilie Fortin (Université Laval), Jane Fry (Carleton), Irina Gherasim (Polytechnique Montréal), Kathy Szigeti (Waterloo), Eugene Barsky (UBC, guest), Daniel Bromby (Bishop's), Carrie Breton (U of G), Kaitlin Newson (SP), Alicia Cappello (SP)

Number of participants on call: 43

Agenda & Notes

1. Housekeeping
   - Notetaker: Alicia
   - Slides and agenda/notes on DV contacts resources page:
     https://spotdocs.scholarsportal.info/display/DAT/Institutional+Contacts+Resources
2. Scholars Portal Dataverse Service Updates
   - Slides:
     https://docs.google.com/presentation/d/1WrDBRkPcLn3MZ3bYPDeBgTsP5O-0r0GX-nhXv8eRh8s/edit?usp=sharing
   - Moving SP DV over to cloud storage, using the same infrastructure as ORLC. This means all data will be stored in triplicate across 3 of the 5 nodes in the cloud network.
     i. All the data centres are located in Canada: York, Toronto, Queen's, Ottawa, and Guelph.
   - Globus:
     i. Integration ongoing.
     ii. Working on resolving issues with restricted access as Globus uses a separate authentication system than SP DV.
   - SP DV is now discoverable through the Harvard DV!
     i. https://dataverse.harvard.edu/dataverse/scholars_portal_harvested
   - Harvard's Dataverse Community Meeting coming up June 15th to 17th.
     i. https://projects.iq.harvard.edu/dcm2021
     ii. Take a look at the Google Group for Harvard Dataverse to see details regarding requesting speakers/presenters.

- CTS Project Update:
  i. Portage announced the funding recipients and members of their larger cohort at the beginning of the month.
  ii. SP DV CTS Project members received a survey to obtain their opinions on events and activities to hold with the cohort. Email with results sent a few weeks ago.
  iii. SP DV CTS project cohort added 3 new members from 2 institutions this month.
  iv. Portage's CTS project organizing committee working on scheduling an introductory meeting in June - day and time TBD.
  v. SP DV team working on user guide, terms of use, FAQs, and Tech info sheet.

3. Community Topics
   - Dataverse North (John Huck, Chair, Dataverse North)
     i. Slides
     ii. Amber and John have been having discussions about bringing this community and the Dataverse North group closer together to work more collaboratively. Stay tuned for more info at the next meeting.
     iii. Dataverse North is currently a working group within the Portage Network of Experts.
     iv. Three sub-groups/committees: Metadata, Policy, and Training.
        - All members of the DVN working groups are members of at least one sub-groups.
     v. DVN Group formed in 2017. Report released in Year 1 recommending stakeholders should work together to establish a national Dataverse service.
     vi. DVN Group undergoing a review of the purpose and objectives of the group going forward.
        - Group is now designated as an Expert group, rather than a Working group.
        - Name (i.e., Dataverse North) may change.
   - Dryad2Dataverse (Eugene Barsky and Paul Lesack)
     i. Slides: https://spotdocs.scholarsportal.info/display/DAT/Institutional+Contacts+Resources?preview=/204079128/240583006/Dryad2dataverse_SP_27May2021.pdf
     ii. Source: https://github.com/ubc-library-rc/dryad2dataverse
     iii. Docs: https://ubc-library-rc.github.io/dryad2dataverse/
        - Research Commons Team
        - Dryad2Dataverse
     iv. Problem to be solved: Researchers tend to deposit data where they're most comfortable with, which are usually a domain-specific repository. The long-term sustainability of those repositories is questionable and may not allow data to be discoverable as widely.
     v. Created a software pipeline from Dryad software to (any) Dataverse software.
        - Dryad has over 600 datasets from UBC researchers.

- UBC became a member of Dryad so researchers can deposit their data for free.
- UBC wanted to become the data stewards of their own researchers.

vi. Tool developed by Paul that would move data and associated metadata from one place to another. He also wanted the software to be reusable by other institutions.
- Standalone tool also has a Python library that can be used for development.
- Software is not server-based, it's very simple and easy to run on different infrastructure (including a cell phone but not recommended!).
- Basically the software connects to Dryad using an API and then to SP DV using an API, with a database as a holding place for the data between the two systems.
- Metadata fields from Dryad were mapped to SP DV metadata fields. Mapping not available on documentation website. Contact Paul if you want a copy.
- Software tool only uses Python code and has very limited dependencies.
- Python library has 3 components: **serializer** (metadata mapping), **transfer**, and **monitor**.
- **Serializer** essentially takes a Dryad JSON metadata file and converts it to a SP DV JSON metadata file.
- **Transfer** -- due to API limits, can't transfer data from Dryad to SP DV directly. It needs to be downloaded to another database first, then uploaded to SP DV in a separate step.
- **Monitor** component monitors new datasets in Dryad to highlight which data needs to be transferred to SP DV. Monitors both data and metadata.

vii. Main benefits of putting data in SP DV is that it keeps more of UBC's data in one location and SP DV data is discoverable in more places (i.e., FRDR, Harvard DV, etc.).

viii. All data in Dryad has a CC0 license, so there's no concern over transferring the data.

ix. Many of the datasets in Dryad contain very large files. API transfer has a 3GB limit. Approx. 0.5% of the datasets were too large to transfer via software pipeline.

x. All datasets in Dryad already had DOIs assigned.

xi. Dryad2Dataverse tool can be used to move data to any DV instance, not just SP DV.

xii. UBC runs this software about once a week.

xiii. Limitations:
- Intent of the tool is to mirror the dataset in Dryad in SP DV. But SP DV's standard upload process unzips files, which wouldn't allow the datasets to be mirrored.
- SP DV also processes Excel files and converts them to TAB files and the tool isn't really able to bypass this.

- File sizes
- Publication Date - is also the ingest date. This date can only be changed if you go into the DV database directly, which we cannot do. Therefore, publication dates in Dryad do not match the publication dates in SP DV.
- DOIs - a new DOI is generated when the dataset is uploaded into SP DV. Therefore, each dataset has at least 2 DOIs.

xiv. Bulk release utility now available as a standalone program, draydd.py.

xv. https://dataverse.scholarsportal.info/dataverse/UBC_DRYAD

xvi. Contact Paul if you have any questions.

xvii. Questions:
- John H (U of A) - Is there an attempt to link the datasets from Dryad to a specific researcher's account in SP DV?
  a. No, there's limited info about the depositor on Dryad.
  b. Note, this data transfer hasn't been announced or communicated to researchers because the data is all CC0. This process is more of a tool that UBC uses to make the data more discoverable.
- William R (Queen's) - Does the monitoring component of the tool link to versioning in SP DV?
  a. Versioning is not the same between the 2 systems. The first transfer from Dryad to SP DV is Version 1 in SP DV. Subsequent transfers of the same dataset will create a new version. But the version could be completely different in Dryad.
  b. Tool is open-source and can be updated/altered if needed.

4. Open Q&A
   ○ No questions.
5. Next meeting: June 24th (Kaitlin to send updated calendar event)