A Metadata-Driven Data Curation Tool

By: Amber Leahey & Kevin Worthington, Scholars Portal
Background

- Scholars Portal is a digital library project of the Ontario Council of University Libraries (OCUL), consortium of 21 university libraries.
- In 2008, SP launched ODESI a social science survey data repository [http://odesi.ca](http://odesi.ca)
  - Built with the Nesstar repository
  - DDI-XML metadata store (variable-level) in MarkLogic
  - Collection curated by OCUL libraries
- In 2012, SP hosts Dataverse for Ontario libraries and researchers
  - Open-source repository for research data
  - DDI is supported; not full DDI
  - Not as much curation support
- Would like to bring curation + functionality to Dataverse for research data!
The University of Illinois’ Graduate School of Library and Information Science defines data curation as “the active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for reuse over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation.”

In OCUL libraries -- data curation involves managing data with a view to reuse, and often can focus on documentation, metadata enhancement, and quality review, therefore it makes sense to develop tools to assist researchers, data curators, and librarians, with these tasks.
Who are the data curators?

- Librarians
- Library support staff
- Student support staff
- Researchers
- Research managers
- Research assistants
- Data analysts

Really isn’t one job title for data curators!
Data Documentation Initiative (DDI)

- Internationally recognized metadata standard for describing social science and behavioural research data
- Heavily used by data archives (ICPSR, GESIS, UKDA, CESSDA)
- Increasingly used by researchers / large research groups for improved metadata capture and reuse (MIDUS, SOEP, CLOSER)
- XML and RDF representations
- Implemented and supported by a number of repository tools (Nesstar, Dataverse, Colectica)
Dataverse

- Open-source repository developed at Harvard University’s Institute for Quantitative Social Science (IQSS)
- Offers flexible data deposit and access models (self deposit, mediated, curated, restricted)
- Supports general and disciplinary metadata standards (Dublin Core, Social Science, Health, Life Science, Astronomy)
- API-driven
- Modular application support
Dataverse APIs

- APIs are sets of defined protocols and terms for programatically getting two systems to talk to one another
  - Useful for automating workflows (such as data upload / ingest)
  - Useful for developing applications (such as new discovery tools that rely on data / metadata retrieval from one system to another)
- SWORD API (upload)
- Data Search API
- Data Access API
- OAI-PMH (open archives standard for metadata harvesting)
Modular Applications / Integrations

- Modular configuration of the Dataverse code + APIs for data access allow for really awesome new and custom developments that integrate with DV!

  At SP we’ve been able to develop:

  - Dataverse-specific metrics tool
  - Data Explorer
  - Data Curation Tool

Other integrations: OSF, Archivematica, TwoRavens, WorldMap, OJS, and others...

https://dataverse.scholarsportal.info/metrics/?selection=
Data Explorer

- **Framework:** AngularJS
- **UI:** Bootstrap
- **Chart:** Google Charts API
- **Dataverse API:** Data Access
Data Curation Tool (Prototype)

- Framework: Angular 5
- UI: Google Angular Material
- Chart: d3
- Dataverse API: Data Access
View Frequency Metadata

Add Group | liberal
---|---
**All**

- **v134466**
  - Q17: 17. Is Canada becoming more conservative or is it becoming more liberal?

- **v134800**
  - Q4: 4. Do you approve or disapprove of the job Justin Trudeau is doing as leader of the Liberal Party?

---

**Q17: 17. Is Canada becoming more conservative or is it becoming more liberal?**

- **More conservative**: 669
- **More liberal**: 433
- **Neither**: 456
- **Don't know**: 183
### Group Variables

**Forum Research Political Poll - Federal Issues (Canada) 2015**

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Label</th>
<th>Weight</th>
<th>View</th>
</tr>
</thead>
<tbody>
<tr>
<td>v134374</td>
<td>Q1A</td>
<td>1a. If a federal election were held today, which party are you most likely to vote for?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v134345</td>
<td>Q1B</td>
<td>1b. Even though you may not have made up your mind, which party are you leaning towards at this time?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v134615</td>
<td>Q2</td>
<td>2. Do you approve or disapprove of the job Stephen Harper is doing as Prime Minister?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v134478</td>
<td>Q3</td>
<td>3. Do you approve or disapprove of the job Tom Mulcair is doing as Leader of the Opposition?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v134634</td>
<td>Q5</td>
<td>5. Which party do you expect to win the next federal election?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v134600</td>
<td>Q4</td>
<td>4. Do you approve or disapprove of the job Justin Trudeau is doing as leader of the Liberal Party?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>v134338</td>
<td>Q6</td>
<td>6. Do you agree or disagree: prostitution should be legal for sex trade workers?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Edit/Enrich Variable Metadata
Select Weight Variable(s)

Variable Information

ID: v134351
Name: D2
Label: D2. How old are you?
Literal Question
Interviewer Instructions
Universe
Notes
Group
Weight Variable
Is Weight
<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>v134374</td>
<td>Q1A</td>
<td>1a. If a federal election were held today, which party are you most likely to vote for?</td>
</tr>
<tr>
<td>v134345</td>
<td>Q1B</td>
<td>1b. Even though you may not have made up your mind, which party are you leaning towards at this time?</td>
</tr>
<tr>
<td>v134015</td>
<td>Q2</td>
<td>2. Do you approve or disapprove of the job Stephen Harper is doing as Prime Minister?</td>
</tr>
<tr>
<td>v134478</td>
<td>Q3</td>
<td>3. Do you approve or disapprove of the job Tom Mulcair is doing as Leader of the Opposition?</td>
</tr>
<tr>
<td>v134600</td>
<td>Q4</td>
<td>4. Do you approve or disapprove of the job Justin Trudeau is doing as leader of the Liberal Party?</td>
</tr>
<tr>
<td>v134634</td>
<td>Q5</td>
<td>5. Which party do you expect to win the next federal election?</td>
</tr>
<tr>
<td>v134353</td>
<td>Q6</td>
<td>6. Do you agree or disagree prostitution should be legal for sex trade workers and their clients?</td>
</tr>
</tbody>
</table>
## Search and reorder variables

### Forum Research Political Poll - Federal Issues (Canada) 2015

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>v134466</td>
<td>Q17</td>
<td>17. Is Canada becoming more conservative or is it becoming more liberal?</td>
</tr>
<tr>
<td>v134600</td>
<td>Q4</td>
<td>4. Do you approve or disapprove of the job Justin Trudeau is doing as leader of the Liberal Party?</td>
</tr>
</tbody>
</table>

*Search: liberal*
Curation / Metadata Editing Interface


DEMO time!
Discovery & Reuse Use Cases

- View question text, universe statements, coding instructions, questionnaire skip logic, etc. in context of variables
- Apply / tag weight(s) variables in a dataset (used for analysis)
- Group variables to make it easier to find a use large datasets with many variables
- Search on richer sets of metadata; e.g. question text provides original research question vs. label for variable
- Metadata is structured in DDI standard and can be exported for reuse in other data curation projects
Possible curation system architecture

Dataverse

Curation copy (Version 2)

Data Explorer

DDI XML Export

Data Download & DDI Codebook (PDF)

DDI XML upload

Dataset upload (Version 1)

Curator / admin initiates “curate”

DV Data Access API

/ DV Data SAVE API (?)
Next Steps

● Connect to Dataverse DDI metadata saving API
● Develop phase two enhancements
  ○ Multilingual support
  ○ DDI xml upload / reuse
  ○ Additional DDI fields?
  ○ Enhancements to Data Explorer to support discovery and reuse

Conduct UX testing!!
Questions?

amber.leahey@utoronto.ca