Qualitative Data Management for Interdisciplinary Research

KRISTAL JONES, RESEARCH SCIENTIST, NATIONAL SOCIO-ENVIRONMENTAL SYNTHESIS CENTER (SESYNC), UNIVERSITY OF MARYLAND-COLLEGE PARK

STEVEN M. ALEXANDER, MITACS SCIENCE POLICY FELLOW AND SCIENCE ADVISOR, FISHERIES AND OCEANS CANADA

DataONE webinar, April 10, 2018
What is qualitative data?
What is qualitative data?
What role does qualitative data play in interdisciplinary research?

Qualitative data can...

- Provide temporal insights
- Form case comparisons
- Scale up patterns
- Scale down interpretations
- Broaden the evidence base
What are the benefits of sharing and re-using qualitative data?

Scientific
• Transparency and triangulation

Descriptive
• Expansive, inclusive, and varied aspects of a phenomena
• Opportunities for teaching & learning

Material
• Reduce the burden on individuals and communities
• Increased return on investment for funders and institutions
• Access for knowledge users outside of research institutions
Is qualitative data currently being shared?
What challenges exist to accelerating qualitative data sharing and re-use?

Practical challenges
- Resources: Time, expertise, financial support
- Infrastructure: Where to deposit?

Ethical
- Confidentiality, representation, consent

Epistemological
- Spectrum from pure positivism to pure constructionism, with lots of pragmatic space in the middle
Research Data Lifecycle

- Plan & Design
- Collect & Capture
- Interpret & Analyze
- Manage & Preserve
- Release & Publish
- Discover & Reuse
Who plays a role in addressing these challenges?

- Researchers
- Research institutions
- Journals and publishers
- Funders
- Data repositories
Want a summary of the benefits, challenges, resources and recommendations?
What resources exist to address qualitative data management challenges?

**Primary data lifecycle:**

**Plan and Design**

- Data management planning: [https://qdr.syr.edu/guidance/managing/planning-data-management](https://qdr.syr.edu/guidance/managing/planning-data-management)
- IRB: [https://qdr.syr.edu/node/20260/](https://qdr.syr.edu/node/20260/)
- Develop shared protocols: [https://www.atkinson.cornell.edu/collaborations/oxfam-cu.php](https://www.atkinson.cornell.edu/collaborations/oxfam-cu.php)
- Teaching qualitative data management webinar from IASSIST: [https://www.youtube.com/watch?v=aATIKsF96Ro&feature=youtu.be](https://www.youtube.com/watch?v=aATIKsF96Ro&feature=youtu.be)
What resources exist to address qualitative data management challenges?

**Primary data lifecycle:**

**Collect and Capture, Interpret and Analyze**

- Myriad resources from environmental anthropology and sociology, human geography


- Sharing code from qualitative data software: RQDA [https://cran.r-project.org/web/packages/RQDA/RQDA.pdf](https://cran.r-project.org/web/packages/RQDA/RQDA.pdf)

  Code books from Atlas.ti, NVivo, MAXQDA
What resources exist to address qualitative data management challenges?

**Primary data lifecycle:**

**Manage and Preserve**

- IRB and de-identification: [https://qdr.syr.edu/node/20260/](https://qdr.syr.edu/node/20260/)
- File formats: [https://qdr.syr.edu/guidance/managing/formatting-data](https://qdr.syr.edu/guidance/managing/formatting-data)
- DDI metadata standards: [https://www.ddialliance.org/sites/default/files/AQualitativeDataModelForDDI.pdf](https://www.ddialliance.org/sites/default/files/AQualitativeDataModelForDDI.pdf)
What resources exist to address qualitative data management challenges?

**Primary data lifecycle:**
Release and Publish, Discover and Re-use

*What is unique about publishing qualitative data for re-use in interdisciplinary research?*
What resources exist to address qualitative data management challenges?

**Secondary data lifecycle:**
Plan and Design, Collect and Capture

*What is unique about interdisciplinary research using secondary qualitative data?*
What resources exist to address qualitative data management challenges?

**Secondary data lifecycle:**

**Interpret and Analyze**


- Text mining: [http://tm.r-forge.r-project.org/](http://tm.r-forge.r-project.org/)

What resources exist to address qualitative data management challenges?

**Secondary data lifecycle:**
Manage and Preserve, Release and Publish

*What is unique about publishing qualitative data for re-use in interdisciplinary research?*
What resources exist to address qualitative data management challenges?

From a researcher’s perspective, how do we operationalize sharing (preserving and publishing) oriented toward re-use (discover and design)?
What resources exist to address qualitative data management challenges?

- What kind of metadata is necessary?
- Has the data been cleaned and made anonymous?
- Will the data be discoverable?
  - What counts as data?
  - How does epistemology shape what can be shared and re-used?
## Linking qualitative data sharing and re-use: Levels of access

<table>
<thead>
<tr>
<th>Level of access</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A – Open</td>
<td>Data is freely available for use in accordance with general use agreement of repository and standard citation practices</td>
</tr>
<tr>
<td>B - Restricted</td>
<td>Data is available for use when user meets standard criteria set by data repository to ensure ethical use of data (could include use agreement, obtaining IRB or accessing data through virtual environment)</td>
</tr>
<tr>
<td>C - Controlled</td>
<td>Data is available for use by trained users in a controlled environment (access could depend on secondary research questions and intended analysis, controls on access method and amount of data shared is decided by original researcher)</td>
</tr>
<tr>
<td>D - Closed</td>
<td>Data deposit and citation exist for archival purposes but no data are currently available (could be embargoed until publication of results, change in sensitive situation, death of a participant, or certain duration of time from collection)</td>
</tr>
</tbody>
</table>
## Linking qualitative data sharing and re-use: Levels of processing

<table>
<thead>
<tr>
<th>Level of processing</th>
<th>Definition</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – Raw data</td>
<td>Full text, image or audio No aggregation or analysis</td>
<td>No redaction - all identifiers included No additional information about context and methodology</td>
</tr>
<tr>
<td>1</td>
<td>Full text, image or audio No aggregation or analysis</td>
<td>Redaction for direct identifiers Idiosyncratic information about context and methodology</td>
</tr>
<tr>
<td>2</td>
<td>Full text, image or audio No aggregation or analysis</td>
<td>Redaction for direct and indirect identifiers Standardized information about context and methodology</td>
</tr>
<tr>
<td>3</td>
<td>Excerpted text, image or audio Thematic or topical aggregation</td>
<td>Redaction for direct and indirect identifiers Standardized information about context and methodology</td>
</tr>
<tr>
<td>4 – Research output</td>
<td>Summarized text, image or audio Thematic or topical analysis</td>
<td>Redaction for direct and indirect identifiers Summarized information about context and methodology</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>0 [raw]</td>
<td></td>
<td>Public policy documents</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>Public policy documents with search terms as metadata</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Public policy documents with code for web scraping</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Public policy documents organized by theme and with code for thematic analysis</td>
</tr>
<tr>
<td>4 [Research outputs]</td>
<td></td>
<td>Descriptive summary of themes within policies with methodology explained</td>
</tr>
</tbody>
</table>
Takeaways

Researchers should consider data sharing and re-use across all stages of the lifecycle

- This often requires more planning at the outset (IRB, metadata documentation, etc.) for qualitative data than quantitative data.

Interdisciplinary research requires working across systems, vocabularies, tools, etc.

- There are many resources out there, but they often aren’t used in an integrated workflow.
Support for this work comes from the National Socio-Environmental Synthesis Center (SESYNC), which is supported under funding received from the National Science Foundation DBI-1052875.

SESYNC Qualitative Data Initiative: https://www.sesync.org/for-you/cyberinfrastructure/research-and-tools/qualitative-data-initiative

Contact information:
Kristal Jones kjones@sesync.org
Steven M. Alexander s22alexa@uwaterloo.ca