A Pilot Study on Social Science Graduate Students’ Data Core Competency

Xuan Zhou
Data Curation Specialist
Texas State University
TCDL, May 17, 2023

Introduction

What is research data management (RDM)?

- RDM refers to the effective management of research data throughout the entire research cycle, from data entry to distribution and to archiving results (Wilms et al., 2020).

Why RDM is important?

- Appropriate RDM practices is essential to enable data sharing, as mandated by research stakeholders.
- Well-documented and preserved data is easily accessed, interpreted, and used by other interested researchers.
- When informational data and materials can be incorporated into research publications, it makes research more efficient.
Data core competency

Data core competency refers to the skill set needed to access, assess, convert, manipulate, summarize, preserve, and present data, based on a theoretical core concept of socio-educational conditions.

Table 1. Examples of data core competencies proposed by different researchers.

<table>
<thead>
<tr>
<th>Disciplines</th>
<th>Data core competencies</th>
<th>Target audiences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sciences and engineering</td>
<td>• Data culture; data ethics; discovery and collection; data management and organization; data quality; data analysis; data visualization; metadata</td>
<td>Students and faculty</td>
</tr>
<tr>
<td>Sciences (STEM)</td>
<td>• Skills in collecting, processing, managing, evaluating, and using data for scientific inquiry</td>
<td>Students</td>
</tr>
<tr>
<td>Engineering</td>
<td>• Domain-specific competencies; developer competencies; soft competencies; business-oriented competencies; and analytical competencies</td>
<td>Big data professionals</td>
</tr>
<tr>
<td>Social sciences</td>
<td>• Digital materials in the holdings remain accessible, complete, uncorrupted, and usable over time</td>
<td>Data archives</td>
</tr>
<tr>
<td>Humanities</td>
<td>• Humanities data collections need to be available, usable, accessible, interpretable, and critically addressable</td>
<td>Data librarians</td>
</tr>
<tr>
<td>Business</td>
<td>• Data organization and storage; understanding data used in business context; evaluating the quality of data sources; interpreting data; data-driven decision making; communicating and presenting effectively with data; and data ethics and security</td>
<td>Business students</td>
</tr>
<tr>
<td>Multidisciplinary</td>
<td>• Interpersonal and communication skills; Knowledge of research lifecycle and funding agency policies; knowledge on data usage and metadata; and knowledge of information technologies</td>
<td>Data librarians and curators</td>
</tr>
</tbody>
</table>

Carlson et al., 2011; Qin & D’Ignazio, 2010; Gurcan, 2019; Kim, et al., 2011; Gutmann et al., 2006; Padilla, 2016; Pothier & Condon, 2020; Lee & Stvilia, 2017)
Research Gaps & Aims

- There has been a longstanding need to provide ongoing data management instruction for graduate students (Johnston & Jeffryes, 2014; Raszewski et al., 2021).
- RDM instruction for graduate students is still in an early phase (Xu, 2022; Xu et al., 2022).
  - Only a few studies reported specifically tailored RDM instruction for social science graduate students.
- Literature lacks sufficient evidence on RDM instruction needs from the perspectives of student researchers or their level of preparedness in data core competency.
  - Particularly in the social sciences.
- The present study aims to address this research gap by evaluating social science graduate students' perceptions of the significance and preparedness of their data core competency.
Research questions

1. Do social science graduate students’ self-rating importance of data core competency differ based on their department affiliation and years of study, while accounting for gender and race?

2. What are the data core competencies of social science graduate students in areas such as data discovery; data conversion and interoperability; data management plans; data organization, documentation and processing; data analysis and visualization; and data storage, data sharing, and preservation?
Research Design

- A **pilot** study to offer a four-hour instruction about RDM to graduate students in the College of Education.

- The data core competency assessment was done after intervention
  - To assess their fundamental knowledge and skills in managing research data.
  - To identify areas of strength and weakness
  - Findings help design a customized RDM training program for social science research
    - This program can be utilized in the future to enhance participants' data management capabilities and improve the overall quality of their research.
Participants

Four departments:

- Educational Psychology (EPSY)
- Teaching, Learning, and Cultural (TLAC)
- Education Administration & Human Resources (EAHR),
- Health and Kinesiology (HLKN)

Table 2. Demographic information of participants.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Group</th>
<th>Department</th>
<th>Total N = 40</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EPSY N = 22</td>
<td>TLAC N = 10</td>
<td>EAHR N = 2</td>
</tr>
<tr>
<td>Years in program</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 years or less</td>
<td>14</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>&gt;2 years</td>
<td>8</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>21</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Male</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Race/ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White and Asian</td>
<td>17</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Underrepresented</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>minorities</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Underrepresented minorities include students African American, Hispanic or Latino American, Native American, and Pacific Islander.
Measures

● 16 items, designed by the research team to cover every aspect of the research data life cycle
● Part 1 (Q1): overview question, covered 12 aspects of data competency
  ○ Participants' self-perceived importance of being data competent in all aspects of RDM
● Part 2 (Q2-16): Four sections of RDM
  ○ data discovery, data conversion and interoperability
  ○ data management planning (DMP)
  ○ data organization, documentation and process, and data analysis and visualization
  ○ data storage, and data sharing and preservation.
Sample questions of part 1

- 12 items
- Rating from 0 (not important at all) to 10 (Essential)
- Self-perceived importance of being data competent
Results of part 1
Self-perceived importance of being data competent

More experienced the graduate students are, the less important they believe it is for them to be knowledgeable in data competency by the time they graduate.

<table>
<thead>
<tr>
<th>Variables</th>
<th>b</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Department</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TLAC vs EPSY</td>
<td>1.08</td>
<td>1.00</td>
<td>1.07</td>
<td>.290</td>
</tr>
<tr>
<td>EAHR vs EPSY</td>
<td>-1.25</td>
<td>2.06</td>
<td>-0.60</td>
<td>.550</td>
</tr>
<tr>
<td>HLKN vs EPSY</td>
<td>0.49</td>
<td>1.15</td>
<td>0.43</td>
<td>.673</td>
</tr>
<tr>
<td>Years in program</td>
<td>-2.02</td>
<td>0.89</td>
<td>-2.27</td>
<td>.030*</td>
</tr>
<tr>
<td>Gender</td>
<td>2.63</td>
<td>1.43</td>
<td>1.83</td>
<td>.076</td>
</tr>
<tr>
<td>Race</td>
<td>-0.84</td>
<td>0.92</td>
<td>-0.91</td>
<td>.371</td>
</tr>
</tbody>
</table>

Students who studied 2 or fewer years in the graduate program had higher self-perceived importance of being data competent.
Discussions for findings of part 1

- Including RDM as a crucial component of graduate education programs

- Possible reasons of this finding:
  - academic environments focus too much pressure on research, publications, and grants, but not research data management
  - graduate students are influenced by peer network and their supervisor
  - the resources and opportunities from their institution or department

- Our recommendations:
  - Including RDM training in the core curriculum of the research methods course
  - Equip graduate students with the basic knowledge and skills of research data management.
Findings of Part 2

Social science graduate students' data core competencies
Data discovery

- ERIC, EBSCO, APA PsycInfo, ProQuest, Google Scholar, PubMed, and Embase are mentioned as frequently used databases.
- A considerable number of participants demonstrated a lack of familiarity with the concept of databases.
  - Misconceptions: perceived *Google* or *institution’s libraries website* as databases.
  - Need of information literacy training in the curriculum.
Data conversion and interoperability

- Social science research is the activity of gathering, analyzing, and interpreting information for a variety of social, economic, educational and political purposes.
- Involve mixed methods in doing research, both quantitative and qualitative.
- A trend towards quantitative analysis in social science was witnessed.
  - Data formats: spreadsheet, text, and audio/video files
  - Data collection
    - EPSY, TLAC, & HLKN: assessments/tests, surveys/questionnaires, and interviews
    - EAHR students listed interviews, focus groups, & observations
- Proficiency in transforming data from one format (proprietary) to another (nonproprietary)
  - Lack the experience in data transformation and sharing
Data management plan (DMP)

- Students who had >2 years of study are more familiar with the requirements of DMPs. But overall, both groups lack DMP experience.
- No significant differences were found by department.
- Graduate programs should consciously provide more opportunities for graduate students

Table 4. Students’ familiarity with the requirements and/or sharing plans.

<table>
<thead>
<tr>
<th>Familiarity</th>
<th>Years in the program</th>
<th>Total</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 or less</td>
<td>&gt;2years</td>
<td></td>
</tr>
<tr>
<td>Not at all</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>Somewhat</td>
<td>5</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>Very familiar</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>13</td>
<td>39</td>
</tr>
</tbody>
</table>

Note. df=2.
Data organization, documentation, & processing

- 67.5% felt somewhat familiar with documenting and/or creating their own data (n = 27).
- No statistically significant differences in years of experience or in department affiliations.

Fig. 1. Data documentation practice by department.  
Fig. 2. Data documentation by years in program.
Data analysis and visualization

- 76.67 % of participants indicated they do not write code or develop software.
- No significant variations by department or years in the program were identified.
- Open-access software was used more frequently by the senior graduate students.

Fig. 3. Data analysis and visualization by department.

Fig. 4. Data analysis and visualization by years in program.
Data storage

Data Storage:

- 43.59% unknown the size of digital research data they were storing.
- 30.77% in gigabyte range; 17.95% in megabyte range; 7.69% in terabyte range
- No departmental or years of study differences.

Methods of data storage or backup:

- Top ranked:
  - Desktop/laptop computers
  - Internet-based storage (i.e., Dropbox, Google Drive)
  - External hard drives (i.e., USB drives)
- Students with <2 years in program use departmental or university servers to store data
- Students with >2 years in program use external hard drives more frequently in addition to desktop and internet-based storage.
Data sharing and preservation

- Top ranked approaches:
  - E-mail upon request
  - Supplementary materials linked to journal articles (e.g., PloS Journals, Dryad)
- No trend found to share the data with the public voluntarily for social science graduate students
- Limited knowledge of data repositories:
  - 52.5 % were not aware of it
  - 97.37 % have not deposit data in any data repositories
    - only 4 of them provided correct information (i.e., TDR, OpenNeuro, Qualitative Data Repository, and Health and Medical Care Archive).
- Interested in starting deposit data in a data repository
  - No statistical differences were identified between departments or years of study.
    - somewhat interested (84.21 %)
    - very interested (13.16 %)
    - only 1 was “not interested
Theoretical and practical implications

- Graduate students overall did not appreciate RDM practice
- Incorporate RDM components into their core curriculum
- It is essential for academia to incentivize and reward best practices in RDM
  - Include open access as a value of promotion and tenure guidelines to reward OA scholarship
- Mandate the practice of data sharing.
  - Researchers need to shift their mindset from considering data sharing as a burden to recognizing it as an integral part of a research project
References


Questions

Xuan Zhou, x_zhou@tstate.edu

Zhihong Xu, xuzhihong@tamu.edu
1. Please indicate how important you believe it is for you to be knowledgeable in each competency by the time you graduate
   (Essential, very important, somewhat important, not important, I don’t know, or N/A)
   
   - Databases and data formats: appropriate data types and formats, concepts of relational databases
   - Discovery and acquisition of data: locating external sources of data, downloading and using data files
   - Data planning and organization: developing data management plans, creating standard operating procedures, file naming conventions, file versioning, tracking components of data sets
   - Data Conversion and Interoperability: standard data formats, migrating data from one format to another
   - Quality Assurance: data consistency and completeness, data corruption/loss, data security and backup
   - Metadata and Data Description: annotation for data understanding and reuse, metadata schemas, controlled vocabularies and ontologies/classifications, reproducibility
   - Data Curation and Reuse: data lifecycle from raw stage to outputs, value beyond initial purpose, funder data sharing policies, identifiers
   - Cultures of Practice: discipline-specific norms, standards and practices for managing research data
   - Data Preservation: benefits and costs, technical and resource considerations for long-term storage of data
   - Data Processing and Analysis: workflows and analysis tools, repetitive tasks automation, data summary and calculations
   - Data Visualization: types of visual data representations, avoiding misrepresentation
   - Ethics and Attribution: privacy and confidentiality, intellectual property, citing data
2. Data discovery, and data conversion and interoperability
   • Please list up to three databases you use most often for your research

   • Please list up to three of the most common formats you collect or generate from your research (e.g., spreadsheets, text, images, videos, audio files, instrument files, photographs, physical samples/specimens, etc.)

   • Please list up to three instruments used most frequently in your research (e.g., assessment/test, survey/questionnaire, interview, focus group, etc.)

   • Are you proficient in transforming data from one format (proprietary) to another (nonproprietary)?
     • Not at all proficient
     • Somewhat proficient
     • Very proficient
     • Not sure

3. Data Management Plan (DMP)
   • How familiar are you with the requirements and/or sharing plans as components of many funding applications (e.g., NSF, NIH, IES)?
     • Not at all familiar
     • Somewhat familiar
     • Very familiar
4. Data organization, documentation and processing, and data analysis and visualization

- How familiar are you with documenting and/or creating data for your data (i.e., so the contents of datasets can be understood by others and/or computer-readable)?
  - Not at all familiar
  - Somewhat familiar
  - Very familiar

- What supporting documentation do you generate about the data so that the data are interpretable, replicable and/or reproducible? (Choose all that apply)
  - Codebooks
  - Notebooks
  - Readme files
  - Use of naming conventions
  - (Electronic) lab notebooks/logbooks
  - Field notes
  - Documentation of the used or created hardware/equipment
  - Documentation of the used or created software
  - Other (Please specify)

- Do you write code and/or develop software? If so, how do you make it understandable and reusable? (Choose all that apply)
  - I do not write code or develop software
  - I write code and/or develop software, but don’t know of any best practice
  - I use comment in code
  - I make scripts publicly available
  - I use version control to manage code
  - Other (Please specify)

- What software do you use to analyze/visualize your data? (Please specify)

5. Data storage, and data sharing and preservation

- Approximately how much digital research data are you currently storing?
  - Megabyte range
  - Gigabyte range
  - Terabyte range
  - Don’t know

- What is your current method of data storage or back-up? (Choose all that apply)
  - Desktop/laptop computers
  - External hard drives (including USB drives)
  - Department/university server
  - Internet-based storage (Dropbox, Google Drive, etc.)
  - Lab books/field notes/other printed or handwritten materials
  - CDs/DVDs/tapes
  - Other (please specify)

- How do you share your data with others? (Choose all that apply)
  - E-mail upon request
  - Supplementary material linked to journal article (e.g., PLoS journals, Dryad)
  - Data repository (e.g., Texas Data Repository)
  - Department/University website
  - Personal website
  - Others (Please specify)

- Are you aware of any disciplinary repository for your research field? (Please specify if yes; put n/a if not applicable)

- Do you currently deposit some or all of your data in a data repository (e.g., ICPSR, TDR)?
  - Yes
  - No

- If yes, how interested are you in starting to deposit your data in a data repository?
  - Not at all interested
  - Somewhat interested
  - Very interested