FINAL PROPOSAL

Name: Amor Ai

Project title: Facilitating Data Reuse Among Researchers

I. Research questions/hypotheses

I intend to use my research to gain a better understanding of the factors and roadblocks that may prevent existing data from being discovered, utilized, and reused. Reusing data brings about an abundance of benefits—"maintains research continuity, saves time and resources, fosters interdisciplinarity," to name a few (Emerson, 2020). In addition, the large amounts of data produced in research comprises a wealth of valuable information and resources that can be used by both researchers who collected the data and those who did not collect the data to extract new information for data-driven decisions or new discoveries. As a result, when obstacles prevent existing data from being extracted and used, a vast amount of valuable resources—such as time, money, and labor—are being neglected and ultimately lost. Because of the many challenges that prevent existing data from being found and reused, it is crucial to better understand how researchers and research support professionals discover data—if their individual demographics or perceptions play a factor—and consequently locate the various obstacles that may hinder future data reuse. Furthermore, analyzing the ways in which different factors precipitate data reuse can help research support professionals better understand how to assist researchers looking for data.

The goal of my theory-building research is to analyze the following: How do factors such as a researcher's demographic information and their perceptions about sharing and using data influence and contribute to their behavior or success in data reuse? Are there any patterns that could help researchers benefit more from reusing data or inform research administrators of new ways to facilitate data reuse? The foundation of this research will stem from participating in the Statistics & Data Science project with the CMU Libraries, exploring an existing survey dataset and finding potential patterns that could explain data reuse behavior of researchers and research support professionals. I will be working with 3 team members, Breana Valentovish, Xavier Xia, and Xiaohan Liu, in order to extract important analyzes from the survey in order to assist research support professionals in finding and utilizing more effective instruments to facilitate data reuse in the future.

While a previous publication has done exploratory data analysis on the same survey dataset, they focus primarily on "the data needs and discovery strategies of respondents, [proposing] a typology for data reuse, and [probing] the role of social interactions and literature search in data discovery" (Gregory et al., 2020). While incredibly insightful, more in-depth analysis is needed beyond the findings from this paper. Therefore, my research is novel as it specifically analyzes the researchers and research support professionals' demographics, along

with their perceptions on data—providing a more thorough understanding of the specific factors that either prevent or positively influence data reuse.

II. Data and Methods

In 2020, a large survey asked respondents' data needs, their data discovery practices, and their methods for evaluating and making sense of secondary data. The data from this survey was published in the DANS data repository, which I will be using to conduct my exploratory data analysis (Gregory). Questions in this survey include asking about the respondents' data needs, their data uses, their data discovery practices, their methods for evaluating and making sense of secondary data, and many more variables.

From this survey, I will be using the questions and corresponding categorical data relating to the respondents' indicated demographics (experience, country of employment, discipline) and their perceptions about sharing and/or reusing secondary data. Utilizing exploratory data analysis techniques using R software, I will be able to produce several visualizations and conclusions about the patterns that arise from my observations. The visualizations may include bar charts, histograms, pie charts, classification trees, word clouds—all of which seek to make sense of any relationships between the variables. These visualizations will shed light on the factors and correlations that influence data reuse, and by the approach of theory-building, my research can hopefully inform research support professionals, such as librarians, on ways and necessary resources to better assist researchers in reusing data.

III. Expected Outcomes

When data is not being reused by researchers, large amounts of valuable information and resources are simply wasted. However, the exact barriers preventing researchers from assessing and reusing data are still unclear. Therefore, using an existing survey about data practices will allow me to first understand how researchers discover, make sense of, and reuse data for research. Consequently, I will be able to create visualizations and identify correlations using R software in order to discover concrete patterns about the relationships between the respondents' demographics and their perceptions about sharing data and reusing data.

While Professor Peter Freeman will evaluate my final work, I will be periodically meeting with a PHD student, Alan Mishler, who will assess my progress (and our team's) and discuss any problems and results. In the end, my analyzes will be presented to (via presentation slides) and evaluated by members of the CMU Library, specifically Huajin Wang and Sarah Young. The overarching goal of my research is to predominantly focus on building knowledge and helping people like research administrators and librarians gain a better understanding of the many factors that may influence data reuse behavior in research.

References

[1] Emerson, K. (2020). *The Benefits of Reusing Research Data*. Research Data Services. https://researchdata.wisc.edu/uncategorized/the-benefits-of-reusing-research-data/.

[2] Gregory.(2020). Data discovery and reuse practices of researchers and research support professionals.[Data set]. DANS-EASY. DOI.

[3] Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2019). Lost or found? Discovering data needed for research. arXiv preprint arXiv:1909.00464.