Texas A&M University
(Department of Agricultural Leadership, Education, and Communications)

Data Management Plan: From assessment to implementation: Creating a standardized data competency measure and discipline-based RDM module

Project Overview and Expected Data Types

This three-year project will use a mixed-methods research approach to design a standardized measure/survey for data core competency, design a discipline-based and evidence-based research data management (RDM) module, and conduct an experimental/intervention study via pre- and post-design in graduate students' research methods courses. The goal is to equip social science graduate students (specifically, in library information science and education) with the necessary skills and knowledge to manage research data effectively and efficiently.

The project will employ a systematic approach to develop a standardized data core competency measure/survey. To achieve this, the project will conduct a systematic literature review to identify the relevant literature concerning data core competency and corresponding assessment for data core competency. Simultaneously, the project will conduct in-depth interviews with social science faculties and focus group interviews with graduate students to identify the components of data core competency. To validate the designed standardized measure/survey, the project team will collect survey data and conduct psychometric validation. This will involve item-level descriptive analysis, parallel analysis, and exploratory factor analysis. To assess the effectiveness of the designed standardized measure, the project team will design a RDM module and embed it in selected graduate-level research methods courses. An experimental study will be conducted via pre- and post- assessment, using ANOVA and regression analysis.

In the research process, the project team will produce various types of data:

1) research data such as surveys, one-to-one faculty interview, focus-group graduate students' interview, and experimental study data. Comma Separated Value (.csv) files will be generated to store tabular data collected from Qualtrics surveys or experimental study data. Plain Text (.txt) files and audio data (.mp3) will be generated to store interview data.
2) education data such as RDM module, curriculum, number of students enrolled/participating in the project, etc. PDF/A (.pdf) files and MPEG-4 (.mp4) will be generated to store course curriculum, and course materials.
3) research products data such as conference proceedings and publications. This project will generate conference and journal publications for dissemination purposes. All the research products will be preserved in PDF files.
4) digital metadata, including readme files and codebooks, will also be created for each type of primary data by the project team in (.txt) file.

Sensitive Information

To protect the privacy of participants in the survey and interview process, the project will implement a de-identification process for personal information obtained in the data collection. Firstly, the project will restrict access to the raw data and only allow PI or Co-PI to access sensitive information. Next, the project will utilize pseudonymization to remove identifying information from raw data. This will involve replacing identifiable information, such as names, age, and gender, with randomly generated strings. The project team will ensure that the identity of the data subject and the data about them is impossible to link together. The consent form will inform participants that their data will be shared with the public after de-identification. By implementing these measures, the project team will ensure the privacy of participants is protected, while promoting transparency and open access to research data.

Requirements and Dependencies

To promote long-term preservation and open access to the data generated in the project, the project will use open, non-proprietary formats to store all the documents. The use of open source software such as R and RStudio will be prioritized throughout the project, and R.script will be generated and stored for all data processing and analysis. Additionally, the project will create digital metadata, including readme files and codebooks, for each type of primary data.

Documentation

To ensure the research process is transparent and reproducible, the project team will capture consent agreements, data documentation, codebooks, metadata, and analytical and procedural information through the research process. All documentation will be stored in open formats, such as .csv or .txt file, and in digital format.

The published documentation will include the information necessary to read and interpret the data, such as file structures and instructions for R.script. Raw data will be preserved in a way that includes self-identifying information, which helps preserve provenance. Additionally, the project will use file naming conventions that provide a second level of visibility into the data's provenance. To ensure that the data is fully described and understood, the project team will manually generate descriptions and other metadata as appropriate.

Post-Project Data Management

All quantitative project data will be shared via the Texas Data Repository (TDR). TDR provides long-term preservation of digital objects using an off-site backup and assigns a Digital Object Identifier (DOI) for citations and discoverability. By default, data is shared with a CC0 public domain dedication. Data will be accompanied by documentation, metadata, and code to facilitate reuse and provide the potential for interoperability with similar data sets. The repository provides bit-level preservation and ensures ongoing access to research data, including associated metadata and documentation for a minimum period of ten years after it is deposited. All data will be retained for a minimum of three years after the conclusion of the project or public release (publication), whichever is later. Data related to a student's research work will be retained at least three years after the student has graduated. Access to the database and associated software tools generated under the project will be available for educational and research purposes.

Review and Monitoring

The PI and Co-PIs will oversee the data management plan, with updates included in progress reports and a final report that details all managed products. Intellectual property rights will be retained by the universities and investigators. Curators at the University Libraries will ensure compliance with data sharing requirements to make data FAIR (Findable, accessible, interoperable, reusable).

All the data generated from the project will be stored and shared by the TAMU team. Dr. Zhihong Xu (PI) will serve as data manager, with assistance of the graduate student in the project, and engage in quarterly quality assurance checks with team members to ensure data integrity and document the chain of custody among different datasets.