Integrated Research Infrastructure for Social Science Project Report

ARDC HASS RDC and Indigenous Research Capability Program

5.2 WP5 Technical report_ Schema for Systematic Management of Survey Data

30/11/2023

LEAD ORGANISATION: Australian National University (ANU), Australian Data Archive (ADA), ANU Centre for Social Research and Methods (CSRM), Research School of Social Sciences (RSSS)

PARTNER ORGANISATIONS:

- University of Queensland, Institute for Social Science Research (ISSR)
- University of Melbourne, Melbourne Institute
- Australian Urban Research Infrastructure Network (AURIN)
- Australian Consortium for Social and Political Research Inc (ACSPRI)

PROJECT LEAD: Assoc. Prof. Steven McEachern (Project Lead)

PROJECT COORDENATOR: Keila Silva

A Schema for Structured and Systematic Management of Survey Data

Abstract

In social science research, surveys are undertaken to collect information regarding specific social phenomenon (Mogalakwe, 2006). Social science constructs have complex meanings and this makes it difficult to capture all the dimensions of a concept (Babbie, 2020). However, it is a complex and challenging task to handle survey data due to the volume of information and the complexity of relationships between different codes, questions, variables and other items. Previous researches have used well-designed database schema to provide a structured approach to managing social science research data, allowing the researchers efficiently query and analyse the data (George, Voorhis, & Lee, 1994).

In this paper, we propose a cross-referencing schema to manage survey data. The schema includes entities related to surveys. The schema stores information about each item in the survey, and also allows for the creation of cross-references between items, enabling researchers to analyse relationships between different items as well as managing a list of controlled vocabulary terms used to standardize the terms used in surveys.

Our cross-referencing schema provides a structured and systematic approach to manage survey data, and is also capable of efficiently querying the survey items. The schema can be used by researchers in a variety of fields, including social science. The schema can also be expanded to allow for user-generated cross-referencing, providing users with the ability to create their own cross-references between survey items.

The proposed schema has the potential to improve the quality of survey data and analysis by using a structured and standardized approach to managing survey data. Future research could focus on testing the schema in a variety of settings and comparing its value to other survey data management approaches.

Introduction

Surveys are a common data collection method used in social science research. They are used to gather information about people's attitudes, beliefs, and behaviours on various social phenomenon (Mogalakwe, 2006). Managing and organizing survey data is a complex task, especially when dealing with large datasets. It is important to have a structured and systematic approach to manage survey data in order to ensure data quality, reduce errors, and increase efficiency in data analysis.

This schema provides a way to organize survey data in a structured and systematic manner. It includes tables to store information about surveys, survey items, controlled vocabularies, and cross-references between survey items. The schema allows for easy querying and retrieval of survey data, enabling researchers to perform analysis efficiently.

One of the key features of this schema is its ability to handle cross-referencing of survey items. Cross-referencing allows researchers to identify and compare similar concepts across different surveys. This can help identify trends and patterns across surveys, as well as facilitate meta-analysis.

In addition to the cross-referencing functionality, this schema also includes a controlled vocabulary table. This table allows for standardized and consistent terminology to be used across surveys. By using a controlled vocabulary, researchers can ensure that data is consistent and comparable across surveys, further improving the quality and validity of the data.

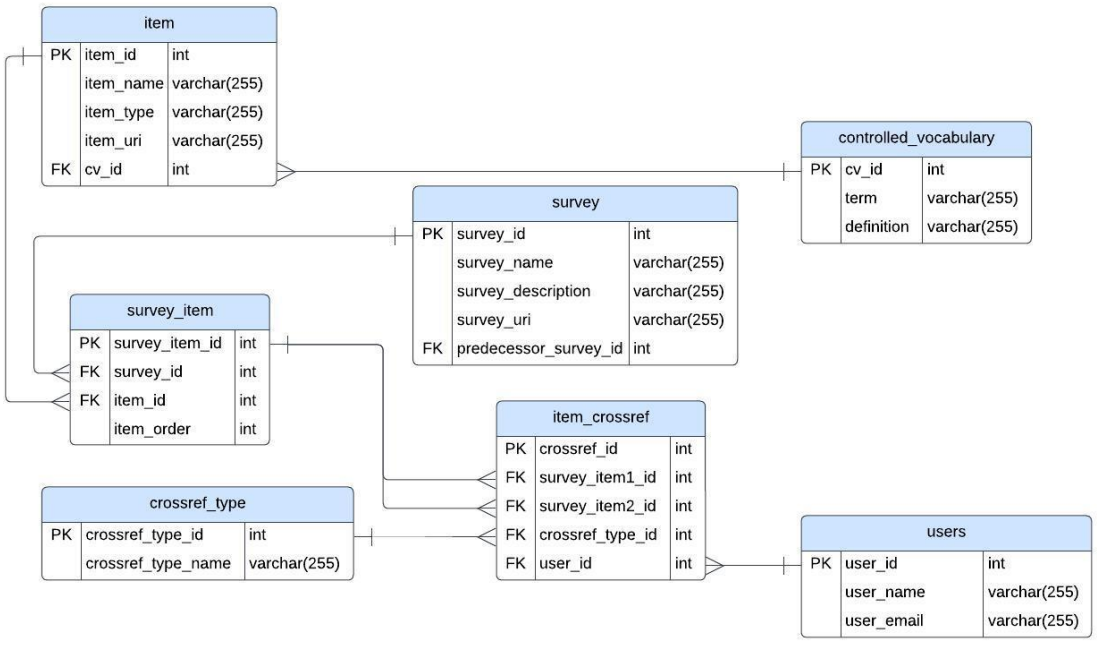
This paper provides a detailed description of the schema, along with other existing tools which can be used to populate data in this schema. We believe that this cross-referencing schema has the potential to improve the management and analysis of survey data, and provide researchers with a valuable tool for data analysis in social science research.

Schema Design

The importance of schema design for managing survey data has been emphasized in previous research (John, 2022). Our proposed schema design can be used for managing survey data using a cross-referencing approach.

The use of cross-referencing in survey data management can help in improving data quality and completeness. Our proposed schema design employs cross-referencing in the ITEM_CROSSREF table to enable the identification of common items in surveys and facilitate data harmonization.

The use of controlled vocabularies in survey data management can be used to enhance data standardization and comparability. Our schema design includes a CONTROLLED_VOCABULARY table to store information related to controlled vocabularies used in the survey items.



User-defined Cross-reference and Controlled Vocabulary

The cross-referencing schema provides a structured and systematic approach to efficiently organize and retrieve survey data, making it an ideal platform for data management and analysis.

The user-defined cross-reference functionality enables the users to create their own mappings between items in different surveys beyond the predefined cross-references. This feature provides

users with the flexibility to tailor the cross-referencing functionality to their specific needs and research questions.

The user-defined cross-reference functionality is implemented through the ITEM_CROSSREF table, where users can insert their own mappings between items. The table has three columns: crossref_id, survey_item1_id, and survey_item2_id, where survey_item1_id and survey_item2_id are foreign keys that reference the SURVEY_ITEM table. Users can insert new cross-reference mappings by adding a new row to the ITEM_CROSSREF table, with the corresponding survey_item1_id and survey_item2_id values.

The user-defined cross-reference functionality is particularly useful when comparing surveys that have unique or specialized items, or when conducting research that requires customized cross-referencing. For example, if a user wants to compare two surveys on mental health, but the surveys have different items related to mental health, the user can create their own mappings between the two surveys' mental health items. This allows the user to make meaningful comparisons between the two surveys, even though they were not originally designed for direct comparison.

The CONTROLLED_VOCABULARY table provides a mechanism for standardization of terms and definitions, which is essential for data analysis across multiple studies. This feature enables researchers to more easily compare and synthesize data from different studies, providing a foundation for future research and analysis.

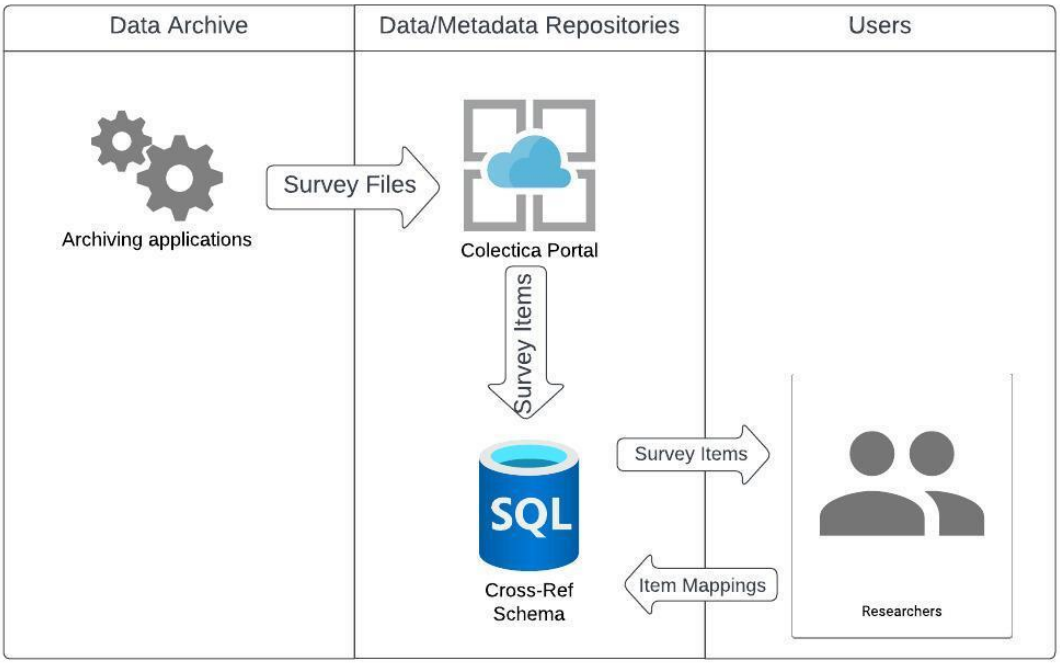
Overall, the proposed schema provides a comprehensive platform for data management and analysis, offering a flexible and systematic approach to organizing and retrieving survey data. The user-defined cross-referencing feature, in particular, offers a unique opportunity for researchers to identify relationships between items that may not have been previously considered, opening up new avenues for research and analysis.

Using Colectica to Populate the Schema

Colectica is a software tool that provides comprehensive solutions for metadata management based on DDI Lifecycle metadata standard (Iverson & Smith, 2014). It can be used to populate the proposed schema with metadata about surveys and survey items.

Colectica provides an intuitive user interface to import survey files into the repository. The software can be used to import existing surveys. Once the survey is imported into Colectica, Colectica can be used to extract metadata about survey items, such as their name, type, label, description, and response options. Colectica can export survey metadata in various formats including DDI. The survey metadata can be exported in a format that is compatible with the proposed schema. The exported survey metadata can be imported into the SURVEY, ITEM, and SURVEY_ITEM tables of the proposed schema.

Using Colectica to populate the cross-referencing schema can help ensure that the metadata is accurate and complete. Colectica's user interface provides a user-friendly way to import survey files and extract survey metadata, making it easier for researchers to create and manage complex surveys.



Conclusion

The use of controlled vocabularies helps researchers in being more specific when tagging information, which makes searches more precise for future (Lykke, Høj, Madsen, Golub, & Tudhope, 2011). The use of a relational database model is also well-established in data management, the proposed cross-referencing schema provides a structured and systematic approach to managing survey data. By using a relational database model, it ensures that data is consistent, accurate, and easily accessible. The schema also allows users to define their own cross-references, providing flexibility and customizability to suit different research needs. The data management and analysis benefits of the schema are significant, and its potential applications are vast. By integrating Colectica with the cross-referencing schema, researchers can streamline their survey management and data analysis processes. A powerful tool for managing and analysing survey data, the cross-referencing schema can help researchers generate insights and make informed decisions that drive progress and innovation in their respective fields.

References

Babbie, E. (2020). Chapter 5. In *The Practice of Social Research* (pp. 124-157). Cengage. Retrieved from https://books.google.com.au/books?id=KrGeygEACAAJ&printsec=frontcover#v=onepage&q&f=false

George, R., Voorhis, J. V., & Lee, B. J. (1994). Illinois's longitudinal and relational child and family research database. *Social Science Computer Review, 12*, 351-365. doi:https://doi.org/10.1177/089443939401200302

Iverson, J., & Smith, D. (2014). Colectica 5 and DDI 3.2: The Next Generation of Metadata Tools. North American DDI (NADDI) Conference. Retrieved from https://summit.sfu.ca/item/13922

John, K. (2022). *Designing a Database Schema for Survey Questions.* Thesis, University of Magdeburg, Faculty of Computer Science. Retrieved from https://wwwiti.cs.uni-magdeburg.de/iti_db/publikationen/ps/auto/thesisJohn22.pdf

Lykke, M., Høj, A. L., Madsen, L. N., Golub, K., & Tudhope, D. (2011). Tagging behaviour with support from controlled vocabulary. *Proceedings of the ISKO UK Second Biennial Conference*, (pp. 41-50). London. Retrieved from https://books.google.com.au/books?printsec=frontcover&vid=ISBN1780526148&redir_esc=y#v=onepage&q&f=false

Mogalakwe, M. (2006). The Use of Documentary Research Methods in Social Research. *African Sociological Review / Revue Africaine de Sociologie, 10*(1), 221-230. Retrieved from http://www.jstor.org/stable/afrisocirevi.10.1.221