Put Relational Databases in Your Data Curation Toolbox

Jenna Jordan
University of Illinois at Urbana-Champaign
Champaign, USA

ABSTRACT

This poster paper makes a case for the use of relational databases as a data curation tool, especially for datasets that are published separately but can be used together due to a common identifier scheme and shared attributes. The Correlates of War datasets are used as an illustrative example to show how the normalization process results in a design with greater data reusability, while check constraints and foreign key constraints can improve data quality.

KEYWORDS

Digital data curation; digital humanities; information management; knowledge management; information design; information organization; knowledge organization

ASIS&T THESAURUS

Relational databases; data curation; international aspects

INTRODUCTION

This project began as an attempt to recreate the many disparate datasets in the Correlates of War (CoW) project[1] as one cohesive relational database. The CoW project is a collection of datasets that contains empirical data on large state conflicts, and "dominates the field of quantitative research into the onset of interstate military conflict" (Travlos & Rudkevich, 2016). It is currently maintained and hosted by several institutions in a system of "coordinated decentralization" that is designed to distribute the cost of curating and updating such a large collection of datasets (Izmirlioglu, 2017). In theory, the datasets are subject to strict rules to maintain data integrity and interoperability.

However, the CoW datasets have several organizational differences that makes successfully merging them together difficult (and in some cases impossible). The datasets required significant reorganization and transformation in order to fit the relational model. The process of transforming the datasets revealed unexpected issues with the data quality as well. The resulting database has greater data quality and the normalized structure allows all of the datasets to be used together; creating the potential for researchers to utilize this source of international conflict data in new ways and thus more fully realizing the goal of the Correlates of War project. The database file and data transformation code are publicly available on GitHub[2].

The task of reorganizing the datasets ended up being an exploration of how the strict constraints and normalization standards of well-designed databases can be a useful tool for data curators. The same method – of taking a set of interrelated datasets, designing a relational model that can account for all of the information stored in the datasets, transforming the data to fit the database design, and then loading the data into the database with strict constraints in place – can be used to improve the quality and reusability of datasets in any field or discipline.

DATABASES AS A DATA CURATION TOOL

In the social sciences, it is becoming more common to share the data that research is based upon. However, this data usually comes in the form of a dataset, which may have been produced by hand – each value typed into a cell. Social scientists may be producing this dataset for one specific purpose or type of analysis, which affects their data design and formatting choices. This means that while the data may be shared, it is not necessarily easily re-usable.

Reusability is one of the areas of broad concern for data curation, and "repositories aiming to support future users may have to curate for data reuse by researchers in other communities" (Chao, Cragin, & Palmer, 2015). The relational database model is a valuable tool for data curators who wish to improve data quality and reusability.

The first step in designing a relational database is to normalize the data, so that the variables are organized into tables based on their functional dependences. For R users, normalized data is 'tidy' data – that is, data structured to facilitate analysis. Wickham's framework of tidy data (each observation a row, each column a variable, each table a type of observational unit) is essentially Codd's third normal form for relational databases, but framed for statisticians (Wickham, 2014). In other words, transforming datasets to fit the relational model results in datasets ready to be analyzed.

METHODOLOGY

In order to design the database, the first task was to identify the functional dependencies. In some cases, the documentation did not make this clear, and the database had

[1] https://www.correlatesofwar.org/data-sets
[2] https://github.com/jenna-jordan/international-relations-database

to be redesigned several times when the data did not reflect the assumed dependencies. Furthermore, many variables had a specific set of possible values which could be enforced through check constraints. This method of designing and filling a database forces the curator to identify and verify every assumption made about the data, without having to personally look through every cell in every table.

The CoW datasets were well-suited to this methodology for several reasons. First, they shared a common identifier scheme for important entities (states and conflicts). Second, the entities (mostly) shared the same attributes across datasets. Third, the datasets all had extensive documentation. All three of these traits are crucial for being able to safely reorganize the data in a meaningful way.

Data Quality

Unlike a spreadsheet, a (well-designed) database does not allow data to be entered that violates the specified constraints. Check constraints can be useful in catching data entry errors. For example, one row in the Intra-State War table lists a day as "-91866", while the following column (which is supposed to contain a year) is blank. Clearly, the person entering the data forgot to tab over to the next column. The day should have been "-9" (which means "unknown") and the year "1866". This mistake is easy to make in a program like Excel, and hard to detect after the fact. However, if the data were entered into a database, it would be found (and trigger an error) immediately. The "day" column could be specified to be a number with two digits, and a check constraint could further restrict that number to being between 1 and 31. While this method may not catch all such data entry errors, it will catch the most obvious.

Foreign key constraints can help catch errors that center on how different variables relate to each other. For example, the Territorial Change table has three columns (Gainer, Loser, TerritoryID) that relate to a single identifier column in the Polity table (PolityID). However, there is one territory present in the Territorial Change table that is not present in the Polity table – '822' – while the territories '8221', '8222', '8223', '8224', and '8225' are present in the Polity table. The original CoW documentation for the Territorial Change dataset makes clear that this is the pattern followed when a territory is broken up into component territories for an updated version of the dataset. So, this missing identifier '822' is evidence of a data versioning issue – a data quality issue revealed by the use of foreign key constraints.

Data Reusability

One of the advantages of relational databases is that they are self-describing: which combination of variables forms the primary key, which variables have relationships with other variables (thus creating variables to merge on), and even what the possible values for variables are.

Spreadsheets, however, are not self-describing – codebooks are required to understand what is happening in a table. However, most codebooks do not explicitly state the functional dependencies – that's up to the data curator to guess and check. Finding the functional dependencies and designing the CoW database was an iterative process.

The four War datasets (intra-, inter-, non-, and extra-state war) form the core of the CoW project – however, despite representing the same entity type (conflict), they are organized in one of three different ways: one row per war; one row per war and country; or one row per war, a side A country, and a side B country. These datasets cannot be used together as they are now, so researchers are limited to studying one type of war at a time. Furthermore, this difference in organization has resulted in other variables being inconsistent between tables. For example, the 'Outcome' column is either coded '1' for 'side A won' and '2' for 'side B won', or '1' for 'this side won' and '2' for 'this side lost'. These inconsistencies also prevent merging.

To solve these problems (and others), I designed a set of war tables – one for the wars (with a war ID primary key), one for the war participants (with a primary key of the war ID, polity ID, and start date), one for the multivalued region variable, and one for the recursive relationship of war transitions. This design is consistent with Codd's third normal form (Codd, 1970). The variables were standardized across tables, no information was lost, and the new format is both more flexible and space efficient, with the added advantage of allowing for analysis across war types.

CONCLUSION

During the process of transforming the CoW datasets into a single database, I caught and corrected data entry errors, discovered data versioning issues, and reconciled different organizational designs used for the same entity type across datasets. I used functional dependencies to reorganize datasets into normal (or 'tidy') form, with relationships maintained by foreign key constraints. The unforgiving nature of database management systems forced every assumption to be checked and verified. This process would be a useful addition to any data curation workflow.

REFERENCES

Chao, T. C., Cragin, M. H., & Palmer, C. L. (2015). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. *Journal of the Association for Information Science and Technology, 66*(3), 616-633.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13*(6), 377-387.

Izmirlioglu, A. (2017). The Correlates of War Dataset. *Journal of World-Historical Information, 4*(1).

Rudkevich, G., & Travlos, K. (2014). Do We Know Too Much About Military Conflict? A Rapid Evidence Assessment of Quantitative Explanations of Interstate Conflict Onset. In *International Studies Association Convention*.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59*(10), 1-23.