Tools for integrating data by complex, dynamic categories

AUTHORS SECTION

Hruschka, Daniel Arizona State University, USA| dhruschk@asu.edu

Cheng, Yi-Yun Rutgers University, USA | yiyun.cheng@rutgers.edu

Hsiao, I-Han Santa Clara University, USA | ihsiao@scu.edu

Bischoff, Robert Arizona State University, USA| bischrob@gmail.com

Peeples, Matthew Arizona State University, USA | Matthew.Peeples@asu.edu

Kasi, Harsha Santa Clara University, USA | hkasi@scu.edu

Huang, Cindy Arizona State University, USA | CindyHYHuang@asu.edu

ABSTRACT

A key challenge in conducting comparative analyses across social units, such as religions, ethnicities, or cultures, is that data on these units is often encoded in distinct and incompatible formats across diverse datasets. This can involve simple differences in the variables and values used to encode these units (e.g., Roman Catholic is V130 = 1 vs. Q98A = 2 in two different datasets) or differences in the resolutions at which units are encoded (Maya vs. Kaqchikel Maya). These disparate encodings can create substantial challenges for the efficiency and transparency of data syntheses across diverse datasets. We introduce a user-friendly set of tools to help users translate four kinds of categories (religion, ethnicity, language, and subdistrict) across multiple, external datasets. We outline the platform's key functions and current progress, as well as long-range goals for the platform.

KEYWORDS

Ontology matching, knowledge organization, cultural informatics, data integration

INTRODUCTION

Scientists and policymakers are increasingly leveraging complex, multi-scale data from diverse, worldwide sources to understand the causes and consequences of economic development, social stratification, climate change, cultural diversity, and violent conflict. This work frequently requires integrating data across diverse datasets by complex, dynamic categories (e.g., ethnicities, languages, religions, subdistricts). However, different datasets usually encode corresponding categories in disparate formats and at different resolutions (e.g., Guatemala Indigenous vs. Maya vs. K'iche') that must be translated across these datasets before bringing them together for analysis (Cheng & Ludäscher, 2020a, 2020b; Hruschka, Bischoff, Peeples, Hsiao, & Sarwat, 2022). Large-scale data with thousands of categories leads to combinatorial complexity hence making it challenging for manual data reconciliation, data sharing, and data reuse (Faniel, Kriesberg, & Yakel, 2012; Sakai, Miyata, Yokoi, Wang, & Kurata, 2020).

Our team has developed a user-friendly, web-based app (CatMapper, https://www.catmapper.org/js) to help users translate four kinds of categories (ethnicity, language, religion, subdistrict) across multiple, external datasets (Hruschka et al., 2022). CatMapper's key functions include: (1) **explore** contextual information (e.g., geospatial location, population estimates, language, religion, alternative names, and hierarchical relationships) about specific categories, (2) **translate** new sets of categories from existing datasets and published studies, (3) **integrate** novel combinations of datasets for researchers' custom needs, including automatically generated syntax (e.g., R, Stata) to merge datasets of interest, and (4) **share** merging templates for public re-use and open science. Rather than storing observational data, CatMapper is based on a dynamic, interactive dictionary of keys that help users integrate observational data from diverse external datasets in disparate formats, thereby complementing and leveraging a fast-growing ecology of datasets storing observational data. CatMapper is designed to grow as registered users submit new categories, translations, and merges.

PROBLEM

To illustrate the key issues the platform aims to address, suppose a user would like to study how ethnic-based discrimination is associated with child development across a broad range of ethnicities worldwide. For measures of child development, the user aims to analyze individual-level data on infant survival and child growth from hundreds of demographic and health surveys which code individuals and households by ethnicity (Corsi, Neuman, Finlay, & Subramanian, 2012; UNICEF, 2015). For ethnic-based discrimination, the user would like to triangulate measures from ethnic-level datasets (EPR (Vogt et al., 2015), AMAR (Birnir et al., 2018)) and nationally-representative attitude surveys (Kamau, 2023). In this case the researcher may want a dataset with: (1) one row for each child from stacked demographic and health survey datasets with columns for needed variables (e.g., child growth, household

wealth, mother's education), (2) merged with aggregated variables from other datasets (Birnir et al., 2018; Kamau, 2023; Vogt et al., 2015) by a child's ethnicity.

Datasets	Category Name	Category Key
DHS Cote D'Ivoire 1994	Agni	v131 = 7
Afrobarometer 8	Akan	Q81 = 260
MICS Cote D'Ivoire 2015	Akan	HC1C = 11
Ethnic Power Relations	Other Akan	gwgroupid = 43704000

Table 1. Diverse encodings for similar categories across datasets

To stack and join data across more than a thousand ethnicities encoded in different ways across hundreds of datasets, the researcher must deal with several challenges. First, each dataset has its own system for encoding ethnicities using data-specific names and keys. For example, consider the case of "Agni" in Cote D'Ivoire illustrated in Table 1. Each dataset has unique ways of identifying an Agni respondent, sometimes with different names, sometimes lumped together with other groups (e.g., as "Akan" or "Other Akan") and always with a different key.

Each dataset also has a different temporal and spatial focus, and the researcher may want to exclude ethnic category matches that are too far apart either temporally or spatially. Thus, the researcher must make a number of choices about how to match these categories across datasets with a combination of one-to-one and many-to-one mappings. Ideally, the researcher also records and annotates these decisions in a systematic and transparent format for inspection and re-use by other researchers. Recall that this is only one of over a thousand ethnic categories that the researchers must manage. Once the researcher has made these decisions, they must still write and debug code to merge the multiple variables from these diverse datasets into a single dataset for their analysis. Manually, all of these tasks are time-consuming and create many opportunities for human error.

TOOLS TO ADDRESS PROBLEM

CatMapper aims to improve the efficiency, accuracy and transparency of the key steps in the reconciliation and merging process by (1) automating tasks when possible, and eliciting (and documenting) user input when ambiguities arise, (2) maintaining a well-documented and expandable repository of categories and translations so users can build from prior work rather than duplicate effort, and (3) documenting user decisions in a common machine readable form for easy inspection and re-use by future users. CatMapper does this with four sets of tools aimed at (1) exploring contextual information about a specific category, (2) translating new classification schemes to existing ones stored in the platform, (3) integrating data from multiple external datasets by categories, and (4) documenting and sharing researcher decisions when integrating data for their specific study. The ultimate product of a project is a "merging template" which encapsulates the key decisions needed to replicate a merge across multiple datasets (Figure 1). CatMapper provides these functions via a javascript interface, neo4j database, and python APIs. Future users interested in using or modifying that merge for other analyses can access key components of the merging template in several ways, including link files and R syntax to replicate the merge on a personal computer where external datasets are also stored.

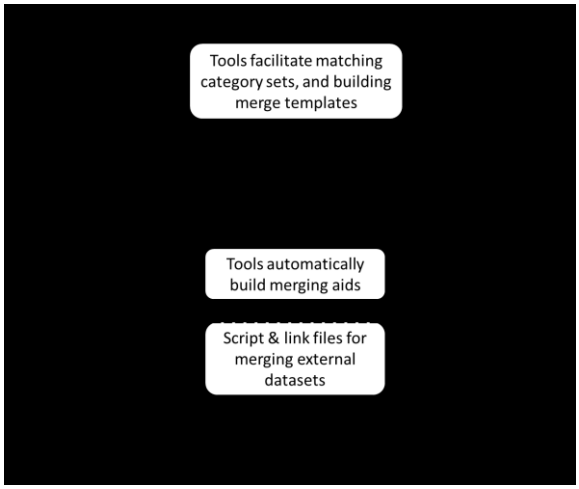

Figure 1. CatMapper database structures and functions

PROGRESS

The platform currently stores contextual information on over 14,000 ethnic categories and over 200,000 subnational regions, including provenance metadata on the sources of this information. In addition, the platform contains keys to how over 2400 datasets encode these sociopolitical categories, as well as functions to assist researchers in bringing together data on ethnicities, languages, religions and districts across diverse datasets.

Even though the platform's beta version has only recently come online, it has already assisted a number of research and infrastructure projects. It has helped researchers and students in synthesizing data for three dissertations, four undergraduate theses, and one published paper. Teams with federally-funded projects at Center for Archaeology and Society and cyberSW have used the platform, to organize sites and artifacts in their repositories and datasets. Since

January 2022, the platform's beta version has attracted 3,900 unique users. And the platform now sends links to over 45,000 publicly facing urls at other sources, including Wikidata (Vrandečić & Krötzsch, 2014), Wikipedia, eHRAF (Ember & Fischer, 2017), Database of Religious History (Slingerland, Monroe, & Muthukrishna, 2023), DPLACE (Kirby et al., 2016), and Glottolog (Hammarström, Forkel, & Haspelmath, 2018).

CONCLUSION

We introduce a user-friendly set of tools to help users translate four kinds of categories (religion, ethnicity, language, and subdistrict) across multiple, external datasets. The platform continues to grow with user input, but it currently contains keys to how over 2400 datasets encode over 14,000 ethnic categories and over 200,000 subnational regions, including provenance metadata on the sources of this information. CatMapper's tools for creating and sharing merges and its expanding repository of linkages across datasets enables users to bring together data across diverse datasets to support novel analyses. In future work, we will be developing cases studies that illustrate using the platform to build datasets for new cross-cultural and comparative analyses.

GENERATIVE AI USE

We confirm that we did not use generative AI tools/services to author this submission.

AUTHOR ATTRIBUTION

DJH: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – original draft, and writing – review and editing. YYC: conceptualization, and writing – review and editing. IHH: conceptualization, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing – review and editing. RB: conceptualization, data curation, formal analysis, investigation, methodology, resources, software, validation, visualization, writing – review and editing. HK: software, validation, visualization, writing – review and editing. CH: data curation, investigation, validation, visualization, writing – review and editing.

ACKNOWLEDGMENTS

We gratefully acknowledge the grant from NSF (BCS-2318505, BCS- 2051369).

REFERENCES

Birnir, J. K., Laitin, D. D., Wilkenfeld, J., Waguespack, D. M., Hultquist, A. S., & Gurr, T. R. (2018). Introducing the AMAR (All Minorities at Risk) Data. *Journal of Conflict Resolution, 62*(1), 203-226.

Cheng, Y. Y., & Ludäscher, B. (2020a). Reconciling taxonomies of electoral constituencies and recognized tribes of indigenous Taiwan. *Proceedings of the Association for Information Science and Technology, 57*(1), e248.

Cheng, Y. Y., & Ludäscher, B. (2020b). Through the magnifying glass: Exploring aggregations of COVID-19 datasets by county, state, and taxonomies of US regions. *Proceedings of the Association for Information Science and Technology, 57*(1), e355.

Corsi, D. J., Neuman, M., Finlay, J. E., & Subramanian, S. (2012). Demographic and health surveys: a profile. *International journal of epidemiology, 41*(6), 1602-1613.

Ember, C., & Fischer, M. D. (2017). Using eHRAF World Cultures with other cross-cultural samples.

Faniel, I. M., Kriesberg, A., & Yakel, E. (2012). Data reuse and sensemaking among novice social scientists. *Proceedings of the American Society for Information Science and Technology, 49*(1), 1-10.

Hammarström, H., Forkel, R., & Haspelmath, M. (2018). Glottolog 3.4. In: Jena: Max Planck Institute for the Science of Human History. Available ….

Hruschka, D., Bischoff, R., Peeples, M., Hsiao, I.-H., & Sarwat, M. (2022). CatMapper: A user-friendly tool for integrating data across complex categories. *Socarxiv, n6rty*.

Kamau, P. (2023). Social Cohesion, Politics and Governance in East Africa: Evidence from Afrobarometer Surveys. In *State Politics and Public Policy in Eastern Africa: A Comparative Perspective* (pp. 231-253): Springer.

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., . . . Ember, C. R. (2016). D-PLACE: A global database of cultural, linguistic and environmental diversity. *PloS one, 11*(7), e0158391.

Sakai, Y., Miyata, Y., Yokoi, K., Wang, Y., & Kurata, K. (2020). Data integration as the major mode of data reuse. *Proceedings of the Association for Information Science and Technology, 57*(1), e391.

Slingerland, E., Monroe, M. W., & Muthukrishna, M. (2023). The database of religious history (drh): Ontology, coding strategies and the future of cultural evolutionary analyses. *Religion, Brain & Behavior*, 1-30.

UNICEF, G. (2015). Multiple indicator cluster survey (MICS).

Vogt, M., Bormann, N.-C., Rüegger, S., Cederman, L.-E., Hunziker, P., & Girardin, L. (2015). Integrating data on ethnicity, geography, and conflict: The ethnic power relations data set family. *Journal of Conflict Resolution, 59*(7), 1327-1342.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledge base.