Question-Generating Datasets: Facilitating Data Transformation of Official Statistics for Broad Citizenry Decision-Making

Rahul Yadav[1], Patricia Snell Herzog[2], Davide Bolchini[1]
[1] Indiana University School of Informatics and Computing at IUPUI
[2] Indiana University Lilly Family School of Philanthropy at IUPUI
{rayada, psherzog, dbolchin}@iupui.edu

Abstract

Citizenry decision-making relies on data for informed actions, and official statistics provide many of the relevant data needed for these decisions. However, the wide, distributed, and diverse datasets available from official statistics remain hard to access, scrutinise and manipulate, especially for non-experts. As a result, the complexities involved in official statistical databases create barriers to broader access to these data, often rendering the data non-actionable or irrelevant for the speed at which decisions are made in social and public life. To address this problem, this paper proposes an approach to automatically generating basic, factual questions from an existing dataset of official statistics. The question generating process, now specifically instantiated for geospatial data, starts from a raw dataset and gradually builds toward formulating and presenting users with examples of questions that the dataset can answer, and for which geographic units. This approach exemplifies a novel paradigm of question-first data rendering, where questions, rather than data tables, are used as a human-centred and relevant access points to explore, manipulate, navigate and cross-link data to support decision making. This approach can automate time-consuming aspects of data transformation and facilitate broader access to data.

Keywords: *Official Statistics; Geospatial Data; Big Data Methods and Automation; Data Economy; Data Access; Data-Based Decision-Making*

1. Introduction

This paper describes an automation process designed to generate questions from datasets. Questions are key to problem solving, and data-based problem solving is crucial for making informed decisions (e.g. Boss, 2016). Sometimes the power of data is harnessed when a new analysis is run, but often the power of transformation occurs in framing the question. Yet, answers to an unasked question are irrelevant (Gutiérrez, 2013, citing Niebuhr, 1943).

The big data in large administrative datasets can help to answer public policy questions. (e.g. Connelly et al., 2016; Chetty et al., 2014; Chetty, 2013). Harnessing their insights requires understanding the distinctions of the "wide data" arising from few observations relative to variables, often garnered from Internet sources such as Amazon clicks and Twitter likes, from the "long data" with many observations relative to variables, often garnered from administrative databases of official statistics such as tax records and censuses (Chetty 2020). Additionally, asking relevant questions of the big data generated from official statistics has challenges (Connelly et al., 2016; Taylor et al., 2014). Notably, users often need domain-specific knowledge to understand the structure and semantics of these datasets and be able to pose questions their data can answer (Europa, 2017).

Advances in natural query language technologies can automate facets of the question-generating process. For example, Salesforce is using machine learning to facilitate everyday people in querying databases using their natural language (Mannes, 2017). Automating portions of the data access process can aid broader utility of the information embedded in these big datasets for public issues. Despite the many benefits of automating the process of database querying, there are several issues that could impede its development. For one, the domain knowledge embedded within large datasets needs to be treated carefully in the automation construction process in order to facilitate broader accessibility and applicability.

To lessen barriers to data access, this project aims to extract user-relevant meaning from datasets in forms of basic factual questions by automating some of the facets of the data extraction and linking processes. Since geographies are particularly relevant for public policy questions, this project focuses on automating extraction of the geospatial data that are pervasive in datasets of official statistics.

2. Exemplifying Scenario

The following scenario illustrates data access issues surrounding use of official statistics in citizenry decision-making. Residents in a city that lacks funding for adequate public transportation want a new transit line added to alleviate traffic congestion and promote

environmental sustainability. Due to the limited infrastructure, there are many opportunities for locations within the city for this new line. Which location is of highest priority?

While transportation services aim to address multiple priorities in the long-term, the limited resources necessitate short-term prioritising of a single new line in a strategic location. In reviewing previous efforts, city officials note that past transit efforts were severely under-funded because many voting citizens declined ballot budget referendums. The resulting lack of funding caused bus delays, driver turnover, and other infrastructure issues. Thus, the goal for the strategic location of the new line is to avoid repeating these previous problems.

To be strategic in selecting the location, city officials decide to target a higher-income area, aiming for wealthier residents to then vote in support of additional funding. The location thus needs to target a location that can facilitate greater commuting to work and cultural activities among high-income residents. To make a data-informed decision about the location that best targets this goal, decision-makers need to know four data points within a set of concurrent geographic units: median household income (higher than city average), median work commute time (higher than within-city average), number of existing public transit lines (lower than city average), and fuel consumption (higher than city average).

In this scenario, the metrics need to be available within-city, at relatively small geographic units that can be mapped along the main roadways. To assess whether the new line is effective, these geospatial data also are needed over time, before and after the new line. Most importantly, decision makers need to be able to access and link relevant data rapidly. While relevant data exist, the wrangling required to extract and merge it is a major barrier.

3. Related Work

Several existing approaches are relevant for these issues. For example, scientists have established that geospatial metadata can be useful for assessing the utility of spatio-temporal data for decision makers (Meeks & Dasgupta, 2004). However, the last decade of attention to how geospatial data contribute to broad citizenry decision-making has centred heavily on the spike in availability of geographic information system (GIS) data generated from user devices (e.g. Bishr & Kuhn, 2007). Indeed, major breakthroughs have occurred in humanitarian assistance as a result of the disaster data that can be rapidly spatially located from widespread use of hand-held devices (Ortiz, 2020). These are important advances.

Yet, advancements from the administrative databases of official statistics have not kept pace. For example, population demographic characteristics, such as median household income levels, remain aggregated within geographic units (GEOIDs: U.S. Census 2018; ANSI: U.S. Census 2019). In these, tracts are a geographic unit that was developed to meaningfully approximate neighbourhoods. Tracts are smaller than metropolitan areas and

counties, but larger than city blocks or block groups, and thus tract-level data provide aggregate statistics at within-city geographic units. In the scenario, tract-level data would help to answer questions regarding the most strategic location for the new transit line. One approach to addressing barriers to broader access to relevant official statistics is to visually represent data through an interactive map (e.g. Cartwright et al., 2013). However, not all relevant data are available within existing maps. For example, Policy Map or Social Explorer are widely used map tools, and SAVI is another tool, which attends to a particular city. In all three of these interactive maps, median household income and work commute time are available from official statistics in the U.S. Census. Yet, number of existing transit routes is only available within SAVI, and average fuel consumption is not available in any.

The U.S. Census uses Federal Information Processing Standards codes (FIPS) to uniquely identify aggregated geospatial data, for example through states, counties, and tracts. In some datasets, FIPS identifiers may be stored as a single string digit, with all three geographic identifiers appended in a single variable. Whereas, in other datasets each geographic unit is stored separately, for example in three variables, each representing state, county, and tract in turn. Moreover, many datasets do not specify the available geographies within the meta-data, requiring domain expertise to discern what geographies are available. The nuances in how aggregated geo-spatial data is identified present challenges for non-experts to access relevant data, and to know at which geo-unit data can be wrangled. While the exemplifying scenario in this paper focuses on U.S. geocodes, problems with non-standardised units also exist at larger geographies, such as nations (Scott & Rajabifard, 2017). Thus, part of facilitating a more accessible data economy of official statistics (Europa, 2017) is to automate the geospatial data transformation process, and thus more readily generate the kinds of questions that can be answered by linked datasets.

4. Question-Generating Datasets

Figure 1 displays the process involved in automatically equipping geospatial datasets with basic questions they are designed to answer. This includes the following six steps.

4.1. Data Sources

To begin an analysis of datasets, we took into consideration different sources of datasets and how they might be stored. Typically based on the type of datasets and its usage, datasets are stored in relational or non-relational databases. For our analysis we had the following requirements for the dataset format: (1) It should be compatible across different platforms; and (2) It should be widely used across industries. Common file formats matched our requirements (Shafranovich, 2005): the dataset is in common tabular format (CSV) and any information available about the dataset is in plain text format (TXT). The data sources that we leveraged were open source datasets available from U.S. Census Data,

American Community Survey, Our World in Data, and the Indiana Data Hub, all of which have geographic identifiers available within the dataset.

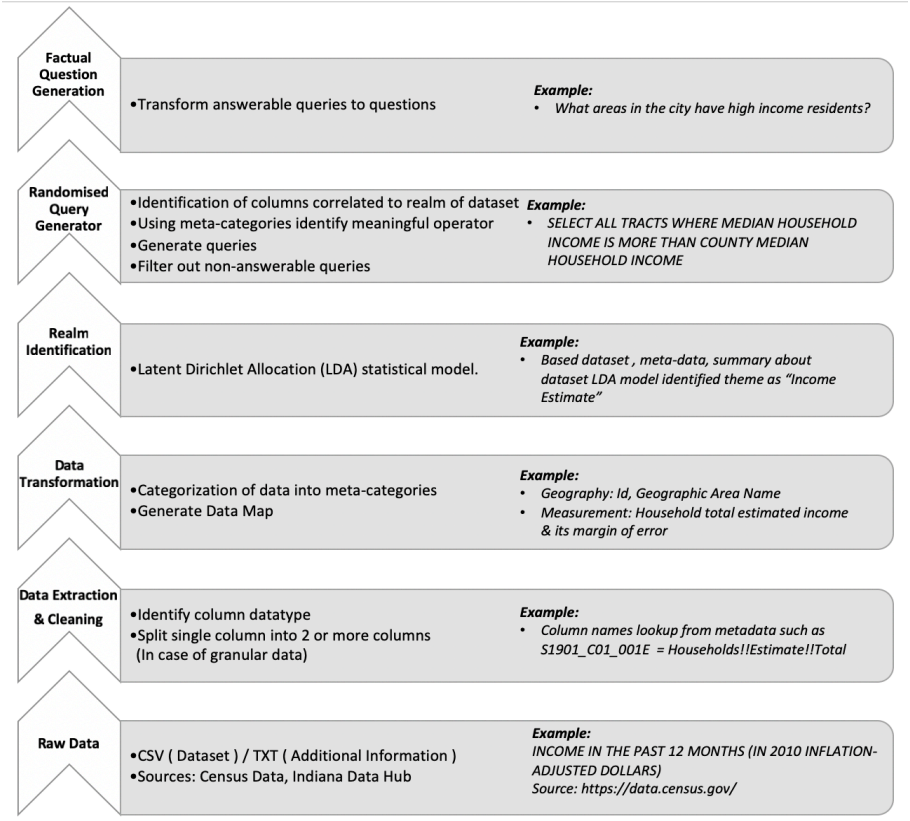

Figure 1. Data transformation process. Source: Author creation.

4.2. Data Extraction and Cleaning

In the process of data extraction (Rahm & Do, 2000), we encountered three issues. First, CSV files do not retain column datatypes. Second, column names often follow patterns that require domain-specific expertise to identify. Third, columns with granular information can be leveraged to gather broader categories. For example, a column with tract FIPS can also contain county and state FIPS within it. Thus, the data extraction process identifies column datatypes and splits granular data into separate variables for state, county, and tract FIPS.

4.3. Data Transformation

Once data are extracted from CSV file, we developed the meta-categories in Table 1, under which most datasets can be represented. We developed a parser (e.g. Srivastava et al.,

2017), which leverages column type and segregates datasets into meta-categories based on patterns and data lookup (e.g. FIPS tables). Once dataset is parsed, the process automatically generates a data dictionary representing the entire dataset in meta-categories.

Table 1. Meta-categories of datasets.

Meta-category	Examples
Time	Date, Period, Duration
Measurement Volume	Length, Weight, Height
Geography	FIPS Code, Latitude, Longitude, Address
Property	Race, Gender, Education

Source: Author creation.

4.4. Identification of Realm

Since we are aiming to generate meaningful answerable questions from the dataset, realm identification is the most important step. With needed meta-data, such as, data summary, data dictionary, and variable description, we were able to correctly identify dataset realms. Realm was processed using a Latent Dirichlet Allocation (LDA) statistical model (Blei et al., 2003) and existing Natural Language Toolkit (NLTK) resources.

4.5. Randomised Query Generator

The next step is to identify columns which are highly correlated to the realm of dataset. Once we select columns which are central to the theme of the data, in conjunction with other randomly selected columns, we leverage meta-category details about these columns to identify meaningful operators for each. Using selected columns and meaningful operators, we generate queries. Once queries are generated, there is a probability that a query might not be answerable by the dataset. Such queries are removed after all queries are generated.

4.6. Factual Question Generation

To render the generated queries as questions, we parsed the queries using semantics (de Marneffe et al., 2006), expressed in Standard Query Language (SQL) into three simple components: Command (action to be taken), Target (specific table) and Additional Clauses (restrictions on the data selection). We replaced the command with question words such as "what" or "which", followed by additional clauses. Examples include:

- Query 1: SELECT RECORDS FROM TABLE WHERE Tract-Level Median Household Income > County-Level Median HH Inc. Q1: ***Which neighbourhoods have high income levels?***
- Query 2: SELECT RECORDS FROM TABLE WHERE Tract-Level # of Public Transit Lines < Median Cty.-Level # of Lines. Q2: ***Which neighbourhoods have low public transit?***
- Query 3: SELECT RECORDS FROM TABLE WHERE Tract-Level Work Commute Time >= Cty.-Level Work Commute Time. Q3: ***Which neighbourhoods have avg.+ work commutes?***
- Query 4: SELECT RECORDS FROM TABLE WHERE Tract-Level Fuel Consumption >= Cty.-Level Fuel Consumption. Q4**:** ***Which neighbourhoods have average+ fuel consumption?***

In this way, datasets can generate examples of the kinds of questions they are equipped to answer, and this automation process can identify those questions in typical semantic terms.

5. Discussion

Traditionally, to interact with datasets, the strategies available to users included: (1) formulating issue specific queries on the data; (2) manipulating and browsing bi-dimensional tables; (3) exploring data through maps and visualisations. Existing paradigms, however, embed a key limitation: they assume that users know or eventually find out which questions to ask the data. However, knowledge workers, who are often not data scientists or domain experts, only vaguely sense the potential value a dataset can hold. As a result, many public datasets are under-utilised. Formulating queries can aid broader access and use.

In this study, we made the first steps towards expanding our understanding of human-data interaction from a data-first to a question-first paradigm. The results of our work exemplify how question-generating datasets can prompt users with examples of potentially relevant questions the dataset can answer. This work sheds light on a potential new class of data-intensive interactive systems, one that endows an available dataset with a suite of available factual questions as the starting point for relevant searches or data exploration.

More broadly, this approach to automating question generation from existing datasets can play an important role in broadening the accessibility and usability of official statistics. The speed at which decisions are made in public and social issues requires that datasets be transformed to aid semantically usable identification of the kind of information a dataset can contribute. In this example, city officials and concerned citizens could more rapidly identify which official statistical databases are equipped to answer their questions, and as a result would be better able to make data-informed decisions regarding where to locate a new public transit line, and its effectiveness in meeting targeted goals over time, for example. Future studies can build upon the automation steps advanced here to improve access for other kinds of data, beyond the geospatial data of this study. Such advancements would improve widespread data access, increase data-based decision-making for public and social issues, and facilitate informed decisions within the rapid durations necessary.

References

American National Standards Institute (ANSI). (2019, April 23). U.S. Census Bureau. https://www.census.gov/library/reference/code-lists/ansi.html

Bishr, M., & Kuhn, W. (2007). Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In S. I. Fabrikant & M. Wachowicz (Eds.), *The European Information Society: Leading the Way with Geo-information* (pp. 365–387). Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Boss, J. (2016, August 3). The Power of Questions. *Forbes*. Retrieved from:

https://www.forbes.com/sites/jeffboss/2016/08/03/the-power-of-questions/

Chetty, R. (2013, October 20). Yes, Economics Is a Science. *The New York Times.* https://www.nytimes.com/2013/10/21/opinion/yes-economics-is-a-science.html

Chetty, R. (2020). *Using Big Data to Solve Economic and Social Problems: The Geography of Upward Mobility in America*. Harvard University, Opportunity Insights, Cambridge, MA. Retrieved from: https://opportunityinsights.org/course/

Chetty, R., Hendren, N., Kline, P., Saez, E., & Turner, N. (2014). Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility. *American Economic Review*, 104(5), 141–147.

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, 59, 1–12.

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006, May). Generating Typed Dependency Parses from Phrase Structure Parses. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). LREC 2006, Genoa.

Europa. (2017). Building a European Data Economy. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Retrieved from: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=41205

Cartwright, W., Crampton, J., Gartner, G., Miller, S., Mitchell, K., Siekierska, E., & Wood, J. (2013). Geospatial Information Visualization User Interface Issues. *Cartography and Geographic Information Science*, 28(1), 45–60.

Mannes, J. (2017, August 29). Salesforce is using AI to democratize SQL so anyone can query databases in natural language. *TechCrunch*. Retrieved from: http://social.techcrunch.com/2017/08/29/salesforce-is-using-ai-to-democratize-sql-so-anyone-can-query-databases-in-natural-language/

Meeks, W. L., & Dasgupta, S. (2004). Geospatial information utility: An estimation of the relevance of geospatial information to users. *Decision Support Systems*, 38(1), 47–63.

Niebuhr, R. (1964). Nature and Destiny of Man, vol. II: Human Destiny (1st Paperback Edition edition). Charles Scribner.

Ortiz, D. (2020). Geographic Information Systems (GIS) in Humanitarian Assistance: A Meta-Analysis. *Pathways: A Journal of Humanistic and Social Inquiry*, 1(2).

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin, 23.

Scott, G., & Rajabifard, A. (2017). Sustainable development and geospatial information: A strategic framework for integrating a global policy agenda into national geospatial capabilities. *Geo-Spatial Information Science*, 20(2), 59–76.

Shafranovich, Y. (2005). Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC, 4180, 1-8.

Srivastava, S., Labutov, I., & Mitchell, T. (2017). Joint concept learning and semantic parsing from natural language explanations. Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, 1527–1536.

Taylor, L., Schroeder, R., & Meyer, E. (2014). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? Big Data & Society, 1(2), 1-10.

Understanding Geographic Identifiers (GEOIDs). (2018, October 10). U.S. Census Bureau. https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html