STRUCTURED OPEN URBAN DATA:

Understanding the Landscape

Luciano Barbosa,[1] Kien Pham,[2] Claudio Silva,[2,3] Marcos R. Vieira,[1] and Juliana Freire[2,3]

Abstract

A growing number of cities are now making urban data freely available to the public. Besides promoting transparency, these data can have a transformative effect in social science research as well as in how citizens participate in governance. These initiatives, however, are fairly recent and the landscape of open urban data is not well known. In this study, we try to shed some light on this through a detailed study of over 9,000 open data sets from 20 cities in North America. We start by presenting general statistics about the content, size, nature, and popularity of the different data sets, and then examine in more detail structured data sets that contain tabular data. Since a key benefit of having a large number of data sets available is the ability to fuse information, we investigate opportunities for data integration. We also study data quality issues and time-related aspects, namely, recency and change frequency. Our findings are encouraging in that most of the data are structured and published in standard formats that are easy to parse; there is ample opportunity to integrate different data sets; and the volume of data is increasing steadily. But they also uncovered a number of challenges that need to be addressed to enable these data to be fully leveraged. We discuss both our findings and issues involved in using open urban data.

Introduction

For the first time in history, more than half of the world's population lives in urban areas[1]; in a few decades, the world's population will exceed 9 billion, 70% of whom will live in cities. The exploration of urban data will be essential to inform both policy and administration, and enable cities to deliver services effectively, efficiently, and sustainably while keeping their citizens safe, healthy, prosperous, and well-informed.[2–4]

While in the past, policymakers and scientists faced significant constraints in obtaining the data needed to evaluate their policies and practices, recently there has been an explosion in the volume of open data. In an effort to promote transparency, many cities in the United States and around the world are publishing data collected by their governments (see, e.g., refs.[5–8]).

Having these data available creates many new opportunities. In particular, while individual data sets are valuable, by integrating data from multiple sources, the integrated data are often more valuable than the sum of their parts. The benefits of integrating city data have already led to many success stories. In New York City (NYC), by combining data from multiple agencies and using predictive analytics, the city increased the rate of detecting dangerous buildings, as well as improved the return on the time of building inspectors looking for illegal apartments.[2] Policy changes have also been triggered by studies that, for example, showed correlations

[1]*IBM Research, Rio de Janiero, Brazil.*
[2]*Department of Computer Science and Engineering, NYU School of Engineering, Brooklyn, New York.*
[3]*NYU Center for Urban Science and Progress, Brooklyn, New York.*

between foreclosures and increase in crime,[9] and the effects of subsidized housing on surrounding neighborhoods.[10]

Open urban data also create new opportunities for the creation of apps that leverage these data. The OneBusAway application[11] provides real-time arrival predictions, and other transit information has been successfully deployed in over 10 cities worldwide. It was shown to not only significantly decrease wait time (real and perceived), but also increase transit usage for noncommute trips, and, remarkably, increase walking—higher confidence in arrival times allowed riders to walk to further stops. Mapumental[12] uses real-time bus arrival times and disruptions in the subway services, provided by the city of London, to calculate the average travel time from any region in London to a given destination using public transportation.

With the growing volumes of open urban data, these success stories can be just the beginning. The challenge now lies in making sense of all the data so that they can be used effectively to answer the right questions in ways that can actually lead to policy improvements or more effective use of resources and infrastructure. In this article, we take a first step toward understanding the current landscape so that we can better assess the challenges and opportunities for finding, using, and integrating urban data. We collected over 9,000 data sets for 20 cities in North America and analyzed different aspects of these data, including their contents and size, how recent and dynamic they are, formats used, quality of the data, and opportunities for integrating disparate data sets.

Our findings are encouraging. Most of the data are available in tabular formats that can be easily parsed and processed, for example, CSV file format. Data sets cover a wide range of topics and new data are constantly being added; in 2013, more data sets were added than in the three previous years combined. However, the total volume of data—around 70GB for all cities—is rather small. As a point of comparison, consider, for example, data about taxi trips in NYC, which is not available as an open data set but that can be obtained from the Taxi & Limousine Commission: each year of data contains approximately 170 million trips and occupies 50GB on disk. Therefore, if these efforts to open data continue to expand, we are likely to see an explosion in the data volume.

An important finding of the study is that there is ample opportunity for integrating the different data sets. We found significant overlap in the schemata of tables. In addition, since most tables contain information about location, location-based joins can be performed to fuse the information across tables, and it is also possible to visually explore them as layers on a map. Our study also uncovered many challenges that need to be addressed for the data to be fully leveraged. While there is overlap across different schemata, there is substantial heterogeneity. For example, we observed occurrences of multiple terms to represent the same concept, as well as the use of a term to represent different concepts. This problem is compounded due to the lack of precise type of information. Thus, sophisticated data integration techniques are needed.[13–15]

"OUR STUDY ALSO UNCOVERED MANY CHALLENGES THAT NEED TO BE ADDRESSED FOR THE DATA TO BE FULLY LEVERAGED."

The remainder of the article is organized as follows: In the section Data Set Description, we describe the data set collected for this study. General statistics about the data are presented in the section Taking a Broad Look: General Statistics, and in the section Examining Tabular Data, we focus on tabular data and analyze structure-related features, including schema size, schema similarity, attribute types, and table sparseness. Finally, in the Discussion and Challenges section, we conclude with a summary of our findings and a discussion of challenges in using open urban data. The code, scripts and the description of the data used in this study are available at: https://github.com/ViDA-NYU/urban-data-study.

Data Set Description

In this study, we focus on data from cities in the United States and Canada. Data sets are available in different platforms, including CKAN[16] and Socrata.[17] We used data published by 20 cities in North America that have adopted Socrata as their publishing platform.[17] The data are diverse, containing information about cities as small as De Leon, Texas (population 2,233), and big cities such as Chicago and NYC (population 2.7 and 8.3 million, respectively). Table 1 gives an overview of the data for the different cities. Data sets are published in different formats: tables, maps, charts, calendars, forms, binary files, documents, and links to external data sets/web sites. In our analyses, we used only structured data sets and downloaded all data sets (in both CSV and JSON format) for all 20 cities in October 2013, totaling 71GB in size. Besides the actual data, Socrata makes available metadata that we use in the study, including category, textual description, date of creation, number of views and downloads, and keywords, and for structured (tabular) data, a list of attribute names.

Taking A Broad Look: General Statistics

How many data sets are available?

Table 1 shows the number of data sets available for each city. They range from 9 (Redmond) to 2,411 (NYC) data sets. Intuitively, a factor that might have some influence in the number of data sets for a city is the size of its population. To verify that, we computed the Spearman's rank correlation coefficient (ρ score) between the number of data sets and the population of each city. The ρ score is 0.81 ($p = 1.33 \times 10^{-5}$), which indicates

Table 1. Number of Data Sets and Top-Three Data Categories for Each City

S. No.	City name	No. of data sets	Top-three categories
1	Redmond, WA	9	N/A
2	De Leon, TX	12	Government
3	Wellington, FL	30	Government, business, personal
4	Salt Lake City, UT	39	Government
5	Madison, WI	58	Property, police, elections
6	New Orleans, LA	66	Geographic reference, administrative data, city assets
7	Honolulu, HI	66	Transportation, public safety, recreation
8	Weatherford, TX	71	Community services, finance & budget, development
9	Somerville, MA	81	311 call center, budget
10	Oakland, CA	98	Public safety, infrastructure, environmental
11	Austin, TX	216	Government, financial, public safety
12	Boston, MA	355	City services, health, public safety
13	Edmonton, AB	395	Demographics, transportation, facilities and structures
14	Raleigh, NC	503	Fiscal year 2014, fiscal year 2013, public safety
15	San Francisco, CA	682	Ethics, public health, geography
16	Baltimore, MD	843	Crime, geographic, financial
17	Chicago, IL	954	Economic development, administration & finance, transportation
18	Seattle, WA	1,044	Public safety, community, permitting
19	Kansas City, MO	1,132	Traffic, census, labor
20	New York City, NY	2,411	Social services, housing & development, city government

that there is a strong correlation between the number of data sets available for a city and its population. A scatter plot (in log scale) of population size versus number of data sets is shown in Figure 1. We also calculated the Spearman correlation coefficient between city per capita income and number of data sets. The resulting ρ score of 0.17 suggests that these variables are not correlated.

What is the nature of the data?

One of the principles of open data is to make them easy to process by a machine.[18] To verify whether this principle is followed, we inspected the formats in which data are published.

FIG. 1. Population size versus number of data sets.

The large majority of the data sets (75%) come in tabular format (see Fig. 2). Socrata allows tabular data to be downloaded in different formats, for example, CSV, RDF, and XML. Relatively fewer data sets are published as pdf (10.6%), zip (8.7%), and other formats (5.7%), for example, XML, KMZ, and XLSX.* While tabular data can be easily processed by applications, pdf and zip files are more challenging to deal with.

As Figure 3 shows, the proportion of tables is not uniform across cities. Whereas Weatherford, Somerville, Madison, Seattle, and Wellington have a high proportion of tables (100%, 98.7%, 98.2%, 93.7%, and 93.3%, respectively), Kansas City and Boston are the least "friendly" cities for automated data processing—they have the smallest proportions of tables, 35% and 29%, respectively.

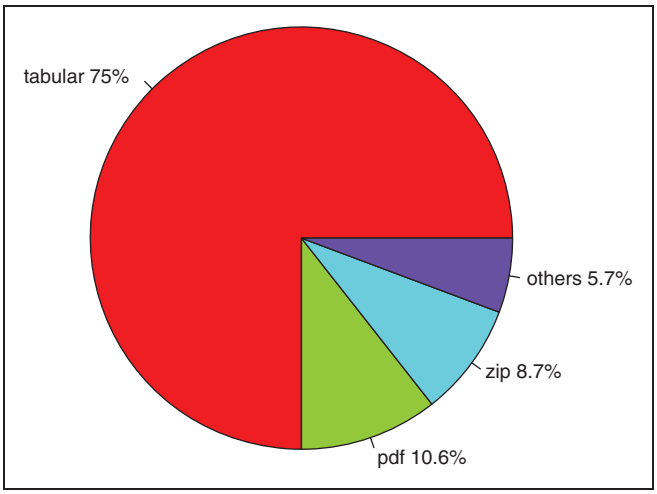

FIG. 2. Proportion of format types.

*XML, KMZ, and XLSX can encode data with complex structure, and they are not classified as tabular data in the Socrata API.

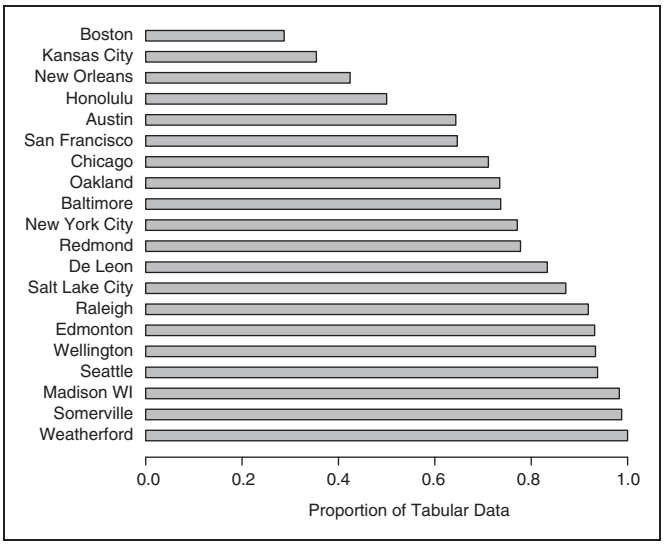

FIG. 3. Proportion of data in tabular format.

How big are the tables?

Table 2 shows the distribution of table sizes with respect to the number of records. Most tables are small—more than 60% of tables have less than 1,000 rows. Only a very small proportion of them (0.3%) have more than 1 million rows. We inspected the content of some of the small tables and found that they usually contain aggregated statistics. For instance, the NYC table "d4uz-6jaw"[19] has 10 rows with the number of inmates arrested by year in NYC from 2001 to 2010. The biggest table in the collection is the Chicago Traffic Tracker table with 6.7 million rows, which reports historical estimated congestion.

What are the data about?

The data sets cover many different topics and categories. To better understand what is covered, in Figure 4a we present a tag cloud containing keywords in the metadata associated to the data sets. Examples of high-frequency topics include service requests, crime, and traffic. The distribution of topics, however, is not the same for all cities. To illustrate this, we show in Figure 4b–e tag clouds for four different cities—NYC, Kansas City, Seattle, and Chicago—which have very different profiles. Tables related to 311[†] and service requests are very frequent in NYC; in Kansas City, tables related to the Land Development Division and traffic are dominant; Seattle has a large number of tables associated with police and crime; and for Chicago, many tables are related to sustainability.

Table 2. Table Size Distribution

No. of records	*Percentage of total*
0–1K	65.3
1K–10K	17.0
10K–100K	11.7
100K–1M	5.5
1M–10M	0.3

How popular are the data sets?

In the metadata associated with each data set, there are two statistics that are useful to assess their popularity: number of views and downloads. In Figure 5, we present the distribution of the number of unique views and downloads for tables since they were created. Tables seem to be visited fairly often. Almost 43% of them were visited more than 100 times since their creation. The most visited table, with more than 250,000 visits, contains a list of severe weather alert systems throughout Missouri provided by Kansas City.

One interesting fact is that the number of table downloads is much smaller than the number of views. Almost 87% of tables were downloaded less than 100 times. Seattle's 911 dispatches, with 438,000 downloads, is the table with the highest number of downloads. These numbers suggest that there is interest in these data (large number of views), but the data sets are still not widely used by third-party applications (small number of downloads).

"IN THE METADATA ASSOCIATED WITH EACH DATA SET, THERE ARE TWO STATISTICS THAT ARE USEFUL TO ASSESS THEIR POPULARITY: NUMBER OF VIEWS AND DOWNLOADS."

In an attempt to understand what brings more attention to these data sets, we generated tag clouds for data sets that have a large number of downloads. Figure 6a–c shows the tag clouds for data sets that have download counts greater than 100, 500, and 1,000, respectively. All cities have data sets that have been downloaded at least 100 times, but only half of the cities have data sets that were downloaded more than 1,000. The keywords "Geographical Information System" and "shape files" are the most common tags in all three sets. This suggests that people are interested in data sets that contain location information.

Note that a large number of views and downloads for a data set is also related to its age—older data sets are likely to have accumulated more accesses than new ones. Furthermore, they can also be the result of programmatic access by applications.

[†]311 is a popular service that allows city residents to submit requests about nonemergency issues.

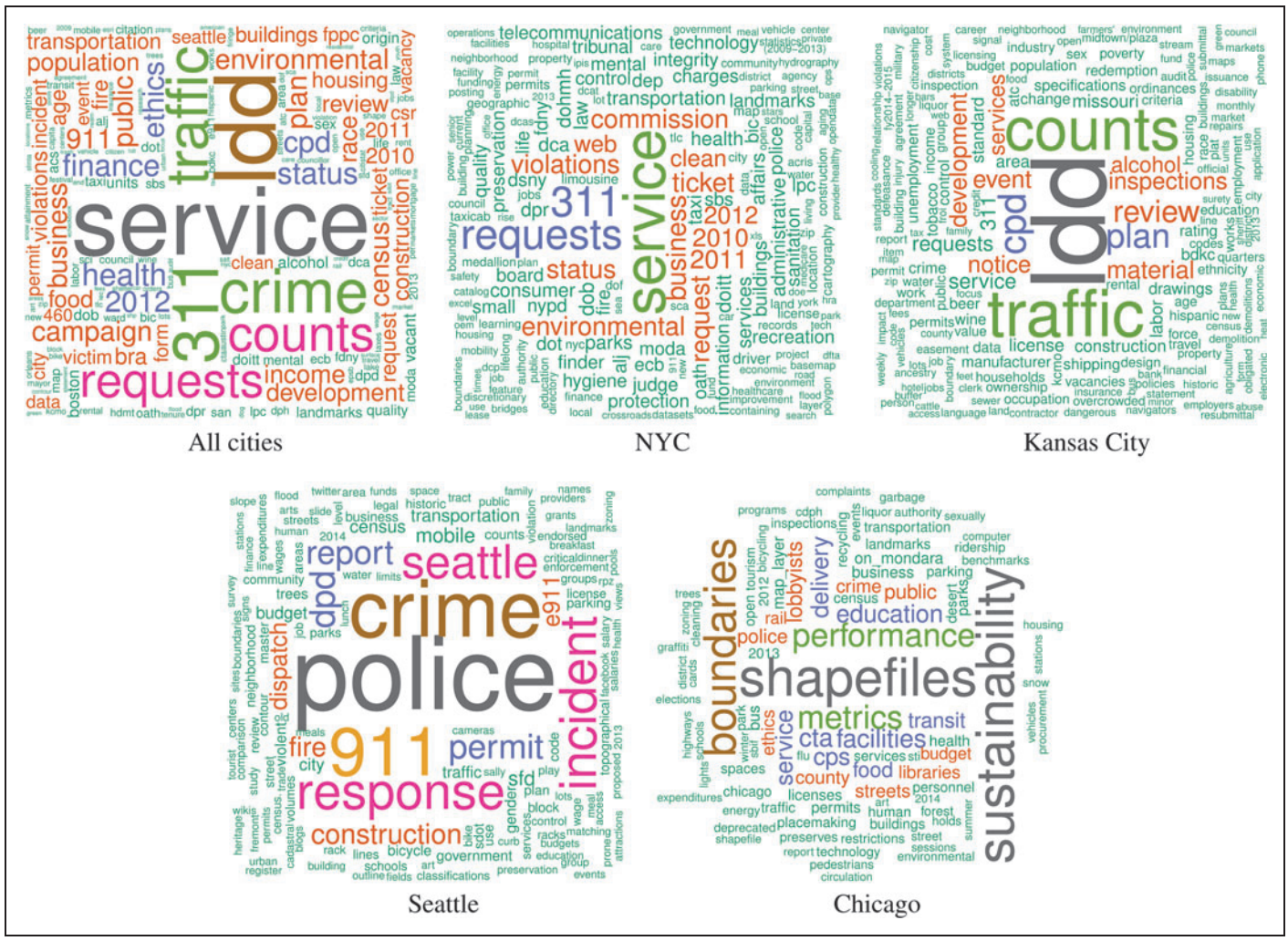

FIG. 4. Tag clouds derived from the keywords associated with the data sets.

How old are the data sets?

Table metadata includes the date of their creation. Using this information, we plotted in Figure 7 the distribution of the age of these tables in months from the day we obtained these numbers (10/30/2013). A table with age zero means that it was created in October 2013. Note that the distribution is skewed toward more recent tables (small ages). The trend line in the plot confirms this. In fact, most of the tables are 1 year old or younger. This shows that cities are increasingly making more data sets available. The oldest table is "Seattle Crime Stats by 1990 Census Tract 1996–2007," created in November 23, 2009 (about 48 months old).

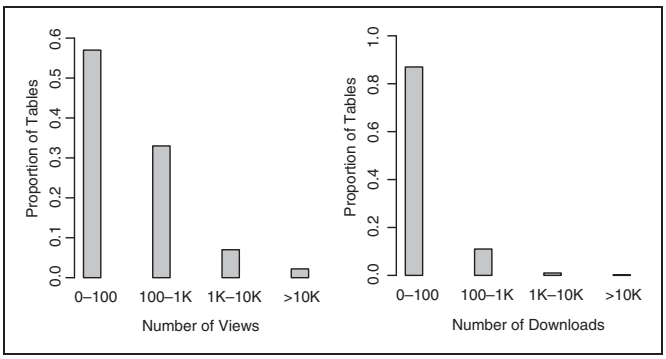

FIG. 5. Distribution of data set views and downloads.

The metadata also includes the last-modified date for each table. We monitored this information daily for all tables during 30 days (from October 1–30, 2013) and computed the change frequency ratio, that is, how many times a table changed in this period. The results are shown in Figure 8. A ratio of 1 means that a table was modified every day, and 0 means that it was never modified. The great majority of tables (71%) were never modified. Only 2.5% of the tables changed daily. An example of a highly dynamic table is the 311 data from Kansas City. We also found data sets whose descriptions indicate that they are updated daily but, in

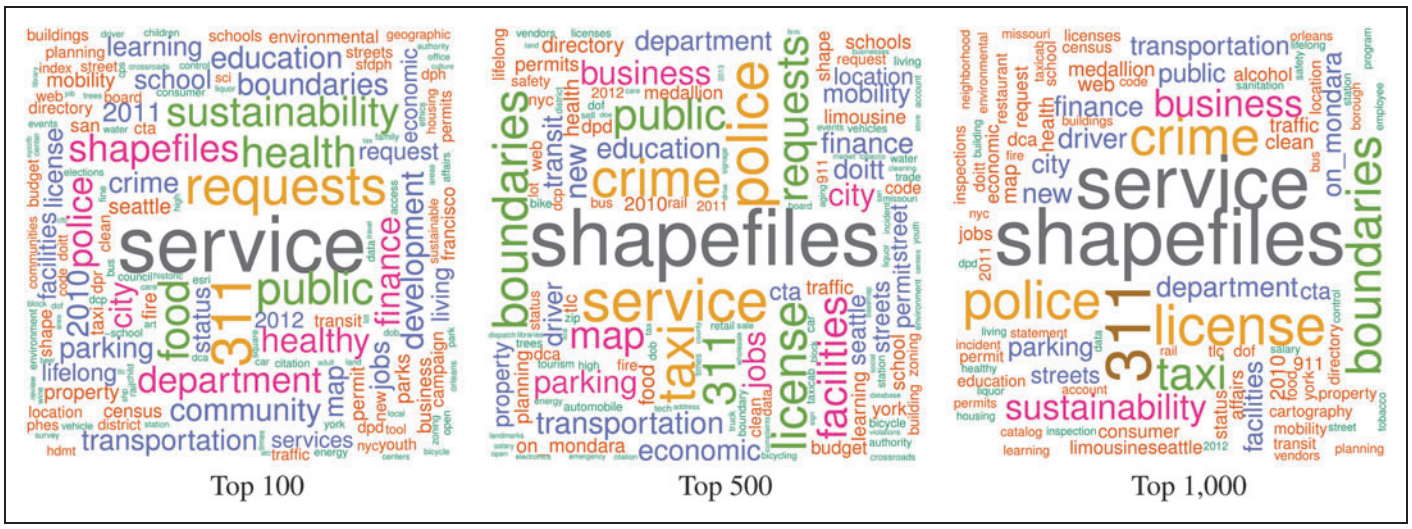

FIG. 6. Tag clouds derived from the keywords associated with the most popular data sets, that is, data sets with the largest number of downloads.

reality, are not. One example is the 311 data from NYC, whose change rate is 0.78.

Examining Tabular Data

We now focus on tabular data and analyze aspects that are important from a data integration and management perspective. Besides characteristics of the schemata (e.g., size and types of attributes), we also explore data quality issues and heterogeneity across data sets.

How big are the schemata?

Figure 9 presents the distribution of schema sizes for all tables. The numbers show that most of the tables have a small schema; more than 80% of them have schemata with fewer than 20 attributes. The proportion of tables decreases as the number of attributes increases. The table with the biggest schema, with 299 attributes, was the Internet and Global Citizens table[20] from Austin. This table has answers to questions for a "study administered through the City's Office of Telecommunications and Regulatory Affairs (TARA) to better understand community technological needs and desires."*

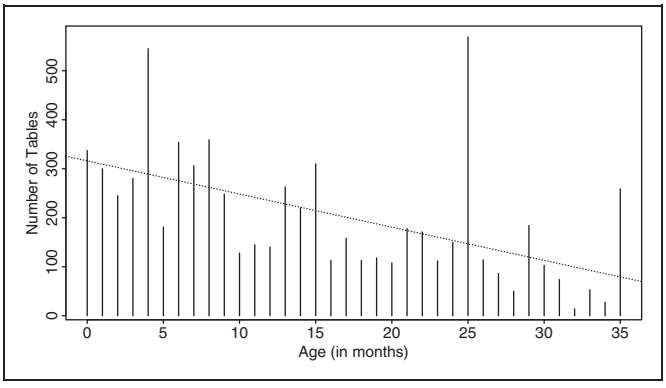

FIG. 7. Distribution of tables' age in months. The inclined horizontal line is the trend line for this distribution.

How similar are the table schemata?

A benefit of having a large number of open data sets available is the ability to integrate them and derive value-added information. Thus, an important question is whether there are opportunities for joining different tables. Linguistic matching based on attribute name is one of the most common techniques used for matching table schemata.[21] If two schemata have similar attribute names, they are likely to match. Thus, to estimate the potential to integrate different data sets, we

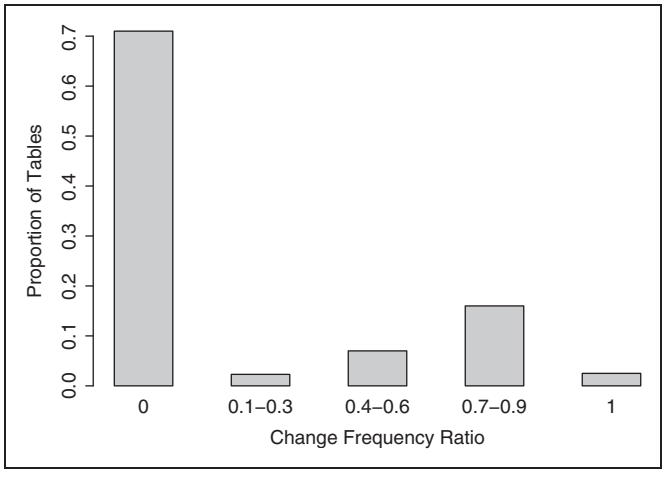

FIG. 8. Change frequency ratio of tables over 30 days.

*https://data.austintexas.gov/dataset/The-Austin-Internet-and-Global-Citizens-Project/gt3n-akq9

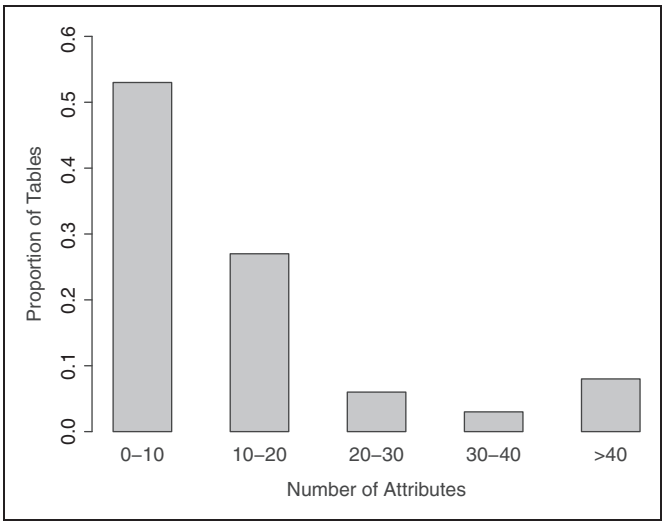

FIG. 9. Distribution of schema sizes.

have examined the diversity of schemata with respect to their attribute names.

To compute the similarity between the schemata of two tables, we applied the hierarchical agglomerative clustering (HAC) using Jaccard similarity.[22] In the first step, the HAC algorithm considers each individual schema as an initial cluster. Then, it iteratively combines the two most similar clusters. This process stops when there is a single cluster.

We ran HAC over the set of tables published by each city. Figure 10 shows the percentage of initial clusters with different similarity values for 5 cities. The remaining 15 cities follow a similar pattern, and they are thus omitted. When the similarity value is 1 (a perfect match), the algorithm joins two tables with the exact same attribute names. After this point (value <1), the algorithm starts grouping schemata with smaller overlap.

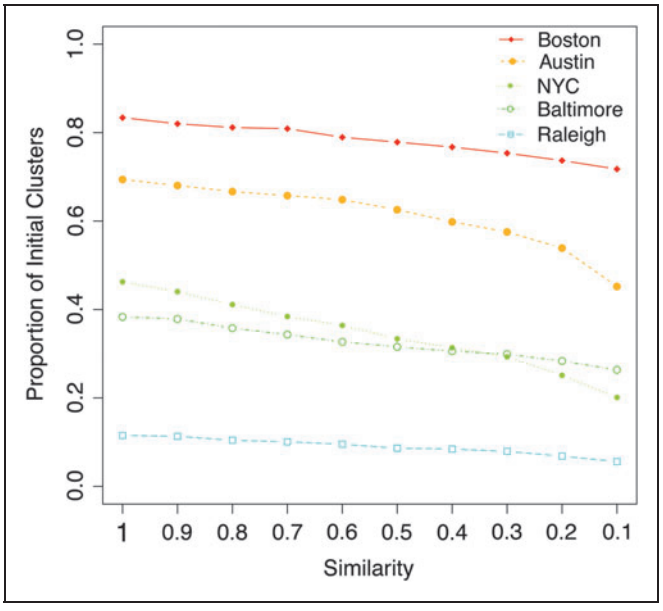

FIG. 10. Schema diversity for tables in five cities.

The schemata of tables in Boston are the most diverse: when the similarity value is 1, 83% of the initial clusters remain; after lowering the similarity to 0.1, still 72% of the initial clusters remain. On the other hand, the schemata of Raleigh's tables are the most homogeneous. Only 11.5% of the initial clusters remain with similarity 1, and 5% with similarity 0.1. Baltimore and NYC also have a small percentage of the initial clusters for similarity 1 (38% and 46%, respectively). One reason for this homogeneity is that these data sets contain many variations of the same tables. For example, the NYC collection has many versions of the 311 data set, covering different complaint categories.

Another interesting observation from Figure 10 is that the variation of percentage (from similarity 1 to 0.1) of initial clusters provides an idea about smaller overlaps. The curves of Figure 10 show small variations for different similarity values, indicating that the overlap across tables is small. The NYC data sets are the ones that present the highest variation (26%), which indicates that their schemata might be more easily integrated because there is a large overlap with respect to attribute names.

To get a different view of attribute overlap across tables, in Figure 11 we show the similarity matrix of tables in Boston and NYC (without 311 tables). Each cell in the matrix represents the similarity between two tables. A dark green cell indicates that the corresponding tables have similarity equal to 1; that is, they have the same schema. When the similarity between the two tables is less than 1, a lighter green is used. The fact that there is a large number of green cells shows that there is a substantial overlap across tables, indicating that there is great potential for integrating these tables. For NYC, we removed the 311 data sets because, when present, they led to a very large dark green square that obfuscated the other overlaps.

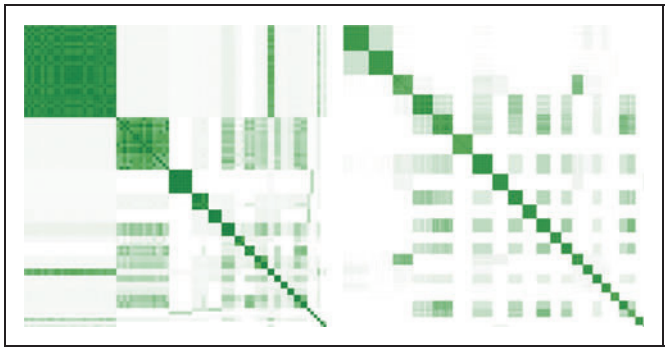

FIG. 11. Similarity among data sets taking into account their schemata and overlap of attribute names.

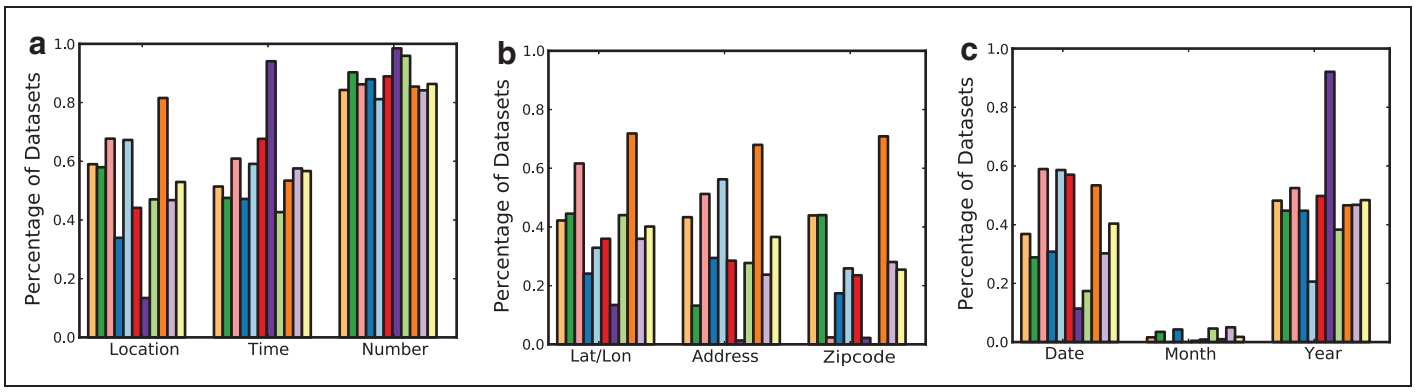

FIG. 12. Proportion of different types in data sets for 10 cities.

What types occur in the tables?

Besides the attribute names, another feature that can be used as an indicator for integration potential is attribute types.[13] Tables in our collection contain column types in their metadata, but in many cases, generic types such as text and number are used. To obtain more specific type information, we built detectors for types that denote location and time, as well as finer-grained types, including latitude/longitude, address, date, month, and year. The detectors apply regular expressions and rules that use both the name and values of a given attribute to determine its type. We extract a sample of attribute values by extracting the first 100 non-null values in each data set.

Figure 12 shows the proportion of tables that contain a given type for the 10 cities with the largest number of data sets: Austin, Baltimore, Boston, Chicago, Edmonton, Kansas City, Oakland, Raleigh, San Francisco, and NYC. As a point of comparison, we also show the proportion when considering the tables for all cities ("All Cities"). Spatiotemporal types are prevalent in these data sets. Latitude/longitude are the most frequent types among the location-related ones: they are present in more than a half of the data sets (52.9%). For some cities (e.g., Seattle and Boston), more than 60% of the tables contain latitude/longitude columns, whereas for Raleigh, they are present in only 13% of the tables. For the time-related types, there is a considerable proportion of tables that have date and year information, 40.4% and 48.4%, respectively. Month is present in many fewer tables, but dates often contain information about month. There is thus great potential of joining tables on spatiotemporal attribute.

How sparse are the tables?

Null values represent missing data, and a high proportion of nulls might indicate data quality problems. Common values we observed in these tables to indicate missing information include "null," "unspecified," "unknown," or "N/A." We examined the proportion of null values in the tables, and Figure 13 summarizes our findings. The great majority of tables have very low sparseness; for example, 63% of them have sparseness between 0 and 0.1; that is, at most 10% of the values are null. There are, however, cases where tables have many null columns, that is, columns where all or most of the values are null. For instance, the San Francisco table "p4sp-es3b"[‡] has 71 null columns out of 86 (82.6%). A considerable number of tables for the different cities have null columns, ranging from 1.9% (Raleigh) to 31.1% (NYC).

How informative are the attribute names?

A good practice in designing a database is to follow name conventions.[23] An important rule is to have meaningful names for table columns, since it makes it easier for users to

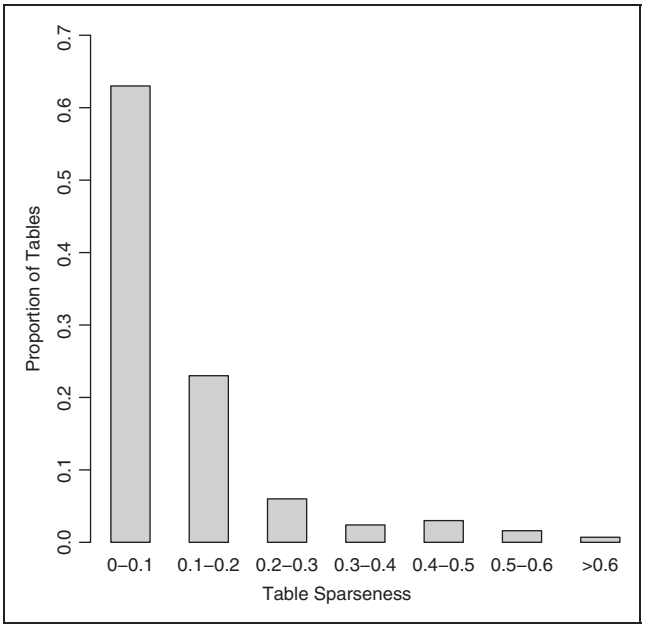

FIG. 13. Distribution of table sparseness: proportion of null values in tables.

[‡]This data set includes all e-filed on Fair Political Practices Commission (FPPC) Form 496 "Part 3" itemized contributions of $100 or more received since 2009.

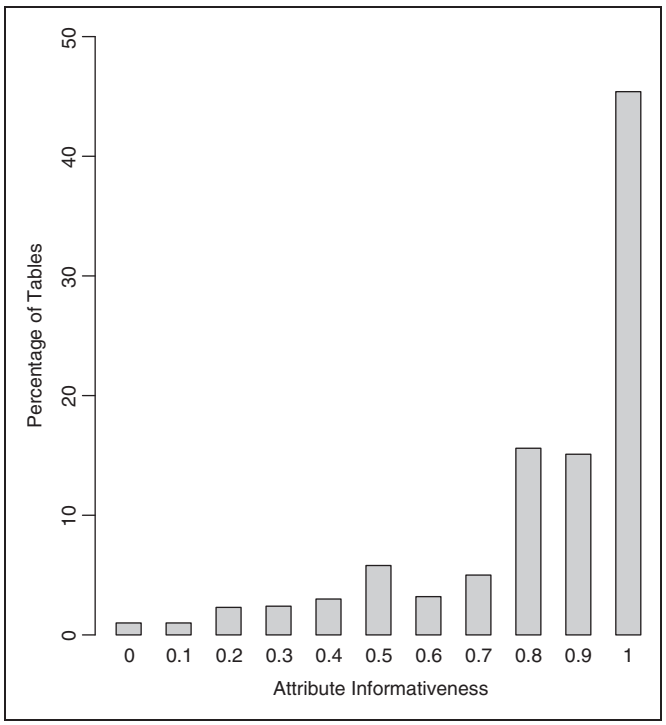

FIG. 14. Degree of informativeness: proportion of columns whose names contain words in the English dictionary.

understand the semantics of the tables in the absence of detailed descriptions for the attributes. Meaningful names are also helpful while integrating multiple data sources.[24] We measure how informative (meaningful) a column name is by checking whether the name is described using words in the English dictionary. For each table, we measured the proportion of informative columns, which we call degree of informativeness. In order to compute this value, we first tokenize the column names with underline characters, and then check whether the tokens with more than two characters are present in the Wordlist dictionary.[25]

Figure 14 presents the distribution of the degree of informativeness for all tables. Interestingly, most of the tables present a high degree of informative fields; about 76% of the tables have degrees of informativeness higher than 0.8, which means that at least 80% of their field names were present in the dictionary. Note that these figures represent a lower bound for informativeness, since some field names have words concatenated (e.g., "citylocation" and "creationdate"). The table Internet and Global Citizens from Austin, Texas, is an example of a table with a low degree of informativeness. The majority of its columns have names such as q69a, q8a7a, and q8c1. These columns contain answers to a questionnaire about Internet access, which is hard to infer by looking at the column names.

What is the geographical coverage of the data?

As discussed in the section titled What types occur in the tables?, many tables in our data set contain location information. Another interesting question that arises is how much of a city is covered by the data; that is, what is the geographical coverage of these tables? To answer this question, we converted attributes with location type to zip codes. The heatmaps in Figure 15a–b show the frequency of references to the zip codes in Chicago and NYC.

These maps suggest a correlation between the number of references to zip codes in these cities and their population. To verify this observation, we collected the population size for the zip codes* and ran a statistical test (Spearman correlation) between the two variables: zip code references in the tables and the population in the zip code. The correlation is very strong (0.86 and 0.88 for Chicago and NYC, respectively), indicating that highly populated zip codes usually have a large number of references in the data sets. Figure 16

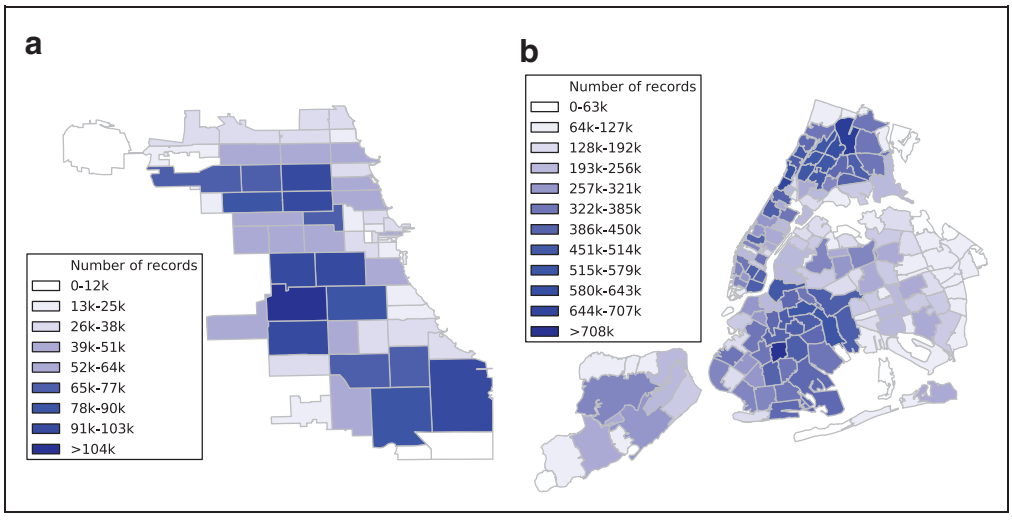

FIG. 15. Heatmaps of the geographical coverage of the data sets for **(a)** Chicago and **(b)** NYC.

*This information was obtained from www.zip-codes.com

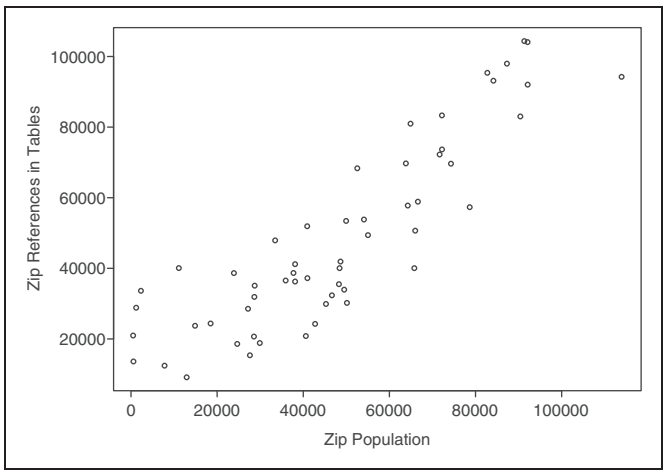

FIG. 16. Population of zip code regions versus references to the zip codes in Chicago data sets.

shows the scatterplot of population versus number of zip code references for Chicago.

Discussion and Challenges

There is great value in opening up urban data as a means to both increase transparency and enable new uses and applications for the data. The steady increase in the volume of open urban data over the past few years provides a good indication that this trend is here to stay. As we observed, cities of all sizes are opening their data, although bigger cities publish a larger number of data sets. Furthermore, a large percentage of the open data come in easy-to-parse formats, and cover a wide range of topics and city affairs. But making data accessible in machine readable format is just a starting point. There are many challenges involved in actually using these data. One indication is the number of downloads for the different data sets, which is relatively low, with most tables having been downloaded less than 100 times. Below, we discuss some of these challenges.

"THERE IS GREAT VALUE IN OPENING UP URBAN DATA AS A MEANS TO BOTH INCREASE TRANSPARENCY AND ENABLE NEW USES AND APPLICATIONS FOR THE DATA."

While many data sets are available, they can be hard to find. Publishing platforms such as CKAN[16] and Socrata[17] provide only simple keyword-based searches over the metadata; consequently, users are not able to identify data sets based on their content, for example, to select all data sets that cover a given time period or a region. To facilitate data discovery, Socrata defines a set of categories for the data sets. For NYC, there are 21 categories, for example, education and community service.* A more comprehensive taxonomy, with finer-grained categories like DMOZ,** would provide a better mechanism for users to browse and explore the data sets. In addition, it could also serve as a means to catalog data sets from multiple cities.

Currently, each city publishes its own data independently, on dedicated web sites. This makes it hard to find related data across different cities. Having a shared directory as well as an urban search engine would give users a single entry point to locate relevant data and simplify the process required to perform analyses that require data from multiple cities.

Socrata provides a set of basic filters and visualizations that can be applied to the data sets, allowing users to quickly inspect the data using their web browsers. This can explain the much larger number of views compared to the number of downloads. It also underscores the importance of web-based interfaces and apps that are easy to use and accessible to the general public, and yet provide more sophisticated analysis and visualization capabilities, such as ManyEyes[26] and Tableau Public.[27]

Given the overlap present in the schemata of tables and the pervasiveness of informative attribute names, there is an opportunity to integrate these tables. In addition, the prevalence of location-related attributes suggests that joining tables on location would be a relatively simple form integration. Nonetheless, there is substantial heterogeneity across the data sets; many different terms are used for a given attribute, and a given term can be used to represent different concepts. Mechanisms and tools are needed that assist users in identifying (potential) links between data sets. While there has been substantial research in information integration, we lack usable tools that support on-the-fly integration and at a large scale. Data profiling tools can also aid in the use and integration of open data, since these can automatically derive metadata and enrich the manually derived descriptions that are currently available.[26] Finally, while currently the total volume of data is small, around 70GB for all cities, this volume is increasing steadily. Consequently, there is a great need for scalable and automatic techniques to process and integrate these data.

Acknowledgments

This work was supported in part by the National Science Foundation award CNS-1229185. J.F. was partially supported

*http://www.nyc.gov/html/doitt/downloads/pdf/nyc_open_data_tsm.pdf
**http://www.dmoz.org

by a Google Faculty Research award. J.F. and C.S. were partially supported by the Moore-Sloan Data Science Environment at NYU and by IBM Faculty Awards.

Author Disclosure Statement

No competing financial interests exist.

References

1. The World Bank. Urban Development. Available online at http://data.worldbank.org/topic/urban-development, 2014 (Last accessed on Feb. 2, 2014).
2. Goldstein B, Dyson L. Beyond Transparency: Open Data and the Future of Civic Innovation. San Francisco: Code for America Press, 2013.
3. Höchtl J, Reichstädter P. Linked open data—a means for public sector information management. In: Electronic Government and the Information Systems Perspective, Volume 6866 of Lecture Notes in Computer Science. Berlin: Springer, 2011, pp. 330–343.
4. Shadbolt N, O'Hara K, Berners-Lee T, et al. Linked open government data: Lessons from data.gov.uk. IEEE Intell Syst 2012; 27:16–24.
5. NYC Open Data. Available online at http://data.ny.gov (Last accessed on September 3, 2014).
6. Chicago Open Data. Available online at https://data.cityofchicago.org/ (Last accessed on September 3, 2014).
7. Seattle Open Data. Available online at http://data.seattle.gov (Last accessed on September 3, 2014).
8. NYC MTA API. Developer Resources. Available online at http://web.mta.info/developers/ (Last accessed on September 3, 2014).
9. Ellen IG, Lacoe J, Sharygin CA. Do foreclosures cause crime? J Urban Econ 2013; 74:59–70.
10. Affordable Housing. Available online at http://furmancenter.org/research/area/affordable-housing (Last accessed on Aug. 14, 2014).
11. Ferris B, Watkins K, Borning A. OneBusAway: Results from providing real-time arrival information for public transit. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 2010, pp. 1807–1816.
12. Mapumental. 2013. Available online at http://mapumental.com (Last accessed on Nov. 1, 2013).
13. Rahm E, Bernstein PA. A survey of approaches to automatic schema matching. VLDB J 2001; 10:334–350.
14. Doan A, Halevy AY. Semantic integration research in the database community: A brief survey. AI Mag 2005; 26:83.
15. Halevy A, Rajaraman A, Ordille J. Data integration: The teenage years. In: Proceedings of the International Conference on Very Large Data Bases (VLDB). New York: ACM, 2006, pp. 9–16.
16. CKAN. Available online at http://ckan.org (Last accessed on May 28, 2014).
17. Socrata. Available online at www.socrata.com (Last accessed on May 28, 2014).
18. Open Definition. 2014. Open data definition. Available online at http://opendefinition.org/od/ (Last accessed on May 1, 2014).
19. Available online at https://nycopendata.socrata.com/Public-Safety/Inmate-Arrests/d4uz-6jaw
20. Available online at https://data.austintexas.gov/dataset/The-Austin-Internet-and-Global-Citizens-Project/gt3n-akq9
21. Madhavan J, Bernstein PA, Rahm E. Generic schema matching with cupid. In: Proceedings of the International Conference on Very Large Data Bases (VLDB). New York: ACM, 2001, Volume 1, pp. 49–58.
22. Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. In Proceedings of the ACM SIGKDD Workshop on Text Mining. New York: ACM, 2000, Volume 400, no. 1, pp. 525–526.
23. Coronel C, Morris S, Rob P. Database Systems: Design, Implementation, and Management. Stamford, CT: Cengage, 2012.
24. Huang CCE, Chiang RHL, Lim E-P. Instance-based attribute identification in database integration. VLDB J 2003; 12:228–243.
25. Wordlist Dictionary. 2014. Available online at http://wordlist.sourceforge.net/pos-readme (Last accessed on September 3, 2014).
26. Viegas FB, Wattenberg M, van Ham JK, et al. ManyEyes: A site for visualization at Internet scale. IEEE Trans Vis Comput Graph 2007; 13:1121–1128.
27. Tableau Public. Available online at www.tableausoftware.com/public (Last accessed on Aug. 17, 2014).
28. Naumann F. Data profiling revisited. SIGMOD Record 2013; 42:40–49.

Address correspondence to:

Juliana Freire
Department of Computer Science and Engineering
New York University
2 Metrotech Center, 10th Floor
Brooklyn, NY 11201

E-mail: juliana.freire@nyu.edu