Towards an Ecosystem of Structured Data on the Web

Alon Y. Halevy
Google Inc.

1. INTRODUCTION

We are in the midst of very exciting times in which structured data is having a profound impact on many aspects of our lives. In many countries, citizens take for granted the fact that governments, local authorities, and non-government organizations should make a variety of data sets available to the public. These data sets span a variety of topics such as economic indicators, crime statistics, educational data, government spending and campaign contributions. Journalists and other data aficionados are fueling this trend by turning this data into visualizations and stories that are spread by social networks and seen by millions of people [8]. These visualizations, stories and public attention, in turn, lead to new questions and hence a demand for additional data.

The potential for the future is even more promising. With the proliferation of smart mobile devices, we can now create in a timely fashion databases that were nearly impossible to create before, such as a current map of potholes [5], rural villages with access to free water [7], or the price of a bottle of mineral water anywhere in the world. Hence, in addition to the vast collection of structured data on any topic of interest to mankind that already exists on the Web, we can expect an influx of relevant and timely data in the years to come.

To realize this bright future, data management tools need to rise to the occasion and address some difficult challenges. We need to create an ecosystem of tools that play well together and provide the necessary services to entice data owners to contribute data and others to enhance and manipulate the data and to create visualizations. I briefly highlight some of these challenges below.

2. EASE OF USE

The most fundamental challenge we face is to make data management systems that are easy to use. Simply put, the only way in which we can get more data sets online and make use of them is if more people can do the job. People who have access to interesting data sets and who are most interested in it are rarely the ones with technical skills to use a database system. For example, while some journalists have lightweight programming skills, the vast majority do not, and they seldom have any database experts at their disposal. In addition, in order for their articles to be timely, they must be able to quickly massage a data set into a story when the data is made available to them.

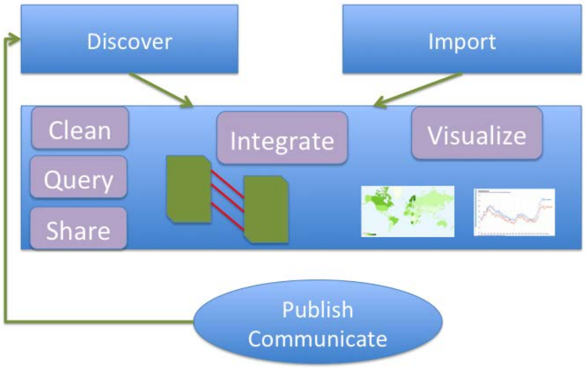

Figure 1: To fully realize the potential of structured data on the Web, we must create an ecosystem around data that includes tools for discovering data on the Web, importing data from existing repositories, tools for easily querying, cleaning, integrating and visualizing data, and mechanisms for easily publishing data to the Web.

Of course, ease of use has been a long standing challenge for the database community and steady (albeit slow) has been made over the past few years. An important observation to keep in mind as we think of ease of use is that the data management systems we develop for “the masses” need not have the same functionality as traditional databases. For example, in Google Fusion Tables [2], a tool that has been particularly popular with journalists, we focused initially on providing an intuitive and fast path from data ingestion to a visualization. Fusion Tables enables users to upload data in multiple formats without requiring the user to declare (or even be aware) of schema. The system then tries to find columns that can serve as keys for visualization (e.g., locations for a map, time points for a time-line), and then lets the user configure a visualization. The general principle here is that we need to think of the common tasks that users face and make sure the system supports these tasks in the most intuitive and efficient way.

EDBT 2012, March 26–30, 2012, Berlin, Germany.
Copyright 2012 ACM 978-1-4503-0790-1/12/03 ...$10.00.

3. CREATE A LOGICAL CLOUD

On the physical level, cloud-based data management services go a long way to increase usability. Users do not need to download and configure a database system or worry about provisioning enough storage for their data. However, the cloud holds another important promise at the *logical* level. Specifically, if many high quality data sets are put in the cloud, the cloud becomes a rich resource for data that can significantly simplify applications by enabling data reuse.

Imagine a scenario in which an analyst is looking at the latest trends in her sales data by county and considering additional locations to focus sales efforts. The analyst may realize quickly that she is missing critical demographic context in her analysis, such as the population of each county or its average income. For the average analyst, this data may be hard to come by. If she works for a large enterprise the data may be available somewhere (but hard to find), but a smaller enterprise is unlikely to even have the data. Individuals performing more transient analyses can benefit even more from such collections of data.

However, population and income data are publicly available from reliable sources. Hence, if her data was in the cloud, she should be able to simply ask for data about population. The system would examine the locations in her database and find an appropriate data set that has population data for these locations. Adding this contextual data should be so easy that the analyst should not even be aware that she performed a join. Importantly, the provenance of the additional columns should be made very clear so she can decide whether it is trustworthy for her purposes.

4. DATA DISCOVERY

One of the main incentives for data owners to make data available is that it be easily discoverable through search engines. However, ranking techniques that work well for text documents do not necessarily transfer over to structured data. There are two main sources of difficulty. The first is that it is hard to decide which data on the Web that has *syntactic* structure (e.g., HTML table) contains high-quality data. In fact, less than 1% of the HTML tables on the Web have good data in them because most uses of HTML tables are exclusively for formatting purposes [1]. In recent work [9] we have shown that detecting semantic coherence of a column in a table is an effective signal for determining whether a table has high-quality relational content. Interestingly, semantic coherence can be predicted from mining text on the Web. Specifically, we can use text mining techniques (such as Hearst patterns [4]) to determine types for entities mentioned on the Web. If a majority of the values in the subject column of a table are found to be of the same type, that is a strong signal that the table contains relational data.

The second challenge is that we know very little about the semantics of structured data on the Web, and therefore deciding whether a table is relevant to a query is hard. In particular, the only schema information we have about tables are column names (at best) but the relations represented by the data are typically described in the surrounding text. Moreover, the semantics are brittle – changing one token on the page (e.g., which year the data is from) can completely change whether a table is relevant to a query.

So far the work on data discovery has focused on finding individual tables that are relevant to a keyword query, but an obvious extension would be to find combinations of tables (e.g., joins, unions) that can answer queries.

5. LEVERAGE COMMUNITIES

There has been a lot of work recently on extending database systems with crowd-sourcing techniques. The key idea is that some database predicates are hard for machines to evaluate but are easy for humans (e.g., *does this photo contain a sunset?*). One of the main challenges in that body of work is that the crowd is often assumed to be low-paid workers with unverified skills, and therefore sufficient redundancy needs to be built into the system in order to ensure that the answers obtained are precise.

In many applications, the crowd need not be a nameless set of individuals. For example, imagine professional communities that would like to collaborate to produce higher quality data that they can share, such as scientists collecting and analyzing data about ecosystems [6], or coffee professionals assembling databases about farms and cafes worldwide [3]. In these contexts, the community should be able to easily identify coverage gaps in their data, places where the quality of their data needs to be improved, or opportunities where resolving semantic heterogeneity would provide high value. Once these issues in the data are identified, the community should be able to efficiently go about filling the gaps in a collaborative fashion. Unlike the typical turkers, here the community is built of motivated individuals (of many levels of expertise and cost), and data collection can be done much more efficiently.

6. CONCLUSION

The topics mentioned above are only a partial list of the challenges we face to make the ecosystem structured data on the Web a reality. Clearly, there is a lot of work needed in the areas of data integration, creating visualization tools that are easy to use and can efficiently handle huge amounts of data, creating mechanisms for determining quality of data sets we find on the Web so users can make an informed decision about when to use them, and generally building systems that try to be proactive and help users with common tasks. To make advances that have real impact, all of these endeavors need to keep a strong focus on supporting users and their tasks.

7. REFERENCES

[1] M. J. Cafarella, A. Halevy, Y. Zhang, D. Z. Wang, and E. Wu. WebTables: Exploring the Power of Tables on the Web. In *VLDB*, 2008.

[2] H. Gonzalez, A. Halevy, C. Jensen, A. Langen, J. Madhavan, R. Shapley, W. Shen, and J. Goldberg-Kidon. Google Fusion Tables: Web-Centered Data Management and Collaboration. In *SIGMOD*, 2010.

[3] A. Y. Halevy. *The Infinite Emotions of Coffee*. Macchiatone Communications, LLC, 2011.

[4] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING*, pages 539–545, 1992.

[5] Otrobache.es. Potholes in spain. http://www.otrobache.com/in/spain, 2010.

[6] PCAST Working Group. Sustaining environmental capital: Protecting society and the economy. http://www.whitehouse.gov/administration/eop/ostp/pcast/docsreports, 2011.

[7] F. Rijsberman. Every last drop: Managing our way out of the water crisis. *Boston Review*, 2008.

[8] S. Rogers. 2011: the year in data, journalism (and charts). http://www.guardian.co.uk/news/datablog/2011/dec/30/top-data-stories-2011?newsfeed=true, 2011.

[9] P. Venetis, A. Y. Halevy, J. Madhavan, M. Pasca, W. Shen, F. Wu, G. Miao, and C. Wu. Recovering semantics of tables on the web. *PVLDB*, 4(9):528–538, 2011.