Data Management for Journalism

Alon Halevy
Google Research

Susan McGregor
Columbia School of Journalism

Abstract

We describe the power and potential of data journalism, where news stories are reported and published with data and dynamic visualizations. We discuss the challenges facing data journalism today and how recent data management tools such as Google Fusion Tables have helped in the newsroom. We then describe some of the challenges that need to be addressed in order for data journalism to reach its full potential.

1 Introduction

For decades, computer-assisted reporting (CAR) has been an essential aspect of enterprise and investigative reporting, using compilations of public records, private databases, and other specialized data sources to reveal newsworthy patterns and anomalies, or simply to identify leads for further investigation using more traditional reporting techniques. Yet while personal computers have been a newsroom mainstay for many years, CAR has remained an essentially niche practice in the newsroom, as most reporters have been ill-equipped to exploit the often powerful resources available on their hard drives. The integration of data analysis into mainstream reporting has been stymied by a range of obstacles: arcane and obscure data formats requiring highly specialized languages and skills to manipulate, limited technologies to support visualization and pattern-recognition, data-literacy and numeracy challenges among reporters, and outdated or difficult-to-obtain data sets. Increasingly, news organizations must compete with social media to break news, and the expertise overhead and time lag involved in obtaining, cleaning, and analyzing data has made it impossible to use for deadline reporting. Recent innovations in Web-based data management tools, however, have begun to dismantle some of these obstacles, giving rise to the broader practice of data journalism, which is quickly becoming a core technique of the 21st century newsroom. In this paper we describe how two such tools - Google Fusion Tables [9] and Google Refine [10] - have impacted data-driven reporting, and we describe the next set of challenges to empowering data journalism.

We begin with an example that illustrates the power of data journalism and its growing role in public discourse. On August 11, 2011, British Prime Minister David Cameron addressed an emergency parliamentary session called in response to the widespread UK riots of the preceding week. In his remarks, he stated that *"Everyone watching these horrific actions will be struck by how they were organised via social media,"* and suggested that the UK government was exploring the possibility of limiting or banning access to social media during periods of public unrest [12].

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

Since the Arab Spring of 2011, the role of social media as an organizing tool for demonstrations and protests had become accepted wisdom in many circles, providing a basis for Cameron's assertion that the government should have the power *"to stop people communicating via these Websites and services when we know they are plotting violence, disorder and criminality,"* [12]. Less than two weeks later, however, a preliminary analysis and visualization of more than 2.5 million tweets by the Guardian UK [4] indicated that riot-related traffic on the Web service tended to spike *after* violence began in a particular neighborhood, as shown in Figure 1. More formal analyses conducted in partnership with the London School of Economics and the University of Manchester over the following months confirmed that messages shared through social media outlets like Twitter and Facebook were not a factor in organizing the riots. In fact, they played a vital role in mobilizing cleanup efforts [1].

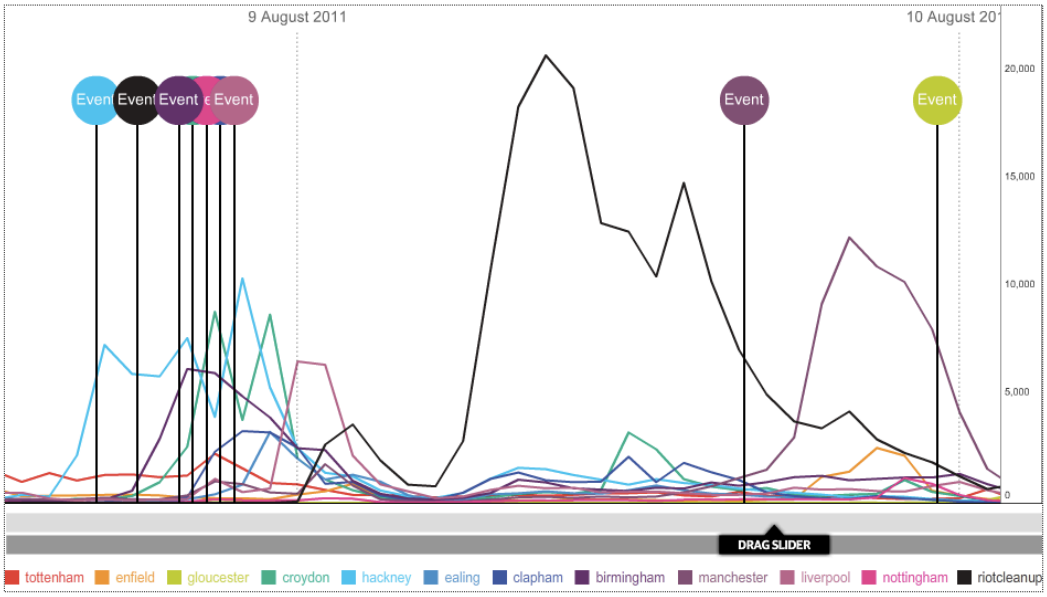

Figure 1: A screenshot of the interactive line chart published by the Guardian on its website on August 24, 2011. The large spike in the middle, shown in black, indicates tweets related to the riot cleanup.

As the Guardian's work demonstrates, there remains an elemental need for professional journalism executed with relevance, rigor and accuracy. Though one can only speculate what actions the government may have taken in the absence of the Guardian's reports, there can be little doubt that the public understanding of the UK riots would have been importantly diminished without them.

This type of insight and public empowerment is at the heart of all great journalism, but is increasingly the purview of data-driven analyses in the newsroom. Whether it's the Guardian's work defending civil liberties in the U.K., the prompting of privacy legislation in the U.S. triggered by the Wall Street Journal's "What They Know" series or the arrests and reform spurred by the L.A. Times' analysis of government payrolls, data journalism leverages the ability to canvass and correlate information with both a scale and detail that would be impossible through traditional reporting. With more than one-quarter of the world's 2.4 billion people creating or consuming content on the Web [24], the volume of information available as digitally-accessible data can only continue to outstrip the knowledge of even expert individuals. In this environment, the practice of data journalism will only grow in its essential role for the industry as a whole.

2 Challenges to Data Journalism

We describe the main challenges faced by data journalists. Many of these challenges are faced by practitioners in other professions as well, but we describe them from the persepctive and expertise of data journalists.

Discovering high-quality data: While many journalists rely on Web searches to locate data sets and other structured information, indexing such information so that it can be discovered through search engines presents significant challenges. Interestingly, humans and computers face two of the same essential difficulties when attempting to locate and identify high-quality information online.

The first challenge is that a great deal of the structured information that is available online resides in the deep Web. Data on the deep Web is typically accessed through HTML forms (e.g., public records and reports, statistical data). Though the data may be well-structured and high-quality, the actual information is sitting in a back-end database and is available as an HTML page only when a user enters a particular query or set of commands. Such content is often described as part of the *invisible Web* because search engines cannot penetrate the forms and crawl the underlying content. Likewise, interface usability issues, specialized query syntax and sui generis data encodings may make the contained data as inaccessible to human users as to their computer counterparts. Though in some cases journalists may be able to call upon system experts (such as librarians or government employees) for assistance, legal and interpersonal barriers can stymie the best efforts at extracting data from the deep Web or similarly-constructed institutional repositories. While there have been some efforts to surface deep-web data [15] and categorize deep-web sites [2], there is still a long way to go to make this data available to journalists.

Even where structured data exists on the so-called surface Web, there remain significant technical challenges to its discovery and use. While more visible to humans, Web pages that contain tables, HTML lists, or have a repeated structure are in fact quite difficult for search engines to index properly and return as well-ranked results for search queries.

The primary difficulty is that it is hard for search engines to determine which pages on the Web contain high-quality data. Less than 1% of the HTML tables on the Web have good data in them, in part because so many of them are used exclusively for formatting purposes [5]. However, recent work [25] has 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. For example, a table whose rows all contain the names of tree species is semantically coherent and therefore probably contains useful structured data. While such insights are useful, the current reality is that we still know very little about the semantics of structured data on the Web. Because the only schema information we have about tables are column names (at best), inferring the broader context of data is quite difficult.

These inferential challenges often apply to human users of structured data sourced from the Web as well. Limited metadata - whether the information is accessed through a form or page - often makes it difficult to determine the usefulness of a given data set in answering a particular journalistic question. In cases where sufficient metadata can be accessed through another part of the page or site, human users still face the complex task of ”scraping” the site to extract the meaningful data from surrounding markup. Tools like ScraperWiki [23] can help support journalists in this task, but rendering the data in a malleable format is still only a first step. Once accessed, there remain the enormously difficult tasks of correcting, editing and making sense of the data - all before its journalistic relevance can even be assessed.

“Dirty” data: Once a data set has been obtained in an appropriate, non-proprietary format, it most often needs to be corrected or *cleaned* before it can be used. Errors, formatting irregularities, missing values, or unknown conventions must be identified, resolved, and systematically addressed before any kind of analysis can occur. For many years, the go-to tool for data cleaning in newsrooms has been Microsoft Excel, at least in part because of its ubiquity as part of the Microsoft Office suite installed on many PCs. Excel provides basic data sorting

and editing features like find and replace, as well as mathematical and statistical functions for calculating sums, averages, and even standard deviations and z-indexes. Advanced features can be used to create basic charts, graphs, macros, and pivot tables for data analysis.

While the spreadsheet interface implemented by Microsoft Excel and other programs is generally accessible even to novice users, effective use of their data-analysis functions still often requires expert instruction and many hours of practice. Moreover assessing the specific size, contents, and parameters of a given data set constitutes a large part of understanding whether it is relevant to a particular journalistic question, and many spreadsheet tools still require that much of this type of analysis be done by inspection.

Google Refine [10], by contrast, has significantly streamlined these fundamental tasks of data analysis. The main interface - while still following a basic spreadsheet layout - immediately displays the loaded data's row count, and updates this figure whenever it is affected by a user action. Text faceting, for example, instantly displays every unique data value present in the selected column. Because the default text-sort in the summary window is alphabetical, minor misspellings or letter-case differences will often appear adjacent to one another, enabling the user to quicky transform them to the same string, a common data cleaning operation. In addition, Google Refine lets users reconcile cell values with entities in Freebase, providing additional context to the data.

Google Refine also provides functions for quickly viewing subsets of even complex data. For every column that is faceted, selecting a value in the summary window filters the data to show only rows matching that value. This process can be repeated across multiple summary windows, and the filtered dataset can be exported at any time without special configuration. Perhaps most importantly, the process is instantly reversible: all filters can be instantly removed and the full dataset restored by simply closing out the summary windows. Finally, Google Refine records the set of transformations performed on the data as scripts. Hence, these transformations can be applied again if the data is reloaded from the source.

The tools of interrogation: Practicing data journalism, at its most fundamental level, means asking journalistic questions of data. Such data interrogation is best achieved through tools that allow data sets to be correlated, queried, manipulated and transformed. Ideally, any such tool will also execute these functions in a fully reproducible, fully recoverable way, so that results can be checked for accuracy, published for transparency, and assumptions can be tested at little or no cost.

Those who are familiar with the principles of databases will quickly recognize that they are by far more appropriate for interrogating data than any spreadsheet program. Indeed, even the faceting feature of Google Refine discussed above essentially implements the simplest type of querying function supported by traditional database systems, and its automatic macro generation constitutes a transferable, stepwise documentation of data manipulation similar to a SQL query. Yet while there is no question that database systems are incredibly powerful, they are notorious for being difficult to use and requiring very skilled and dedicated personnel to use and manage them. Desktop database systems, such as Microsoft Access and SPSS, are costly, and often result in data sets siloed on individual computers. Web-based systems require server space to be purchased, configured, and secured by specialists. Even where these resources are available, traditional database systems ill-suited for relatively transient data management tasks done by people with limited technical or data management experience.

Additionally, most database systems do not have a straightforward method for associating essential metadata with individual data tables or columns. As discussed above, insufficient context around structured data sets are a severe barrier to their effective (re)use. In journalism, the provenance of a data set is paramount. Even where the source is trusted, the recency, measurement units, encodings and other contextual information (e.g., are these employment figures seasonally-adjusted?) that do not comprise the data itself must be evaluated to determine its appropriateness for a given story. Database systems were built with the main goal of supporting high throughput transactions and running complex SQL queries. Though this feature obviates the need to save multiple views of the same data set or email updated files, database tables' lack of contextual information can be profound. An

emailed file has at least some associated source (the sender) and a timestamp (the date of the email). The text of the email itself is likely to contain some contextualizing information. In the absence of cues about origin and context, even technically networked data (e.g., Linked Data [3]) may fail to live up to one of its foremost potentials: recombination and reuse.

Data integration: In today's publishing environment, leveraging the possibilities of networked data has become so important, in fact, that it has helped redefine one of the most significant terms in information technology today: Big Data.

Traditionally Big Data has been characterized solely by its size - specifically, whether it required a supercomputer to process. The most relevant measure of Big Data today, however, is not the size of the file but the extent of the network [7]. Thus, a single YouTube video becomes Big Data by generating 100 million page views, as occurred with the Kony 2012 campaign [14]; likewise, so do the few hundred insurance companies identified by the Sarasota Herald Tribune through analysis of their networks of financial assets and obligations [13]. Because Big Data comprised of networks and relationships embodies phenomena that affect a large number of people, these types of data sets are especially germane to the journalistic enterprise. Exploring these data sets requires powerful tools for integrating data from multiple independent sources, and the main challenges include large-scale entity and schema resolution.

3 New Online Tools

In recent years, a set of new online data management and visualization tools such as Google Fusion Tables [9] and Tableau Public [16] have given journalists more power to discover and tell stories with data. Both of these systems were inspired in some way by ManyEyes [26], an earlier online visualization system. While ManyEyes focused solely on visualization and collaboration around visualization, Fusion Tables and Tableau provide rich data management features. Consequently, users can explore and query their data before the publish a visualization. In what follows we describe Google Fusion Tables and some of its applications within the field (see [8] for a gallery of examples of Fusion Tables in journalism).

Google Fusion Tables is a Web-based tool that combines and extends aspects of spreadsheet, database, graphing and mapping software that supports real-time, responsive, networked data analysis and publishing. Fusion Tables enables easy import of data from CSV and spreadsheet files, and even guesses the data types of each column (which can also be adjusted by the user). On upload, the user is prompted to add meaningful metadata to the table, such as the source name and Web link, as well as a description of the data (with the useful prompt: *"What would you like to remember about this data in a year?"*). Once loaded, individual columns and even cells can be annotated with format and unit information, revision questions, and more.

Fusion Tables uses sharing to help support collaboration and elaboration around data. Imported tables are private by default – only the owner can see the data and make changes. However, the owner can also choose to share the table with collaborators, giving them permission to view, edit, and/or annotate the data. The owner may also choose to make the table public, making it available in both the Fusion Tables search interface, and for indexing by search engines.

Though Google Fusion Tables defaults to a familiar spreadsheet presentation, it supports the selection, projection and aggregation queries usually reserved to databases. Users can also perform simple joins between data sets, as long as they have read permissions to both. Thanks to its sharing model, Fusion Tables makes it possible to integrate data from multiple independent sources. For example, one can join a table containing the coffee production of different countries (coming from the International Coffee Organization) with data about the coffee consumption per capita (coming from Wikipedia).

Google Fusion Tables also provides tools for instant data visualization, an important component of analyzing data for patterns and anomalies. Through the automatic schema-detection, the system tries to identify columns

that can serve as keys for visualization. For example, Fusion Tables will try to recognize columns with locations and columns with time points. When it does recognize such columns, the visualization menu already enables map or time-line views of the data which can be further configured by the user. As such, Google Fusion Tables supports a very rapid flow from data ingestion to visualization in a variety of forms, including bar, line, and pie charts, as well as fully-styled interactive maps. While some of these visualizations can be generated through advanced use of spreadsheet programs, Fusion Tables' mapping feature deserves special consideration.

Much like data journalism and computer-assisted reporting, mapping in newsrooms has long been the purview of a small set of specially-trained reporters, and for many of the same reasons: creating maps required very expensive, complex software and substantial technical skill. Yet in a data set where location is a relevant parameter, the data must be mapped in order to perform any meaningful analysis; there is no universal mathematical construct that can act as a proxy for geography. Google Fusion Tables' location data type, which can interpret many different forms of location information, including country and city names, latitude/longitude coordinates, street addresses and KML, allows a wide range of geographic information to be mapped instantly. For example, see figure 2, an example of data integration where two data sets are shown on the same map.

The map shown in Figure 2 illustrates the power of data integration and mapping. A few days after the earthquake in Japan in March, 2011, the creator of this map combined two data sets that were developed independently: data about earthquakes since 1973 and data about the location of nuclear plants. This visualization helped address the question that was on the minds of billions of people around the world: would an earthquake in their area trigger the kind of disaster that was unfolding in Japan?

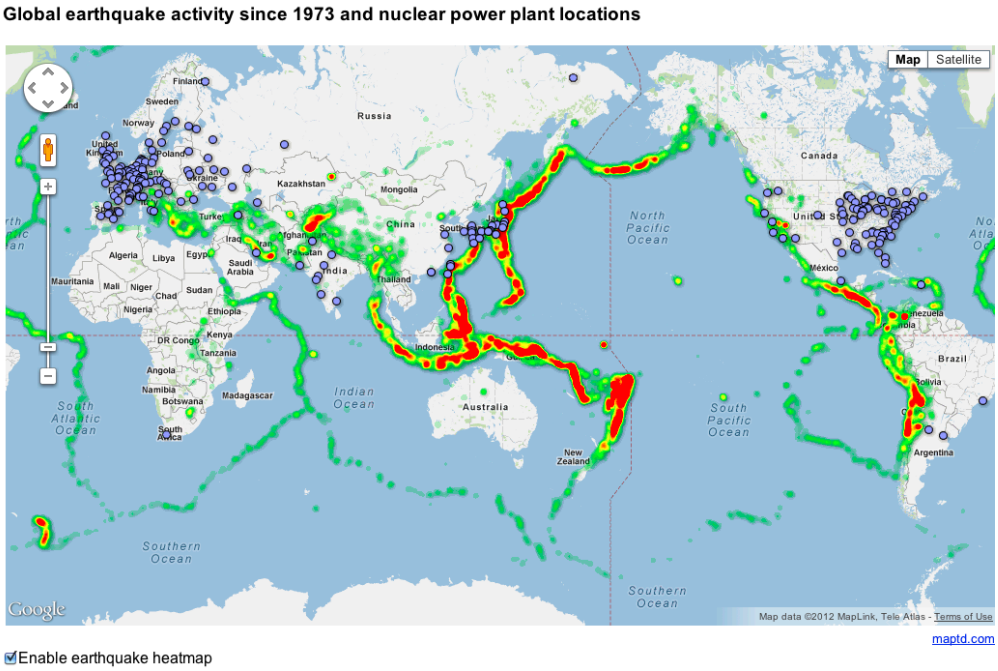

Figure 2: A map that combines two disparate data sources: (1) earthquake activity since 1973, and (2) location of nuclear plants. This map appears on `http://maptd.com`.

4 Remaining Challenges

Thus far, we have touched on some of the tools that are making it possible for more people to ask and answer such questions through the tools and practices of data journalism; below we outline some components needed to realize the full potential of the field.

4.1 Data Literacy

Data literacy needs to be made a priority in all professions, and with it an understanding of the issues around collecting, manipulating, and publishing data. Journalists must cultivate data literacy so that they can evaluate, interrogate and interpret data accurately. Judges must familiarize themselves with the material issues surrounding data formats and accessibility so that they can rule appropriately in data-related cases. Especially in Freedom of Information Act (FOIA) cases, they should require that all metadata and other information necessary to evaluating public data sets be a requisite part of any settlement [17]. As we have reiterated, contextual metadata is the single most defining aspect of a data set's informative value.

4.2 Safety and Privacy

While reporters can now file text, audio, photos – even video – from nearly anywhere, virtually any use of current communication technologies can leave potentially dangerous data traces. While new applications to address these issues are in development [21], both wireless and hard-wired service providers may share user information with government or other entities, endangering sources and journalists. Even if the original requests for user data are later deemed illegal, there may be no legal liability for companies that comply [6]. Used maliciously, this information can create serious threats to civil and human rights.

4.3 Standards and Accessibilty

The sheer number of expensive, complex tools that have been built to work with data is perhaps the clearest indication of its enormous economic value. In the public sphere, however, data should be held to the highest standards of transparency and accountability. Any allegedly public data set should be released in a non-proprietary data format with all metadata intact and relational fields preserved. Unfortunately, this is currently not often the case, as evidenced by the NYPD's release of annual Stop and Frisk data as zipped SPSS archives, and its quarterly reports as pdfs [19].

The journalistic and humanitarian value of standardized data formats is difficult to overstate. For example, many government agencies release weather and other geographic data in KML because it is an accepted global standard. That it is also instantly consumable by Google Maps and Google Fusion Tables means that when there are floods in the midwest, or a tsunami in the Pacific, lifesaving information about what areas are threatened can be published in a matter of minutes [18]. Developing tools and standards that eliminate the need for laborious data cleaning and correlation can improve the speed and accuracy of journalism, increase the transparency and accountability of government, and even save lives.

4.4 The Cloud and the Crowd

Cloud-based data management services such as Google Fusion Tables go a long way to increase the usability of data, in part by making structured data shareable and accessible from anywhere. 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 enhance analysis by enabling data reuse.

Imagine a scenario in which an analyst is looking at the latest trends in health data by county and considering additional locations to focus new efforts. She may be missing critical demographic context in her analysis, such

as the population of each county or its average income - but this type of data is publicly available from reliable sources. If her data is in the cloud, she could potentially just ask for data about population. If the system could examine the locations in her database and find an appropriate dataset that has population data for her locations, her analysis could be dramatically improved.

Adding this contextual data should be so easy that she should not even be aware she performed a database join. Importantly, the provenance of the additional columns should be made very clear, and she should even be able to choose from among any competing data sources. Some of these sources may be branded authorities; others may be crowd-sourced.

While the crowd-sourcing information in this way may seem to contradict our data-provenance requirement, the crowd need not be a nameless set of individuals. Professional communities could collaborate to produce high quality data that they can share through such a system, such as scientists collecting and analyzing data about ecosystems [22], or coffee professionals assembling databases about farms and cafes worldwide [11]. The expertise of these communities might be confirmed by a trusted third party, much like Twitter's "verified user" system; alternatively, user rankings, recommendations or redundant verifications may be used to support sources claims to authority. The project Old Weather [20], for example, successfully crowd-sourced the data entry of naval weather logs by having each scanned record transcribed by multiple users.

With such systems in place, a given community should be able to easily identify coverage gaps and instances where their data quality needs improvement, as well as opportunities to resolve semantic heterogeneity. As these issues are identified and broadcast to the network, the community should be able to go about addressing these issues in a collaborative fashion. Because such a community is comprised of motivated individuals (of many levels of expertise and cost), data collection could be done much more efficiently and effectively than it is today.

5 Conclusions

The latest generation of data management tools has already begun to revolutionize journalism in the 21st century, both in concept and in practice. From the relatively complex, static, and siloed data-manipulation and publishing tools of traditional computer-assisted reporting, the accessible, flexible and networked tools of recent years makes the reach of data journalism virtually as broad as the Web itself. At its core, however, data journalism embodies both the history and the future of the journalistic endeavor. As Pulitzer Prize-winning data journalist Mo Tamman puts it:

"Our job as journalists, more so now than ever, is to help people make sense of what's going on around them. There's so much noise out there it's deafening. Our role is to help them make sense in this deafening noise. It's what we've always done."

References

[1] J. Ball and P. Lewis. Twitter and the riots: how the news spread. *The Guardian*, 2011.

[2] L. Barbosa and J. Freire. Combining classifiers to identify online databases. In *WWW*, pages 431–440, 2007.

[3] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[4] J. Burn-Murdoch, P. Lewis, J. Ball, C. Oliver, M. Robinson, and G. Blight. Twitter traffic during the riots. http://www.guardian.co.uk/uk/interactive/2011/aug/24/riots-twitter-traffic-interactive, 2011.

[5] 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.

[6] Court of Appeals, 9th Circuit 2011 No.09-16676 (Dec. 29)). Hepting v. AT&T. https://www.eff.org/sites/default/files/filenode/20111229_9C_Hepting_Opinion.pdf, 2011.

[7] danah boyd and K. Crawford. Six provocations for big data. In *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.

[8] Fusion tables gallery. https://sites.google.com/site/fusiontablestalks/, 2012.

[9] 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.

[10] Google refine. http://code.google.com/p/google-refine/, 2011.

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

[12] J. Halliday. David cameron considers banning suspected rioters from social media. *The Guardian*, 2011.

[13] P. S. John. Insurers risk of ruin. *Sarasota Herald-Tribune*, 2010.

[14] G. Lotan. Kony2012: See how invisible networks helped a campaign capture the worlds attention. http://blog.socialflow.com/post/7120244932/data-viz-kony2012-see-how-invisible-networks-helped-a-campaign-capture-the-worlds-attention, 2012.

[15] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Y. Halevy. Google's deep web crawl. *PVLDB*, 1(2):1241–1252, 2008.

[16] K. Morton, R. Bunker, J. D. Mackinlay, R. Morton, and C. Stolte. Dynamic workload driven data integration in tableau. In *SIGMOD Conference*, pages 807–816, 2012.

[17] No. 10 Civ. 3488 (SAS) (Feb. 7). National day laborer organizing network, et al. v. U.S. Deptartment of Immigration & Customs Enforcement Agency, et al. http://www.ediscoverycaselawupdate.com/National.pdf, 2011.

[18] NOAA. National weather data in kml/kmz formats. http://www.srh.noaa.gov/gis/kml/, 2012.

[19] NYPD. NYPD stop, question and frisk database. http://www.nyc.gov/html/nypd/html/analysis_and_planning/stop_quest 2011.

[20] OldWeather. Old weather: Our weather's past, our climate's future. http://www.oldweather.org/, 2011.

[21] Open internet tools project. http://openitp.org/, 2011.

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

[23] Scraperwiki. https://scraperwiki.com, 2010.

[24] I. T. Union. Internet users 06-11. http://www.itu.int/ITU-D/ict/statistics/material/excel/2011/Internet_users_ 01-11.xls, 2011.

[25] 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.

[26] F. B. Viégas, M. Wattenberg, F. van Ham, J. Kriss, and M. M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1121–1128, 2007.