# Open Civic Data: Of the People, By the People, For the People

Arnaud Sahuguet arnaud@thegovlab.org
John Krauss john@thegovlab.org
Luis Palacios luis@thegovlab.org
David Sangokoya david@thegovlab.org

The Governance Lab @ NYU
2 MetroTech Center, Brooklyn, NY 11201, USA

**Abstract**

*"Software is eating the world", says Marc Andreessen, with data as its fuel and its by-product. Inspired by the success of various open movements, data is now getting open as well. At the forefront, governments and cities are releasing a trove of civic data with promises of better – data-driven, collaborative and participatory – forms of governance. In this paper, we provide* a *definition of Open Civic Data and motivate what makes it special. We present an overview through stories from the field. We look at current technical barriers; future trends and challenges; and hint at how database research can and should contribute.*

## 1 Introduction

"Software is eating the world", says Marc Andreessen [1]. Data is both its fuel and its by-product. The value of data is a given. In the private sector, fortunes have been made by small and large corporations leveraging big data: Google for search and advertising; Facebook, Twitter and LinkedIn for social media; Amazon for retail. Data has become a competitive advantage and a carefully guarded asset.

The civic sphere – broadly and loosely defined for now – is eager to join the bandwagon. With more than 60% of the world population living in an urban environment by 2050, smarter cities [2] have become a necessity. Complex public problems (climate change, health, transportation, employment, education, etc.) require an effective collaboration between public and private entities. And this kind of collaboration can only happen if the data is made available in order to describe the problem and measure the impact of a solution.

Inspired by the success of various "open movements" – open source for software, open content for creative work, open access for scholarly research, open education for teaching materials – public players have embraced open data as a way to achieve this goal. The White House Executive Order [3] and G8 Open Data principles [4] are just two examples (see Table 1).

"Open data is data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike" [5]. According to McKinsey [6], this can translate into $3 to $5 trillion in economic impact. Open data also creates large opportunities for innovation [7] and social impact [8]. But so far, government – at all levels – has been the main steward for open data using its organizing power as a stick rather than a carrot.


**Bulletin of the IEEE Computer Society Technical Committee on Data Engineering**

| | | |
|---|---|---|
| 1 | Open Data by Default | *"To promote continued job growth, Government efficiency, and the social good that can be gained from opening Government data to the public, the default state of new and modernized Government information resources shall be open and machine readable."* |
| 2 | Quality and Quantity | |
| 3 | Usable by All | |
| 4 | Releasing Data for Improved Governance | |
| 5 | Releasing Data for Innovation | |

Table 1: G8 Open Data Principles (L); White House Executive Order (R).

We define *Open Civic Data* as a subset of open data that either originates from civic sources or is applicable to civic purposes. We also take a very broad definition of *civic* to embrace all things related to the common good or about "doing together what we cannot do alone". In such an open system, "the public becomes part of the data processing system and might process data, enrich data, combine it with other sources, and might even collect their own data" [9].

What makes Open Civic Data different and interesting is the fact that this is data about us that, if handled properly, can have a huge impact on our lives and our immediate environment. Paraphrasing words from a famous address[1], we see Open Civic Data as data *of the people, by the people, for the people*.

Open data as a research topic is still in its infancy, mostly based on anecdotes [10], interviews with practitioners [7], qualitative assessments of practices, descriptive surveys of datasets [11], best practices [12] and decision frameworks [9, 13, 14]. Eckartz et al. [13] propose a decision process that involves answering the questions of ownership, privacy, economic value, data quality and technical aspect. Janssen et al. [9] present an exhaustive list of benefits (political & social, economic, operational & technical) and barriers (institutional, task complexity, use & participation, legislation, information quality, technical) for open data.

Our focus here is on the technology component of open data, which usually goes beyond the narrow technical dimensions mentioned above. Like our fellow researchers, we start from stories from the field to catalog technical barriers and look at future challenges.

The rest of this paper is organized as follows. In Section 2, we present some stories from the field to illustrate examples of Open Civic Data and its applications. In Section 3, we review existing barriers and show how database innovations can address some of them. Section 4 is more speculative and focuses on Open Civic Data 2.0 with related research problems. We finish by presenting our conclusions.

## 2 Stories from the field: Open Civic Data 1.0

"In God we trust; everyone else, bring data" tweeted[2] Mike Bloomberg in 2010, then mayor of New York City, in the context of his campaign against smoking. And leveraging the city open data portal, the quantitative analyst behind IQuantNY [15] investigates city issues such as noise, most blocked driveways, taxi cab tips, the worst places to swim or the most-money-making fire hydrant.

In this section, we present a few stories from the field organized around (1) quality of life, (2) government-to-government, (3) economic development and (4) crisis management.

### 2.1 Quality of life

"Governments should concentrate on the three B's: Buses for transit data, Bullets for crime data and Bucks for budget & expenditure data," said Mark Headd, former chief data officer for the city of Philadelphia. Bullets

[1] Abraham Lincoln, Gettysburg, Pennsylvania November 19, 1863.

[2] https://twitter.com/mikebloomberg/status/114493100541489152

and Bucks are sometimes controversial topics, but Buses – transit data – is a clear success story with hundreds of mobile apps released globally, millions of users and millions of minutes – and maybe dollars – saved every day. The Google-initiated GTFS standard, pioneered in Portland, OR and now used by thousands of cities, was instrumental in creating transit maps and mobile applications.

The EU-funded CitySDK project exposes city data like points of interest for tourists via harmonized APIs, with deployments in Helsinki, Amsterdam and Lisbon. The CityByk.es project out of Barcelona aggregates information about hundreds of bike-sharing programs across the world and powers numerous mobile applications.

Through a partnership between the City of San Francisco and Yelp!, food enthusiasts access restaurant hygiene scores from their mobile phone, with a standardization of such scores in the making.

In Brazil, citizens of Recife can find the closest health-unit based based on their location [16].

All of these examples demonstrate the great benefits of smarter access to information as a way to improve directly or indirectly people's lives. The GTFS data standard is particularly interesting. Four components contributed to its success: (a) good timing, with lack of existing standard at the time; (b) ease of publishing of transit data for cities with a simple CSV-based standard; (c) a high-value use-case for people with a problem – transit – they face usually more than once a day; and (d) a ubiquitous delivery channel, Google Maps. Now everyone is looking for "the next GTFS."

## 2.2 Government-to-Government

Surprisingly, the biggest beneficiaries of government opening its data are actually governments themselves.
The adage that "the left hand doesn't know what the right hand is doing" is often true because of heavily siloed data. Open data helps cities spend less, while generating new revenue and allocating their resources better. In Philadelphia, because data from the Department of Revenue and the Department of Licenses & Inspections is not shared, the city issues building permits and rental licenses to entities who are late with their taxes. This situation amounts to tens of millions in lost revenue. In Baltimore, the city issues permits for occasional events, e.g., street fairs, which sometimes require the presence of additional police forces to guarantee safety. Because the police department does not have access to the calendar of events, assignments are made at the last minute, which translates into overtime and extra cost for the city. Opening the data (a) would let the city know about the late taxes and deny the permits, encouraging tax payment and generating revenue; and (b) would let the police department anticipate the need for extra staff, with likely reduced need for overtime.
In both cases, time and money would be saved.

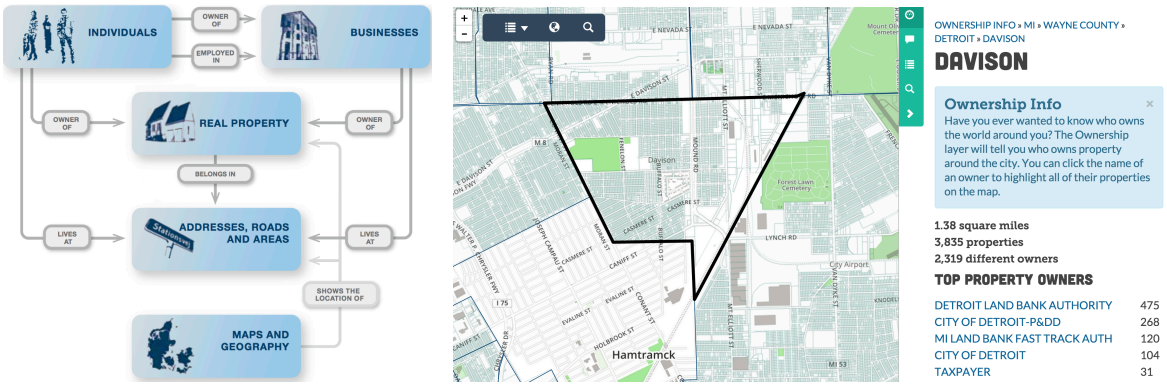
In New York City, leveraging open data, "... the Analytics Team has used data-based targeting to isolate instances of prescription drug abuse, discover cases of mortgage fraud and identify businesses operating with expired licenses [...] worth more than a combined $200 million."

In New Orleans, open data about blighted properties – due to hurricane Katrina in 2005 – lets the city and private partners prioritize which block to renovate and which to demolish.

The fear of being judged, data turf wars, and plain old politics are often barriers to open data in government. But when it opens its data, government highly benefits from it, via cost reduction, better resource allocation and also new sources of revenue. Opening data and seeing others make impactful use of it is a good ice-breaker for future conversations and collaborations.

## 2.3 Economic Development

Census bureaus worldwide have a long tradition of releasing datasets of key socio-demographics for their population. These datasets are heavily used by the private sector when deciding on investment, in such industries as real-estate, insurance, retail and entertainment. Denmark has identified (a) individuals, (b) businesses, (c)




(a) (b)

Figure 1: (a) What is "Basic Data"? [17] (b)"Why Don't We Own This?" in Detroit.

real estate properties, (d) addresses and (e) maps as "Basic Data" that needs to be open to foster economic development (Figure 1(a)).

The UK's Open Address project is creating a free, fresh and accurate database of all addresses in the UK, to remove the legal and cost barriers of using geographic data in the kingdom.

In Detroit, due to the economic downturn, the city lost a large chunk of its population, with many houses being abandoned. The city had to file for bankruptcy with no clear idea of what it owns. Through crowdsourced efforts, a survey of all properties was conducted and the data made open (Figure 1(b)). This helped to reboot the real-estate market.

In France, the OpenFisca project uses open data from the revenue service to provide an open source micro-simulator of the tax-benefit system. Users and corporations can use it to calculate social benefits and taxes paid. Simulations can also be run at the level of an entire population to estimate the consequences of new tax regulations.

The Danish example shows the importance of opening strategic datasets that build the foundation for more open data. This is not surprising: in the mobile space, location-based applications started to flourish when geo-data was made available via APIs such as Google Maps.

## 2.4 Crisis Management

Open data also plays a key role before times of crisis (preparedness, resilience planning), during and after (reconstruction).

Ushahidi pioneered the use of crowdsourcing and open data in the context of the disputed 2007 election in Kenya. The open source platform has been widely adapted and used in over 150 countries since.

In 2014 the UN Office for the Coordination of Humanitarian Affairs (OCHA) launched the Humanitarian Data Exchange portal as an open platform for UN agencies, NGOs and government departments to share data on disaster response. The platform has more than 1,300 datasets that can be compared across countries and crises and is currently being used in the Ebola crisis in West Africa.

Before Hurricane Sandy, open datasets about flooding zones and evacuation routes were used to inform residents whether they should evacuate and optimal routes towards shelters and food distribution centers.

In time of crisis, it is crucial to be able to bring "all hands on deck". Open datasets are key for collaboration between sectors and across country boundaries.

### 2.5 Horror Stories

Open data has also witnessed its share of horror stories.

A recent decision from the Canadian government to cancel the long form census [18] has created a data vacuum that is damaging research, city planning and also the private sector.

Outside of the civic sphere, a seminal example was the release of search logs by AOL for academic research purposes in 2006 [19]. The released data contained some personally identifiable information (PII) on AOL members that made it possible to identify people and reveal their Internet search histories [20]. On the consumer side, advances of Big Data are not sending a reassuring message with the example of Target who can figure out that a teenage girl was pregnant before her father did [21].

Because open civic data is more recent and governments opening their data are extremely risk and publicity averse, we have less examples to choose from. In New York City, taxi data released by the Taxi & Limo Commission was not anonymized properly. As a result, it was used to infer the identity of taxi drivers; in some cases, trips of celebrities were also reconstructed [22].

A few other areas prone to such stories are budget, land and health. Because of its very complicated nature, budget data can be represented in many ways and published data does not always reflect real-time adjustment and modifications. Land data is very tricky to define properly. Health data contains lots of private information.

We should expect more such stories because it is very hard to anticipate how disparate datasets will be combined. Very recently, a curious data enthusiast ran an algorithm combining taxi trip data (time, location) with prayer starting times to guess the religious orientation of some drivers [23].

Open Civic Data success stories are centered around data packaged in actionable intelligence for people, data with strong economic value – either for the government or the private sector – and cases where lives are at stake. There are few horror stories so far because data has been released haphazardly. Also, bigger stories linked to abusive state surveillance have probably eclipsed them. Another fact not to be neglected is the digital divide. By opening the data, we should not create two classes of open data users: the haves – who can get the data and understand it – and the have-nots.

For more examples of open data in action, see Open Data Census, the Open Data Barometer, government using GitHub for data or [24]. Data-Smart City Solutions at Harvard University provide a rich catalog of cases for such categories as civic data, civic engagement, health & human services, infrastructure, public safety, regulation and the responsive city.

## 3 Technical barriers

There are many barriers to open data. Janssen et al. [9] identify the following categories: institutional, task complexity, use & participation, legislation, information quality and technical. In this section, we look at the technical components of these barriers and relate them to the field of data management.

### 3.1 What makes civic data special?

Civic data relates to people's behavior and can therefore be very personal, especially with some obvious privacy issues such as personally identifiable information (PII). By its civic nature, the data is big, increasingly real-time and more and more mediated by the "Internet of Everything" [25]). It encompasses data (e.g., crime stats), text (e.g., legislation), audio (e.g., gunshot-detecting microphones) and video (e.g., traffic cameras, CCTV). It also spans across geo-spatial, temporal and social dimensions. Often times, the data starts at the analog level and needs to be digitized (with the cost and quality loss this can imply); or the data comes from legacy systems

| | |
|---|---|
| Opening | acquisition<br>modeling<br>transformation |
| Publishing | anonymization<br>quality<br>provenance & versioning |
| Using | discovery<br>visualization<br>query & processing |

Table 2: Barriers when opening the data.

where export is also costly. For accountability and transparency reasons, data needs to be versioned and each change documented.

Civic data is often only one part of the equation and requires integration with other non-civic data sources or data owned by a different governmental level (e.g., city vs state). This integration can prove to be difficult, depending on the nature of the data owner itself. Cities are complicated entities with hundreds of years of civic data, a plethora of agencies with sometimes misaligned incentives and a procurement system that makes it hard to bring in innovative solutions [26].

Last but not least, civic data is of interest to a very large audience with different skills and different needs: from the journalist, the investor, the regulator, the decision maker, the consumer, the citizen or the activist. As data is being requested by any of these diverse parties (e.g., through a Freedom of Information Act (FOIA) request in the US), its structure might be different.

### 3.2 Opening the data

Opening civic data costs money. Siloed or legacy data requires special handling. Data needs to be kept fresh. Data needs to follow legal requirements. All of this translates into human and software costs. So far, the stick has been more prevalent than the carrot in terms of opening data with more legal mandates (e.g., Obama's Executive Order [3] and FOIA requests) than scientific studies showing the positive impact of open data. If not done properly, opening data can create legal risks for cities. The most relevant datasets (the 3 B's mentioned in Section 2) are often the most controversial from a political point of view at the local level (turf-war between agencies, political strife between candidates) and beyond (cities competing to attract people, business and funding based on their core civic metrics). As a result, the opening of civic data often optimizes for ease, cost, liability and political kudos rather than positive impact.

A key challenge when opening the data is data modeling. The "runtime model" (used inside a database) might not be the same as the "publishing model" (used on an open data portal). There is no agreed-upon standard for civic data: no preferred format, no pre-defined schemas, globally unique identifiers and taxonomies for categories of civic data. Nor is there a preferred way to publish a data catalog. For instance, the city of Boston uses two different schemas for its summer and winter farmers market. New York City, Portland and Chicago each have each their own different ways of publishing their police precinct data. See Barbosa et al. [11] for an in-depth study of civic datasets.

Getting the data itself is often problematic. Civic data often originates from decades-old legacy systems. Water data from the City of Baltimore needs to be exported from an old mainframe. Data is often siloed and requires proper integration before it is published. The first step in the process often consists in agreeing on a unique

identifier to join separate datasets. Lots of open data work in New York City was done around Boro-Block-Lot (BBL[3]) identifiers.

There are some emerging efforts to address these issues, e.g., Sunlight Foundation's OCDIDs, Schema.org GovernmentService and the UK's ESD standards. Database research on information extraction, data integration, schema mapping and ETL languages is very relevant in this context.

### 3.3 Publishing the data

Access-control and anonymization are critical. Civic data is personal by nature because it describes human behavior. Taxi data from the New York City TLC commission was not sufficiently anonymized (see Section 2). It is also extremely difficult to anticipate how datasets – including external ones – will be combined and guaranteeing that no information will be leaked or inferred is very challenging.

"[...] a city is more than a place in space, it is a drama in time." said P. Geddes in [27]. Civic data keeps evolving. Versioning and provenance of raw datasets can be preserved relatively easily at publishing time using version control systems like GitHub. It is harder when datasets are combined across versions. Also, open data can be reused and modified freely; therefore it becomes very hard to keep track of the various transformations the data went through.

Data quality is also a critical element. In the context of crisis, bad data or stale data can lead to lost lives. The US Department of Commerce imposes the following quality requirements in terms of data: "comprehensive, consistent, confidential, trustworthy and available to all."[4].

Techniques like k-anonymity [28] and differential privacy [29] can of course be applied, but (1) one has to be aware of the issue and (2) one needs the tools to process the data. Practical and theoretical work on data provenance [30, 31, 32] and data quality [33] are also extremely relevant.

### 3.4 Using the data

Usability of the data is a principle in the G8 Open Data charter [4] and a requirement for innovation.

Publishing the data is just the beginning. First, it needs to be discovered. There is no agreed-upon schema to describe data catalogs (DCAT, VoID, etc.). The best way to have something discoverable on the Web so far is to have it indexed by Google.

Once discovered, data needs to be downloaded. Most data portals provide only bulk access to the datasets, often with no data compression and no way to retrieve only a subset of the data. New York City taxi datasets are notorious for taking forever to download. Datasets often come with no metadata or data dictionary. Datasets often require other datasets to be downloaded. A civic hacker often has to find and fetch the dataset, repeat the process for other datasets it depends on create a relational schema for each dataset and finally load the data into a database management system, before she is able to use the data.

As mentioned before, civic data is geo-spatial, temporal and social, which makes it hard to represent on traditional three dimensional interfaces. Canned solutions often do a poor job.

Innovative approaches like TaxiVis [34] leverage database indexing techniques to provide rich and efficient queries and data visualization over large amounts of multidimensional data. Commercial GIS solutions from ESRI or Google Earth Engine address the very common use case of spatial data.

[3]Borough-Block-Lot (BBL) or parcel numbers identify the location of buildings or properties.

[4]`http://www.commerce.gov/blog/2014/06/19/listening-our-data-customers-open-data-roundtable`

### 3.5 From the commercial side

Most vendors focus on one-size-fits-all solutions that do not address the specificities of open data. The IBM SmartCities initiative focuses on: "Do more with less," "Bridge silos in information and operations," "Use civic engagement to drive better results" and "Invest in infrastructure for better management". CISCO's emphasis is more on "connecting people, process, data, and things" with an Internet of everything for cities [25].

Unfortunately, the most popular open data solutions such as CKAN, Socrata or GitHub, offer very limited data management capabilities: no versioning (except for GitHub), no schema validation, and no triggers. None of them allows you do joins between datasets [35].

Civic Open Data today has a lot of challenges, whether you are a producer or a consumer. Fortunately, 20+ years of database research provide some answers to most of them, at least on paper. But these answers have to be propagated to data curators, commercial vendors and civic data scientists. A solution might be for these innovations to be unbundled, packaged for easy deployment and integration with existing solutions, including open-sourced ones. A good analogy would be to do for data what Docker is doing for software. A "Docker for Open Civic Data"[5] would package various datasets, their schema, a fitting datastore, some relevant views that represent the most common facets of the data and a ready-to-go user friendly front-end to let people start "playing with the data" right away.

## 4 Open Civic Data 2.0 & Challenges

A meta-trend we see at the governance level [36] is a push for more data-driven, collaborative and participatory forms of governance. In terms of open data, this means governments pushing the envelope with data, new forms of crowdsourcing for data creation and also the emergence of citizen-science targeted at governance.

In this section we look at where Open Civic Data is going and some of the domain-specific challenges that will emerge.

### 4.1 Richer structure

First, location and time have to be first-class data citizens, at the storage, query and visualization levels.

As civic data gets more interconnected and more social, we need richer models in terms of structure and semantics. Schema.org and ESD standards in the UK are already using Linked Data.

Civic data is social and connected by nature with annotations, comments and relationships between entities. OpenCorporates compiles a map of relationships between corporations, e.g. holdings, subsidiaries, etc. It also provides unique identifiers and taxonomies to help build knowledge on top. The goal of orgPedia project is to map corporate entities to possible labor, environment or export violations.
The natural structure for such data is a graph database, which presents both storage and query challenges [37, 38, 39].

### 4.2 From descriptive to predictive, with high quality and concrete metrics

The current wave of Open Civic Data focuses mostly on the descriptive aspect of running a city with reports, dashboards, etc. The next wave will be about predictive analytics.

City managers will anticipate crime before it happens or dispatch emergency response services to reduce response time. Citizens will access information about the best location for parking or the next train with room to sit, e.g. La Tranquilien in Paris.

[5] Docker4Data project

Such applications will require models to be computed via machine learning techniques and also built-in optimization algorithms. Database architectures like LogicBlox [40] combining OLTP, OLAP, machine learning and optimization around a single datastore could provide a very appealing solution.

All of this assumes a very high level of quality for the data. As we mentioned before, quality is often the reason why data is not made available in the first place. Tools to clean the data but also metrics to measure the level of quality of the data are needed.

### 4.3 Social civic data & civic data science

The next wave of Open Civic Data will be more crowdsourced [41] at the production, selection and consumption levels. Two concrete emerging use cases are population informatics [42] (aka social genome) and precision medicine [43].

More and more data will come from the people directly either via apps like FixMyStreet or Waze, or mediated via wearables or connected sensors [25]. Open data portals need to accommodate individual data contribution and handle issues such as data quality, duplicated data and spam for user contributed content. Research on crowdsourcing [44] and data privacy & anonymization [29, 45] is highly relevant.

People will also want to have their say on which datasets should be made open using legal tools like the Freedom Of Information ACT (FOIA) in the US. Open data portals need to accommodate voting and prioritization to help decision makers pick the next dataset to open.

Tapping into the reservoir of expertise – from citizens, private sector or simply other agencies – implies making Open Civic Data easy to access, process and understand. It also means bringing together the people with data and problems and the people with knowledge and solutions. DataKind, Bayes Impact and Kaggle are examples of such data science marketplaces.

Database management and data science are becoming the two faces of the same coin as recently discussed by Howe et al. [46]. Visualization tools like Tableau or interactive data science frameworks like Trifacta or iPython notebook are good steps towards tools that provide "an immediate connection" to what people look for in the data, as described by Victor [47]. The "Docker-for-data" model mentioned in Section 3 and the PC-AXIS file format are along those lines. The dataset-centric nature of open data makes data immutability a new "inexorable trend" as noted by Helland [48].

Because of a push for more data-driven, collaborative and participatory decision-making and governance at the city level, we should expect Open Civic Data to reflect these changes. This means richer and more social data, contributed directly by people or mediated by sensors. This also means making data more directly accessible and actionable to people in order to tap into their collective intelligence.

We are closing this section by offering a list of 10 "impossible queries" collected through various conversations with people in the field. Answering such queries will require data that does exist, data that exists but cannot be easily joined, query languages that make it easy to express such queries and more.

1. List companies in this geographic area with more than 5 labor infractions?
2. What's the relationship between company X and elected official Y?
3. What are the best default locations for ambulances during the heat season?
4. Where should I send my building inspectors first?
5. What blighted blocks should be demolish and why?
6. What's the impact of releasing (resp. removing) dataset X?

7. Which gas stations are doing price gouging during the current hurricane?
8. What's the impact of car sharing services?
9. Can people affected by flooding afford the surge of insurance premium in areas at risk?
10. Your query here

# 5 Conclusion

Sixty percent of the world population will live in cities by 2050. Better decision-making at the city level is critical to have a positive impact. And Open Civic Data is key.

Civic data is not just data: it is data about us and how we operate, among ourselves and within our environment. Opening the data is not just publishing the data: it means making the public part of the system and tapping into its wisdom to drive actionable decision and impactful outcomes.

There are lots of barriers, some of them technical; and database research can help. There are also some upcoming challenges. These are opportunities for great research. But whatever good research produces, we have to make sure it is properly packaged and can be consumed and incorporated into the the current work streams and work flows used by people running cities and governments. We also have to make sure we do not create an even bigger digital divide for people and cities.

In the not so distant past, database researchers were forced to pick their paper's "motivating example" based on availability of the data – rather than an application domain they care about or makes sense for their research – or to rely on made-up data. With Open Civic Data, this time is over.
Open Civic Data gives our field a chance to apply years of database research – past, present and future – to solve new challenging problems and to have a concrete impact on society and our environment.

# Acknowledgments

The authors would like to thank Michael Flowers, Alberto Lerner and the Data Engineering Bulletin editors for comments on early versions of this paper.

# References

[1] M. Andreessen, "Why software is eating the world'," *Wall St. J.*, vol. 20, 2011.
[2] A. M. Townsend, *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. 2013.
[3] B. Obama, "Executive Order-Making open and machine readable the new default for government information," *Whitehouse.gov*, vol. 9, 2013.
[4] U. K. P. o. G. Cabinet Office, "G8 open data charter and technical annex," in *Open Government Partnership Summit*, 18 June 2013.
[5] D. Dietrich, J. Gray, T. McNamara, A. Poikola, P. Pollock, J. Tait, and T. Zijlstra, "Open data handbook." `http://opendatahandbook.org`, 2009.
[6] J. Manyika, *Open data: Unlocking innovation and performance with liquid information*. McKinsey, 2013.
[7] A. Zuiderwijk, N. Helbig, J. R. Gil-García, and M. Janssen, "Special issue on innovation through open data: Guest editors' introduction," *J. Theor. Appl. Electron. Commer. Res.*, vol. 9, no. 2, pp. i–xiii.
[8] A. Howard, "More than economics: The social impact of open data," *Tech Republic*, 31 July 2014.
[9] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, adoption barriers and myths of open data and open government," *Information Systems Management*, vol. 29, no. 4, pp. 258–268, 2012.
[10] J. Gurin, *Open Data Now*. McGraw Hill Education, 2014.

[11] L. Barbosa, K. Pham, C. Silva, M. R. Vieira, and J. Freire, "Structured open urban data: Understanding the landscape," *Big Data*, vol. 2, pp. 144–154, 1 Sept. 2014.

[12] J. Tauberer, "Open government data." https://opengovdata.io/, 2012.

[13] S. M. Eckartz, W. J. Hofman, and A. F. Van Veenstra, "A decision model for data sharing," in *Electronic Government*, Lecture Notes in Computer Science Volume 8653, pp. 253–264, Springer Berlin Heidelberg, 1 Sept. 2014.

[14] A. Sahuguet and D. Sangokoya, "A "calculus" for open data." https://medium.com/p/p-b-d-c-1218ee894400, Feb 2015.

[15] B. Wellington, "I quant NY." http://iquantny.tumblr.com/.

[16] K. dos Santos Brito, M. A. da Silva Costa, V. C. Garcia, and S. R. de Lemos Meira, "Brazilian government open data: Implementation, challenges, and potential opportunities," in *Proceedings of the 15th Annual International Conference on Digital Government Research*, dg.o '14, (New York, NY, USA), pp. 11–16, ACM, 2014.

[17] The Danish Government, "Good basic data for everyone - a driver for growth and efficiency." http://goo.gl/4V6ZPs, 2012.

[18] T. Grant, "Damage from cancelled census as bad as feared, researchers say." The Globe and Mail, Jan. 2015.

[19] M. Arrington, "AOL proudly releases massive amounts of private data," *TechCrunch: http://www. techcrunch. com/2006/08/06/aol-proudly-releasesmassive-amounts-of-user-search-data*, 2006.

[20] D. Kawamoto and E. Mills, "AOL apologizes for release of user search data," *CNET: http://news.cnet.com/AOL-apologizes-for-release-of-user-search-data/2100-1030_3-6102793.html*, 2006.

[21] K. Hill, "How target figured out a teen girl was pregnant before her father did," *Forbes, Inc*, 2012.

[22] C. Gayomali, "NYC taxi data blunder reveals which celebs don't Tip–And who frequents strip clubs," Oct, year = 2014, url =.

[23] A. Berlee, "Using NYC taxi data to identify muslim taxi drivers." http://theiii.org/index.php/997/using-nyc-taxi-data-to-identify-muslim-taxi-drivers/, Jan 2015.

[24] Sunlight Foundation, "The Impacts of Open Data," tech. rep., Sunlight Foundation, 2014.

[25] S. Mitchell, N. Villa, M. Stewart-Weeks, and A. Lange, "The Internet of Everything for Cities," 2013.

[26] J. Bessen, "The Anti-Innovators: How special interests undermine entrepreneurship." http://www.cfr.org/united-states/anti-innovators/p35910, 2015.

[27] P. Geddes, *Cities in Evolution*. Londres, William & Norgate LTD, 1915.

[28] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, (New York, NY, USA), pp. 49–60, ACM, 2005.

[29] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth, "Differential privacy: An economic method for choosing epsilon," in *IEEE 27th Computer Security Foundations Symposium, CSF 2014, Vienna, Austria, 19-22 July, 2014*, pp. 398–410, 2014.

[30] P. Buneman, S. Khanna, and T. Wang-Chiew, "Why and where: A characterization of data provenance," in *Database Theory - ICDT 2001*, Lecture Notes in Computer Science Volume 1973, pp. 316–330, Springer Berlin Heidelberg, 2001.

[31] T. J. Green, G. Karvounarakis, and V. Tannen, "Provenance semirings," in *Proceedings of the Twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '07, (New York, NY, USA), pp. 31–40, ACM, 2007.

[32] R. Ikeda and J. Widom, "Panda: A system for provenance and data," *IEEE Data Eng. Bull.*, vol. 33, no. 3, pp. 42–49, 2010.

[33] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma, "Improving data quality: Consistency and accuracy," in *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, (Vienna, Austria), pp. 315–326, VLDB Endowment, 2007.

[34] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual Exploration Of Big Spatio-temporal Urban Data: A Study Of New York City Taxi Trips," *IEEE Transactions on Visualization and Computer Graphics archive Volume 19 Issue 12*, vol. 19, pp. 2149–2158, Dec. 2013.

[35] M. Headd, "I hate open data portals." http://civic.io/2015/04/01/i-hate-open-data-portals/, Apr 2015.

[36] B. S. S. Noveck, *Wiki Government: How Technology Can Make Government Better, Democracy Stronger, And Citizens More Powerful*. Brookings Institution Press, 2009.

[37] P. T. Wood, "Query languages for graph databases," *SIGMOD Record*, vol. 41, pp. 50–60, Apr. 2012.

[38] J. Seo, J. Park, J. Shin, and M. S. Lam, "Distributed socialite: A datalog-based language for large-scale graph analysis," *Proceedings VLDB Endowment*, vol. 6, pp. 1906–1917, Sept. 2013.

[39] E. Yoneki and A. Roy, "Scale-up graph processing: A storage-centric view," in *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, (New York, NY, USA), pp. 8:1–8:6, ACM, 2013.

[40] LogicBlox, "Datalog for enterprise applications: from industrial applications to research." `http://datalog20.org/slides/aref.pdf`, 16 Mar. 2010.

[41] D. C. Brabham, *Crowdsourcing*. MIT Press, 2013.

[42] H.-C. Kum, A. Krishnamurthy, A. Machanavajjhala, and S. C. Ahalt, "Social genome: Putting big data to work for population informatics," *Computer*, vol. 47, no. 1, pp. 56–63, 2014.

[43] E. Hafen, D. Kossmann, and A. Brand, "Health data cooperatives - citizen empowerment," *Methods Inf. Med.*, vol. 53, pp. 82–86, Feb 2014.

[44] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the World-Wide web," *Commun. ACM*, vol. 54, pp. 86–96, Apr. 2011.

[45] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *SIGMOD Record*, vol. 33, pp. 50–57, Mar. 2004.

[46] B. Howe, M. J. Franklin, J. Freire, J. Frew, T. Kraska, and R. Ramakrishnan, "Should we all be teaching intro to data science instead of intro to databases?," in *Proceedings Of The 2014 ACM SIGMOD International Conference On Management Of Data*, pp. 917–918, ACM, June 2014.

[47] B. Victor, "Inventing on Principle." http://govlabacademy.org/coaching-programs.html, 2013.

[48] P. Helland, "Immutability changes everything," in *7th Biennial Conference on Innovative Data Systems Research (CIDR)*, 2015.