PUSH, PULL, AND SPILL: A TRANSDISCIPLINARY CASE STUDY IN MUNICIPAL OPEN GOVERNMENT

Jan Whittington, Ryan Calo, Mike Simon, Jesse Woo, Meg Young & Peter Schmiedeskamp†

ABSTRACT

Municipal open data raises hopes and concerns. The activities of cities produce a wide array of data, data that is vastly enriched by ubiquitous computing. Municipal data is opened as it is pushed to, pulled by, and spilled to the public through online portals, requests for public records, and releases by cities and their vendors, contractors, and partners. By opening data, cities hope to raise public trust and prompt innovation. Municipal data, however, is often about the people who live, work, and travel in the city. By opening data, cities raise concern for privacy and social justice.

This article presents the results of a broad empirical exploration of municipal data release in the City of Seattle. In this research, parties affected by municipal practices expressed their hopes and concerns for open data. City personnel from eight prominent

DOI: http://dx.doi.org/10.15779/Z38PZ61

† Jan Whittington is an Associate Professor of Urban Design and Planning in the College of Built Environments, University of Washington, and the Director of the Urban Infrastructure Lab. Ryan Calo is an Assistant Professor in the University of Washington School of Law, and Director of the University of Washington Tech Policy Lab. Mike Simon is Chief Technology Officer for Creation Logic, LLC. Jesse Woo is a Corporate Attorney in Berkeley, California and a Consultant at the University of Washington Tech Policy Lab. Meg Young is a Ph.D. Student in the Information School, University of Washington. Peter Schmiedeskamp is an Interdisciplinary Ph.D. Student in Planning at the University of Washington.

This project was conducted in partnership with the City of Seattle. The authors acknowledge Michael Mattmiller, Ryan Biava, Ginger Armbruster, Bruce Blood, and the many additional employees, residents, and business representatives of the City of Seattle, who generously gave their time to participate in this research. For their comments on this study, the authors would also like to thank the participants of the 19th Annual Berkeley Center for Law & Technology and Berkeley Technology Law Journal Symposium, Open Data: Addressing Privacy, Security, and Civil Rights Challenges, held on April 17, 2015 and Responsible Use of Open Data: Government and the Private Sector, held at New York University on November 19–20, 2015, co-organized by BCLT and NYU's Information Law Institute and Department of Media, Culture and Communication. This project was one of six funded, in part, by Berkeley Center for Law & Technology with a generous grant from Microsoft, with funding also provided by the City of Seattle.

departments described the reasoning, procedures, and controversies that have accompanied their release of data. All of the existing data from the online portal for the city were joined to assess the risk to privacy inherent in open data. Contracts with third parties involving sensitive or confidential data about residents of the city were examined for safeguards against the unauthorized release of data.

Results suggest the need for more comprehensive measures to manage the risk latent in opening city data. Cities should maintain inventories of data assets, produce data management plans pertaining to the activities of departments, and develop governance structures to deal with issues as they arise—centrally and amongst the various departments—with ex ante and ex post protocols to govern the push, pull, and spill of data. In addition, cities should consider conditioned access to pushed data, conduct audits and training around public records requests, and develop standardized model contracts to protect against the spill of data by third parties.

TABLE OF CONTENTS

I. INTRODUCTION

Cities hold considerable information, including details about the daily lives of residents and employees, maps of critical infrastructure, and records of internal deliberations. Cities are beginning to realize that this information has economic and civic value. The responsible release of city information can result in greater efficiency and innovation in the public and private sector. New services are cropping up that leverage open city data to great effect.[1] Activist groups and residents are also placing increasing pressure on state and local government to be more transparent.

There has been little research into the growing area of municipal open data.[2] Cities are beginning to open their data in a way that has never been seen before, and these releases may raise privacy concerns. Scholarly and media attention has focused at the federal level toward the activities of the National Security Agency (NSA), the Federal Trade Commission (FTC), and the White House.[3] Despite the attention given to federal agencies, most personally-identifiable data is collected much closer to home, by the governments of the cities where we live, work, and play.[4]

---

1. *See, e.g.*, Kathleen Hickey, *AppStore Gives Governments Access to Municipal Apps*, GCN (June 4, 2014), http://gcn.com/articles/2014/06/04/granicus-appstore.aspx; Angus Loten, *Entrepreneurs Shape Free Data into Money*, WALL ST. J., Jan. 9 2014; Jason Slotkin, *City Living: There's an App for That*, COMPUTERWORLD (Jan 11, 2013), http://www.computerworld.com/article/2494114/mobile-wireless/city-living--there-s-an-app-for-that.html; Geoffrey A. Fowler, *Apps Pave Way for City Services*, WALL ST. J. (Nov. 18, 2010), http://www.wsj.com/articles/SB10001424052748704658204575611143577864882.

2. For example, Maxat Kassen has observed:

> [I]t is not yet clear how the potential of the open data concept can be realized at the local level as there has been no analysis of current projects so far. The concept is still in its infancy, and in fact it gained a political meaning primarily after the launch of the official U.S. government data portal in 2009. Later, similar data projects were initiated at the local level.

Maxat Kassen, *A promising phenomenon of open data: A case study of the Chicago open data project*, 30 GOV'T INFO. Q. 508, 509 (2013); *see also* Anneke Zuiderwijk & Marijn Janssen, *Open Data Policies, Their Implementation and Impact: A Framework for Comparison*, 31 GOV'T INFO. Q. 17, 17 (2014) ("[V]ery little systematic and structured research has been done on the issues that are covered by open data policies, their intent and actual impact. Furthermore, no suitable framework for comparing open data policies is available."). As recently as 2011, the International City/County Management Agency national survey of e-Government did not include questions on open data. Donald F. Norris & Christopher G. Reddick, *Local E-Government in the United States: Transformation or Incremental Change?*, 73 PUB. ADMIN. REV. 165–175.

3. *E.g.*, DANIEL J. SOLOVE, NOTHING TO HIDE: THE FALSE TRADEOFF BETWEEN PRIVACY AND SECURITY (2011); Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014).

4. *See generally* Bill Schrier, *Chapter 28: Toads on the Road to Open Government Data*, *in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN

This Article is a cross-disciplinary assessment of an open municipal government system. We are a team of researchers in law, computer science, information science, and urban planning that worked hand-in-hand with the City of Seattle, Washington to understand its current procedures around data processing from each of our disciplinary perspectives. Based on this empirical work, we have generated a set of recommendations to help the city manage risk latent in opening its data.

Seattle makes for a great case study. With a population of 650,000 and growing rapidly, Seattle is mid-sized, but not so enormous as to be unwieldy. It is a highly educated, technically savvy city and is often highly ranked among its peers on measures of innovation, creativity, and technology.[5] Seattle was one of the first cities to embrace an open data initiative.[6] Its leadership has publicly stated a need to achieve a balance between privacy and transparency.[7] During our research, we found encouraging signs in what Seattle is already doing and its willingness to adopt best practices, and identified areas for additional improvement.

A. THE MUNICIPALITY IN FOCUS

Municipalities govern a wide array of activities, from police services to building permits to parks and recreational services and facilities. City governments collect and process large amounts of information to support these activities, often with the help of third party contractors. Some of this data is confidential, requiring special handling for security purposes, while other is not confidential, but nevertheless contains sensitive details about residents and employees. If taken out of context or made publicly available, this data could bring about harms to privacy or social equity.

Rapid technological changes pose significant complications for municipalities seeking to govern data in the public interest. Municipalities are eager to become "smart cities" by adopting information technologies

---

PRACTICE 305, 305–313 (Daniel Lathrop & Laurel Ruma, eds., 2010); Kassen, *supra* note 2, at 509; Peter Conradie & Sunil Choenni, *On the barriers for local government releasing open data*, 31 GOV'T INFO. Q. S10, S10–17 (2014).

5. *E.g.*, Boyd Cohen, *The 10 Smartest Cities In North America*, CO.EXIST (Nov. 14, 2013, 7:08 AM), http://www.fastcoexist.com/3021592/the-10-smartest-cities-in-north-america.

6. Press Release, Socrata, Inc., Socrata Strengthens Open Data Market Leadership (Jun. 28, 2011), http://www.socrata.com/newsroom-article/socrata-strengthens-open-data-market-leadership.

7. Press Release, City of Seattle Office of the Mayor, City of Seattle Launches Digital Privacy Initiative (Nov. 3, 2014), http://murray.seattle.gov/city-of-seattle-launches-digital-privacy-initiative.

that promise more effective and efficient delivery of services.[8] Ubiquitous computing includes mobile micro-video cameras, utility meters that discern the use of appliances, and technologies for detecting and tracking residents' whereabouts, energy use, and other information. Each of these technologies has the potential to create real-time, continuous data feeds. As the technologies of data collection, processing, and storage become ever more advanced and potentially intrusive, local governments face the challenge of adapting policies and guidance about privacy and social equity to changing circumstances. In the absence of clear criteria and procedures, municipal agents may resort to ad hoc decision-making. In a federated system of governance, the cumulative implications of multiple data releases may have consequences not anticipated by any individual unit, including the ability to reconstruct the identity of an anonymous resident.

The data generated by municipalities is of interest to many commercial entities, which seek to use the data for purposes that are not necessarily aligned with the public interest. In March 2014, the FTC published a report introducing the data-broker industry, which is built around the collecting, processing and reselling of data about individuals.[9] Brokers aggregate data from public and private sources, index the data into detailed profiles of persons, households, and neighborhoods, and sell it to private and public buyers. Eight of the nine data brokers participating in the FTC study reportedly relied on information supplied by government to identify and profile individuals.[10]

B. PURPOSE, THEMES, AND CONTENT

Our research explored both the mechanisms and consequences of municipal data releases. Our results provide a snapshot of activities and their

8. *See generally* Michael Batty, *Smart Cities, Big Data*, 39 ENV'T & PLAN. B: PLAN. & DESIGN, 191 (2012); Rob Kitchin, *The Real-Time City? Big Data And Smart Urbanism*, 79 GEOJOURNAL 1 (2014); Mike Weston, *'Smart Cities' Will Know Everything About You: How Can Marketers Cash In Without Becoming Enemies of the People?*, WALL ST. J., July 12, 2015, http://www.wsj.com/articles/smart-cities-will-know-everything-about-you-1436740596. Weston writes:

> [M]unicipalities and governments across the world are pledging billions to create "smart cities"—urban areas covered with Internet-connected devices that control citywide systems, such as transit, and collect data. Although the details can vary, the basic goal is to create super-efficient infrastructure, aid urban planning and improve the well-being of the populace.

Id.

9. FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY (2014), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf.

10. *Id.* at 15.

potential implications in a city that is striving to reap the benefits and avoid the pitfalls of data release.

Cities share data in three basic ways: push, pull, and spill. Cities "push" data when they publish databases through online or other portals. Residents and others "pull" data out of the city with public records requests. And cities "spill" data, through accidental exposure, malicious data breach, and the distribution of data by vendors, contractors, and partners. We use the push, pull, and spill taxonomy as a unifying theme throughout our analysis and recommendations.

Whether pushed, pulled, or spilled, the release of municipal data has many consequences. Three questions guided our exploration of the consequences of municipal data releases. Does the availability of open data increase public trust in the effective and efficient delivery of public services? Under what technological, legal, and other circumstances can municipalities govern the release of open data to meet the public need for privacy? What harms could municipal open data lead to, including issues of disparate racial or social impact, physical insecurity, or harm to consumers or the marketplace? We approach these questions across multiple methods and sections of this Article.

The rest of the Article proceeds as follows: We discuss our specific approach to investigating the city's use of municipal data in Part II. Part III summarizes our findings. Part IV consists of seven recommendations for Seattle—and other cities interested in improving open data practices. We recommend: (1) conducting an inventory of data assets, (2) requesting each department to submit a data management plan, (3) establishing nested governance structures to deal with issues as they arise, (4) establishing ex ante and ex post protocols for push, pull, and spill, (5) conducting an audit and training around public records requests, (6) exploring the prospect of conditioned access to some city data, and (7) developing a standardized model contract for data vendors. We understand that Seattle is actively pursuing some or all of these recommendations even as of this writing. Finally, the Article closes with Part V outlining future work suggested by our analysis and findings.

II. OUR APPROACH

There is little empirical work on municipal open data practices to date. However, exploratory research is not without guideposts. A sophisticated and expanding literature investigates the private sector's use of information technology. This literature builds theoretical and empirical accounts and examines how those uses may compromise social norms and features of the economy; features that are prefaced upon the privacy of personal

information, racial and social equity, and the preservation of the public trust in digital or online transactions.[11] This Article seeks to begin a similar line of research aimed at the public sector, starting with municipalities. As subjects of research, municipalities are recent entrants into an ongoing, multidisciplinary conversation about the benefits and pitfalls of data collection, use, release, retention, commercialization, and security. This characterization is especially apt when the aim of research is to orient policy to the public interest.

As the subject of this particular study is municipal open data, we focus on the release of data by or from municipalities.[12] The push, pull, and spill taxonomy assisted us in designing research that would explore current practices while highlighting the potential future effects of such practices on public trust, privacy, and social equity. This required a mixture of research methods, each suited to a likely area of contest or hazard.

Our research methods and findings are described in four parts:

- **Qualitative Assessment 1—Key Stakeholders:**
  We begin with a sense of the hopes and concerns of the parties affected by municipal practices. For this, we carried out focus groups on the topic of pushed, pulled and spilled municipal data, with several types of key stakeholders in the Seattle community. We relay our findings.
- **Qualitative Assessment 2—The City:**
  We then discuss how Seattle itself handles data. We conducted interviews with city personnel involved in the release of data. Interviews spanned push, pull and spill: the intended purpose and use of open data by departments, the circumstances of public disclosure requests, and the involvement of departments in

11. *See, e.g.*, Arvind Narayanan & Vitaly Shmatikov, *Privacy and Security: Myths and Fallacies of "Personally Identifiable Information,"* 53 COMM. ACM 24 (2010), https://www.cs.utexas.edu/~shmat/shmat_cacm10.pdf; Alessandro Acquisti & Jens Grossklags, *Privacy and Rationality in Individual Decision Making*, 3 IEEE SECURITY & PRIVACY, 26 (2005); Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010); Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995 (2014); Chris Jay Hoofnagle & Jan Whittington, *Free: Accounting for the Costs of the Internet's Most Popular Price*, 61 UCLA L. REV. 606 (2014).

12. Other stages in the lifecycle of data matter and, though not central to this study, are just as worthy of research. The results of this study suggest promising future avenues for research in these related areas, including, for example, the potential for upstream decisions about collection and retention to be predicated on the downstream effectiveness of policies restricting the uses of data.

contracts with third parties for information-intensive services. The results indicate the types of data collected and used, the incentives that departments have to release datasets (or not), and the ways in which releases are modified to preserve privacy and social justice.

- **Technical Assessment—Open Data Analysis:**
  We conducted technical analyses of the datasets already pushed to the City's open data portal in order to understand how the City uses the portal and to investigate the extent to which the City's current practices could potentially compromise privacy and social justice.
- **Legal Assessment—Vendor Contracts:**
  Having identified, in departmental interviews, many contracts with third parties involving sensitive or confidential data about residents of the city, we examined these contracts for the kinds of safeguards one might expect in order to prevent, for example, unauthorized spills of this data.

As a collection of exploratory assessments, these research activities provide a broad array of insights into the role of the municipality in the release of data.

III. FINDINGS

This part of the Article presents extensive findings on how a city generates and releases municipal data. This is a vast area for research. As other authors have explained, government departments are created to perform services that markets do not or should not provide, or are difficult or impossible for residents to provide for themselves.[13] For example, municipalities organize to provide regulatory functions to curb the many ways in which the for-profit, self-interested incentive structure of the private sector will "as if by an invisible hand" lead markets to fail to serve the public interest.[14] Within their jurisdiction, municipalities operate monopoly or monopolistic markets for several goods (e.g., water, electricity, roads, lighting), which are often provided through contracts with firms on behalf of residents. In negotiating these contracts, municipalities have

13. *See* Shrier, *supra* note 4, at 311.

14. *See id.* (quoting ADAM SMITH, AN INQUIRY INTO THE NATURE AND CAUSES OF THE WEALTH OF THE NATIONS 423 (Edwin Cannan ed., 1937) (1776)).

substantial leverage on the public's behalf, reducing the transaction costs that would have accrued if members of the public were left to organize and bargain on their own.[15] This bargaining power makes cities powerful market players—an untapped source of influence over privacy and security policy, as we discuss below. Municipalities also provide intergovernmental coordination: the geospatial area or jurisdiction of any given municipality is layered with the jurisdictions of several other governmental entities (e.g., special districts, counties, states, and the federal government). With such eclectic aims, municipalities can appear to be labyrinths of data production and release, bewildering in their complexity.

As a consequence of the enormity of the research task—as well as the inherent subjectivity in terms such as "open" or sensitive—we were forced to make certain assumptions and choices that we try to highlight through our findings. We also lay out an agenda for future work that reflects the realization that there is much more to do. Nevertheless, we attempted to convey and engage with both the breadth and depth of city data in our analysis.

Unlike physical assets, in Seattle as in many other cities, there is no central catalog of datasets and metadata. This research was conducted in partnership with the City of Seattle. The participation of departments in interviews and in the collection of key documents was critical to the success of this research in depicting, *in situ*, the governance of municipal open data.

A. QUALITATIVE ASSESSMENT I: KEY STAKEHOLDERS

Though our subject is municipal data, our backdrop is the people it affects. This section discusses our qualitative analysis of stakeholders' perceptions of open municipal data, particularly its downstream impacts. We understand that cities want to be responsive to their constituents, and

15. As Ronald Coase explains, illustrating with the case of the harmful effects suffered by many from the smoke exhaust of a factory:

> [D]irect governmental regulation will not necessarily give better results than leaving the problem to be solved by the market or the firm. But equally there is no reason why, on occasion, such governmental administrative regulation should not lead to an improvement in economic efficiency. This would seem particularly likely when, as is normally the case with the smoke nuisance, a large number of people are involved and in which therefore the costs of handling the problem through the market or the firm may be high.

Ronald H. Coase, *The Problem of Social Cost*, 3 J.L. & ECON. 1, 18 (1960). On the application of Coase's theory to privacy harm through transactions with personal information, see generally, Hoofnagle & Whittington, *supra* note 11, and Jan Whittington & Chris Jay Hoofnagle, *Unpacking Privacy's Price*, 90 N.C. L. REV. 1327, 1331 n.9 (2012).

we endeavored to gain a sense of the hopes and fears of residents and others around open municipal data. We designed the research question for this component to be open-ended and as inclusive as possible of the range of issues that stakeholders may find relevant to the initiative. Through focus groups and interviews, we asked users for their hopes, concerns, and expectations for Seattle's open data initiative.

1. Methods: Data Collection and Analysis

a) Research Design and Sampling

The data collection for this study included the following stakeholder groups: (1) Seattle residents in general, (2) civic hackers, (3) privacy activists, (4) city employees, (5) an academic, (6) a legal advocate, and (7) industry representatives.[16] Our hope was to talk to those who directly use or would potentially use open municipal data, as well as those who work on closely related issues. Thus, with the exception of the group of "residents in general," respondents were largely familiar with the topic at the time of the focus groups and interviews.[17]

Seattle's local tech economy offers unique access to major industrial players, tech hobbyists, and activists. Data collection for this study was conducted with these existing organizations. For example, the "civic hackers" focus group was conducted with a local hobbyists group which meets weekly to build apps of local interest using open data. The focus group with privacy activists was conducted with members of a community activist organization focused on privacy issues, like the use of police surveillance cameras. The four industry representatives interviewed came from relevant departments in three large local corporations.

Most sampling for the study was purposive, based on respondent membership in relevant organizations or interest in the study.[18] Civic hackers, privacy activists, the legal advocate, academic, and industry representatives were contacted directly for their relevance to the study.

16. We adopt the Value Sensitive Design definition of stakeholders: "Direct stakeholders refer to parties—individuals or organizations—who interact directly with the computer system or its output. Indirect stakeholders refer to all other parties who are affected by the use of the system. Often, indirect stakeholders are ignored in the design process." Batya Friedman, Peter H. Kahn, Jr. & Alan Borning, *Value Sensitive Design and Information Systems*, *in* EARLY ENGAGEMENT AND NEW TECHNOLOGIES: OPENING UP THE LABORATORY 55, 73 (2013).

17. We used a focus group format to collect data from the first four stakeholder types listed. Due to scheduling constraints, data from a legal advocate, academic, and industry representatives was based on interviews.

18. As part of the University of Washington Institutional Review Board (IRB) approval for this study, demographic information about respondents was not collected.

Members of the general public were recruited via fliers and Craigslist.[19] Our hope for the city employees focus group was to speak with workers on the "front-line"—police, fire, waste management, and others who drive fleet vehicles; constraints within the city made this infeasible. The city employees who participated were largely administrative staff; nevertheless, this group was more sensitive to potential privacy issues than we had expected.

b) Data Collection

Data collection for this study was based on focus groups and interviews. The focus group format was piloted twice to make it more neutral. Each focus group had 7–10 members and lasted 60–120 minutes. We used this format for residents, privacy advocates, civic hackers, and city employees.[20] Focus groups are well-suited for understanding unobservable phenomena like attitudes.[21] As a method, focus groups present a risk of respondent bias and group-think; our research design took measures to minimize these risks.[22]

Focus groups began with a 10-minute introduction from the moderator covering relevant background information. The moderator introduced the city's open data portal, the types of data currently available on it, and data types that the city has made available. The moderator introduced the Washington State Public Records Act (PRA), and its strong value on government transparency.[23] The PRA is a state law that establishes broad rights for state residents to request public records. It is intended to promote government transparency and accountability. The moderator explained that while the PRA requires the reactive release of data in light of a public disclosure request, open data is proactively released and not mandated. The presentation discussed how data is anonymized by removing its identifying

19. This group was compensated $15 for their time. No other respondent was compensated. Perhaps because of this means of recruitment, respondents for the general public group happened to be people experiencing instability in employment and housing.

20. In addition, we also interviewed four industry representatives, a legal advocate, and an academic.

21. For a detailed discussion of the strengths and weaknesses of focus groups as a research method see DAVID L. MORGAN, FOCUS GROUPS AS QUALITATIVE RESEARCH 13–17 (2d ed. 1996).

22. *See* Jenny Kitzinger, *Qualitative Research. Introducing Focus Groups*, 311 BRIT. MED. J., 299–30 (1995) ("The method is particularly useful for exploring people's knowledge and experiences and can be used to examine not only what people think but how they think and why they think that way."). *See generally* Jenny Kitzinger, *The Methodology of Focus Groups: The Importance of Interaction Between Research Participants*, 16 SOC. HEALTH & ILLNESS 103 (1994).

23. *See* WASH. REV. CODE § 42.56 (2011) (Public Records Act).

attributes, and under what circumstances data subjects may be re-identified, if any. Focus groups were conducted with a minimal moderation approach.[24]

c) Data Analysis

Transcripts of the focus groups and interviews were analyzed via qualitative coding. The first round of coding used a priori codes based on our research questions. The second round of coding used open and axial coding, in keeping with a grounded theory approach.[25] Analysis was conducted using NVivo 10 qualitative data analysis software.[26] Using this tool, the researcher tags blocks of text with a theme. Based on these tags, the software creates a database of quotes indexed by theme and respondent group. Iterative, inductive coding was formalized as a coding manual, by which data analysis was standardized across respondent groups. In keeping with a grounded theoretic approach, the following results are closely derived from the data.

The results of the stakeholder analysis offered a range of perceptions on the downstream impact of open data. Due to the exploratory, open-ended nature of this study, the analysis covered a broad scope of hopes, concerns and expectations about who will use the data, and to what end. Issues related to public trust, privacy, race, and social justice were of core interest to this work. Additional topics, like safety, commercial actors, and legal issues also emerged in the analysis. In this section, we discuss results by theme, and offer a sense of the inter-group variation on a given issue.

24. Respondents were told that the central goal of the session was to hear as many of their hopes and concerns as possible. Three themes—public trust, privacy, and race and social justice—were of particular interest to this project. Rather than prompting these themes directly, the moderator waited to see if they arose naturally from the conversation. If any of these topics were not addressed, the moderator made a note of this, then directly addressed remaining themes at the end of the session.

25. Qualitative coding is an interpretive process of systematically analyzing a text to surface themes within it. A priori codes are themes that the researcher brings to the text. A grounded theory approach necessitates that these themes arise from the text itself. Open coding is the initial process of capturing each theme from a text; axial coding combines these open codes into groups. For background on these coding methods, see generally Juliet Corbin & Anselm Strauss, *Strategies for Qualitative Data Analysis*, *in* BASICS OF QUALITATIVE RESEARCH: TECHNIQUES AND PROCEDURES FOR DEVELOPING GROUNDED THEORY 85 (4th ed. 2014).

26. *See What is NVio*, QSR INT'L, http://www.qsrinternational.com/what-is-nvivo (last visited Sept. 23, 2015).

2. Findings

a) Effects of Open Data Initiative on Public Trust

Respondents' primary hope for open data was that it would increase transparency in government. Every group touched on this sentiment, although the form it took varied. This included hopes for greater transparency, the democratization of governance, and the hope to build a better society through data-driven policy decisions. Government accountability was of keen interest to those in five of the seven stakeholder groups. This was expressed in many forms, from oversight on police or prison guard actions, to residents fact-checking politicians by looking at the same raw data. Some groups, like the civic hackers, presented this hope with conviction: "Having the data be open is an incredible source of accountability. It is a key to democracy."[27] This group spoke in-depth about opportunities for widespread data-literacy, which was viewed as a key intermediate step to true accountability. Others, especially privacy advocates, and residents in general, held similar hopes while also more ambivalent; we outline these concerns further on.

b) Economic Value Latent in Data

A commonly stated goal for open data is that it can bolster the local economy. Stakeholders—including industry representatives, privacy activists, and civic hackers—shared this goal. Some focused on ways open data can foster new companies and lead to more jobs, or allow existing companies to offer new products. Industry representatives were interested in ways that commercial actors improve the quality of data as they use it, and cited the potential for a "two-way pipe," by which companies could add value to the data—e.g., with real-time data feeds—and give it back to the city.[28] One industry representative said data could be used to target their marketing: "How do you find out which customers are heavy commuters? You just ask the city for all the tapes about license plates."[29] Privacy activists and hackers said that businesses could help interpret and make the data more usable to everyday people. However, one privacy activist thought that while analysis and usability was a valuable role for businesses, it constituted a public good that should not be delegated to private actors. Civic hackers were hopeful that open data could help smaller, more agile companies replace large firms in government procurement.

27. Focus Group, Civic Hacker Organization, in Seattle, Wash. (Feb. 12, 2015).
28. Telephone Interview, Industry Representative #2 (Mar. 27, 2015).
29. Telephone interview, industry representative #1 (Mar. 27, 2015).

c) City Management of Open Data Initiative

Stakeholders asserted a range of expectations for the city in how they proceed with the open data initiative. Every group stated that the data should be anonymized prior to release. In keeping with the spirit of the PRA, there was also a strong conviction that data held by government belonged to the public. The groups who most used this data, like industry representatives, privacy activists and civic hackers, had specific input for the way the data is and should be stored, accessed, formatted, licensed and released. These groups stated that the license terms under which the data was released should be clearer. The legal advocate and academic shared the expectation that the city should limit data collection, and limit its use beyond that for which it was collected. Despite potential risks, civic hackers and privacy advocates were profoundly opposed to the idea of access restrictions, fearing that they would be used against someone with legitimate interest in the data. Often, the scope of this conversation moved into one about the city as a data custodian: its data storage, retention, and deletion processes.

Multiple groups shared a sense of unease about the city's ability to prevent data spill.[30] This concern was echoed by members of the general public, who were acutely concerned about hacking and identity theft. Both industry and city employees said that the city's servers are regularly targeted by Chinese hackers and other international actors. As we discuss further on, both the general public and city employees were concerned that hacked data would be used to threaten critical infrastructure.

There was large variation within and between groups on the feasibility of use restrictions on the data, with an overall sense that restrictions would not be enforceable. Civic hackers and privacy activists noted the practical problems with governing uses of data once it is made open. The legal advocate pointed out that some forms of use restrictions would represent unconstitutional restraints of free speech. Even in the absence of formal use restrictions, industry representatives were sensitive to the way the public

30. One industry representative said:

> They need to follow reasonable baseline data security practices, particularly if the city is going to be a repository of big data. And, if for-profit companies in the health-care sector, for example, have under-invested in data security, then it's a fair bit to say the IT systems of many municipal governments aren't where they should be either.

Id.

would react to different uses.[31] Public-facing organizations, as opposed to organizations that work business-to-business, were thought to use public feedback as a check on data uses. The legal advocate shared this sense, adding that data brokers and less visible actors are less responsive to norms around data use: "Is anyone really comfortable with the variety of awful things that have happened with commercial actors in this space—like companies creating extortion schemes by posting photos of people online that they get via public records?"[32] While use restrictions were generally deemed infeasible, this quote illustrates the ambivalence stakeholders expressed about unintended consequences of data release.

d) Privacy Interests in Open Data

Privacy implications of the open data initiative were a prominent feature in every conversation, with the exception of the civic hackers group. Some respondents among the general public and civic hackers asserted that "privacy is an illusion."[33] Members of these groups strongly believed a data spill was liable to happen eventually. However, they were less concerned about privacy implications than they were that public outcry would slow the momentum of the open data initiative. Civic hackers framed concerns about privacy as important, but coeval with concerns about data inaccuracy and misinterpretation. Overall, this group shared an impetus to get "more eyes on more data"[34]—data in anonymized form. Some respondents in the privacy activist group shared the civic hackers' confidence that data

31. One representative said, "We're very conscious of ethics and big data, civil rights and big data, and trying to be really thoughtful about how we combine data so that it isn't used in bad ways or identifies people." Telephone Interview, Industry Representative #3 (Mar. 27, 2015). Similarly, another industry representative said:

> [I]t could be useful for commercial benefit if you're doing that in a de-identified or aggregated way, and that shouldn't be a problem. If you're doing it in a personally identifiable way—so the people can add factors to your behavioral profile—that's probably going to rub people the wrong way.

Telephone Interview, Industry Representative #1 (Mar. 27, 2015).

32. Interview, Legal Advocate, in Seattle, Wash. (Feb. 19, 2015).

33. Focus Group, Civic Hacker Organization, in Seattle, Wash. (Feb. 12, 2015). One civic hacker said:

> I think that banks and private health care are a much bigger concern for privacy problems than the government; they're a lot more focused. [Governments have] bits and pieces of data all over the place, you'd have to really want to aggregate that stuff in order to really drill down in somebody's privacy.

Id.

34. *Id.*

anonymization processes are resilient to reverse re-identification. Members of the general public and the legal advocate were less confident that data anonymization could protect individuals.

Other stakeholders had more acute privacy concerns. There was a general sense that the city had sensitive data. A privacy advocate said, "I fear the efforts to make data available about the government actually makes data available about the public."[35] The category of what information is or should be "private" varied between groups. Members of the general public framed private data as social security numbers and information related to financial status (e.g., credit rating). An industry representative and civic hackers emphasized that locational data would be a privacy concern, if released in a granular way. The legal advocate favored an approach that would scrutinize any data type as one piece of a larger mosaic: "If it's a sufficient analysis, it's also going to take into account whether this information, when correlated with other data that is available, presents harms."[36] The legal advocate spoke to ways that data could be re-identified; thus, he said that entire record types should be considered sensitive (e.g., police video) and exempt from proactive release or most forms of public records request.

City employees' discussion of what constitutes private information was broader than that of other groups, due in part to the large amount of information the city has in their personnel files. Employees described the different standards of privacy that applied to them as public employees. They recalled the shock of adjusting to having their salaries posted publicly. Members of the group were unaware of whether certain data types were protected from public records request under the PRA, for example, home address, employee benefits, and retirement information. These respondents were also very concerned about the release of insurance information such as the identity of their dependents or other family members.

Multiple respondents within all groups mentioned specific segments of the population they perceived as having special privacy interests. Several groups, including the general public and civic hackers, mentioned the special interests of children and the elderly. One privacy activist said:

> It's a really privileged position to be able to say that everything should be open. People with experiences of different kinds of abuse have had to build hiding into their cultural identity—open is not just going to work for them.[37]

35. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015).
36. Interview, Legal Advocate, in Seattle, Wash. (Feb. 19, 2015).
37. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015).

Safety concerns were the primary reason cited for these special privacy interests.

e) Safety Risks Latent in Data

Concerns about safety were more widely held than we had expected, and came up in conversations with every group. Respondents were concerned about the safety of vulnerable populations. There was concern that children, elderly people, and victims of previous crimes would be specifically targeted by criminals seeking to assault or con them. These concerns were brought up widely, in five out of seven groups. The nature of government services means that those in need will be especially present in the data. One City employee pointed out ways that police officers' route information reveals domestic violence: "you can find safe houses, individuals that are maybe victims that are being involved in their processes and response patterns."[38] Privacy activists noted that governments also have data on foster children and those in child protective services.

Multiple respondent groups were concerned with the safety implications for City employees. First responders were perceived to be at risk of vigilante justice. A privacy activist said:

> People have tried to find out where cops live so they can go to their houses and do stuff to them. Cops still have personal rights and personal privacy rights and stuff too, even though we would default to thinking that they don't go out of their way to respect our own.[39]

This concern for officers' safety co-existed with the respondent's other attitudes about police. City employees even referred to a past PRA request for police officers' home addresses that had been granted. They noted that this incident had led the fire department to take greater precautions with the kinds of identifying information it included in its reports.[40] City employees also raised the possibility that public data could be used to derive route patterns, which could be used by criminals to target officers on their daily routine.

---

38. Focus Group, City Employees, in Seattle, Wash. (Mar. 9, 2015).
39. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015).
40. One exchange in this group illustrates these concerns:

> You might get incident information, but you're not going to have the firefighters' names because then they're easily looked up. They're at Station X, OK—you can see shift details and stuff, so we have to be smart about it. Especially the kinds of shifts firefighters are on—they have to leave their families . . . . They're on 24 hours.

Focus Group, City Employees, in Seattle, Wash. (Mar. 9, 2015).

City employees were also concerned about their safety. Some responded that they felt they could be targeted because of their race or sexual orientation; one person described a city department's LGBT group meeting wherein another city employee tried to use the PRA to request the names of all attendees. The same person reported feeling outed when trying to change his or her official marital status.[41] Other respondents felt personally exposed by ways that public records are indexed and searchable on Google.

Safety risks were perceived to implicate not only individuals, but larger domestic security concerns. Members of the general public, industry representatives, and city employees referred to the potential for open data to be used to target critical infrastructure. This risk was framed as applying to physical infrastructures, like the power grid, as well as servers and other digital assets. To the extent that open data could be used to derive first response patterns, city employees were concerned that this information would be used to divert public safety officers from a planned attack. The academic cited a counterexample of the public safety utility of open data, especially public health concerns like vaccine and disease status.

f) Lack of Public Trust in the Management of the Open Data Initiative

Despite these risks, multiple stakeholder groups were concerned that the government would not open enough data. Civic hackers, industry representatives, privacy activists, and members of the general public shared a concern that open data efforts would fall short of its promise if very little data were released. Members of the general public and privacy activist groups shared a sense that those in city government would selectively record or release data to protect their own image. One privacy activist said, "If the city . . . maintains the ability to selectively refrain from publishing portions of that data, then we're not a whole lot better off than if they just weren't publishing in the first place."[42] Respondents in the civic hackers group and

41. This individual responded:

> It doesn't feel safe to me at all. My being, you know as a, being married, I had to contact a lot of people to get my status change in the city. They didn't, you know, so then I'm thinking okay, let's advertise it even more to everybody. I was certainly in my right so I'm going to do it, but it's pretty public. If I wanted to not tell people I was gay, it would have been impossible because everybody has access to it.

Id.

42. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015). A member of the Seattle residents group said, "This is just something they're doing to appease the general public because there's an outcry in America. But the police is going to be the police . . . as soon as they get some information they don't want to be publicized,

privacy activists were concerned that government actors could edit data, and raised the importance of using metadata or a data signature or hash that would verify its authenticity. While the responses of the general public and privacy activists exhibited low trust in government, civic hackers were more interested with issues of data quality.

Other groups worried that the promise of open data might become encumbered. One industry representative thought agencies might lose sight of the larger goals surrounding open data. He feared open data would become "a compliance exercise where the agencies and the cities will all do whatever they have to [do in order to] stop being bothered about it anymore."[43] This respondent spoke from a sense that unambitious management of the data would pose a missed opportunity. Both civic hackers and city employees noted that governments feared exposing themselves to liability from data release; for the civic hackers, liability and related concerns were framed as barriers to progress.

g) Perceived Social Justice Implications of Open Data

Respondents perceived open data as having promise for social justice issues. Half of the groups explicitly mentioned "social justice" issues without prompting. Even when not referred to explicitly, the implication of open data on social justice issues was present in respondents' ideas about government accountability for misconduct. Other references to social justice included the possibility of communities using data to advocate for themselves (civic hackers), data-driven policy (general public and civic hackers), and crowdsourced service requests (e.g., potholes, streetlight reports) (industry representatives). While some in the general public group felt that open data would have positive and incremental social justice implications, one person thought that little would happen in this vein: "I think the reality of it is, it's not going to really affect anybody that's down and out anyway in Washington State, it's only going to affect the . . . powers-that-be anyway."[44] Racial minorities within the general public group expressed a sense that open data would not be put to work on their behalf.

Other groups raised concerns that open data could have negative racial and social justice implications. Many of these were related to the potential that commercial uses of the data would have a disproportionate impact on marginalized communities. One member of the privacy activists said, "I fear

there's going to be a glitch in it." Focus Group, General Public, in Seattle, Wash. (Mar. 19, 2015).

43. Telephone Interview, Industry Representative #4 (Apr. 6, 2015).

44. Focus Group, General Public, in Seattle, Wash. (Mar. 19, 2015).

that it would be used to lower property values, redline insurance, et cetera, in neighborhoods with high crime rates rather than addressing those issues. I'm worried that data about precincts where people don't vote much could lead politicians to write them off."[45] A member of the general public group spoke to the ways that data, once open, is copied and persists:

> The information they put on there is a detriment to me because I've been trying to get, well I just got out. I was released from a penitentiary and I've been trying to get work and anytime they do a background check it's bringing up shit from like 1996. This is 2015.[46]

Taken together, these responses highlight how uses for open data could reify existing social marginalization.

3. Implications of Stakeholder Assessment

The open-ended nature of the qualitative stakeholder assessment resulted in some findings that we might have expected, some opinions that were more widely shared than we would have expected, and some surprises. For the purpose of our recommendations, we foreground the following results: (1) Multiple groups expressed concern regarding privacy risks latent in the data, especially to vulnerable and marginalized populations and city employees. Not all stakeholders were confident that anonymization would be enough to protect those listed in the data, although each stakeholder listed strong anonymization as an expectation for the city. (2) Stakeholder groups spoke to positive economic impacts from commercial uses of the data, but drew a clear line between these uses and those that were considered overly intrusive. Members of the general public were aware of threats to privacy from data brokers, which the research team did not expect. (3) City employees did not know what aspects of their personal data were protected, and they did not feel safe. (4) In thinking about open data, many groups spoke more broadly about issues of data custodianship; in their eyes the city's responsibility to protect its data and to open it intertwined. (5) Stakeholders were not clear about the terms under which data was released, and asked for data licensing, with more clear terms. (6) Respondents were concerned about ways that governments might prevent data release to protect itself, or might treat different data requestors differently. Our recommendations were shaped in part by the application of these findings.

45. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015).
46. Focus Group, General Public, in Seattle, Wash. (Mar. 19, 2015).

B. QUALITATIVE ASSESSMENT II: THE CITY

Having generated some context for our discussion by connecting with residents and other stakeholders, we turn to a discussion of how the City of Seattle actually processes and shares data. This section discusses the findings of interviews with city departments relevant to municipal data management and release.

1. The City of Seattle as a Case for Study

One underlying premise of this research is the tension or conflict between the adoption of "smart city" technology and the protection of privacy and fairness for the individuals and groups who generate the data. In this respect, recent events have made Seattle an ideal case for study. On February 3, 2015, the City of Seattle formulated and adopted a set of privacy principles, which will guide the actions the city takes when collecting and using personal information. Central to the principles is the following policy statement: "We work to find a fair balance between gathering information to provide needed services and protecting the public's privacy."[47] The six privacy principles adopted speak to the importance of keeping personal information private when collecting it, storing and using only what is needed for city services, and being accountable for "managing your personal information in a manner that is consistent with our commitments and as required by law."[48] Where possible, the City also commits to updating information to be accurate, and notifying citizens on how information is used.

Many Seattle departments have adopted or contracted for the use of various smart city technologies to improve the efficiency and effectiveness of public services. Smart cities have been defined according to their use of large-scale sensor networks to improve the provision of city services.[49] As Rob Kitchin explains,

> The notion of a 'smart city' refers to the increasing extent to which urban places are composed of 'everyware'; that is, pervasive and ubiquitous computing and digitally instrumented devices built into the very fabric of urban environments (e.g., fixed and wireless telecom networks, digitally controlled utility services and transport infrastructure, sensor and camera networks, building

47. CITY OF SEATTLE, PRIVACY PRINCIPLES, http://www.seattle.gov/Documents/Departments/InformationTechnology/City-of-Seattle-Privacy-Principles-FINAL.pdf. Disclosure: One of us assisted Seattle in its formulation of privacy principles through his participation in an advisory board.

48. *Id.*

49. Kitchin, *supra* note 8, at 1–2.

> management systems, and so on) that are used to monitor, manage and regulate city flows and processes, often in real-time, and mobile computing (e.g., smart phones) used by many urban citizens to engage with and navigate the city which themselves produce data about their users (such as location and activity).[50]

The adoption of these technologies amongst Seattle's departments, and the simultaneous adoption and development of citywide privacy principles, signify the tension that exists between the perceived role of the city as a custodian, consumer, and distributor of data about residents. Depending on the perspective one has, or rationale one adopts, the same categories of data may be considered either to be of value to the public—therefore warranting public distribution, or of value to the public—meaning it should be kept in a secure state with strict controls on access.

2. Selected Departments: A Sample Size of Eight

Like virtually all mid- to large-sized municipalities, the City of Seattle functions more as a federated system of departments than a hierarchy.[51] The open data portal in Seattle is the product of activities conducted by the Department of Information Technology, which oversees the third-party contractor who maintains the portal. However, each department in the City governs the data it generates with considerable autonomy.

With regards to the release of data, departments are also subject to many different rules and regulations, from both internal and external sources. The Washington PRA, however, applies to all departments.[52] Thus, many of the City's units are involved in the release of data.

The City of Seattle contains thirty-six departments and agencies.[53] Within this population, we selected eight to research: the Department of Information Technology; the Department of Planning and Development; Finance and Administrative Services; Seattle City Light; the Department of Transportation; the Police Department; Parks and Recreation; and the Fire Department. A few criteria, generally organized around the principles of maximizing internal variation and generalizability, guided our selection. In consultation with City staff, departments were selected to represent the variety of challenges and approaches cities face as data is pushed, pulled, and spilled. Most, but not all of the selected departments, are active users of the

50. *Id.* (internal citations omitted).

51. In comparison to private firms, municipalities appear to be very flat organizations. This is due in part to the sheer number of roles and responsibilities mandated for and by local government.

52. *See* WASH. REV. CODE § 42.56.010(1) (2014).

53. *See Departments and Agencies*, SEATTLE.GOV, http://www.seattle.gov/city-departments/departments-and-agencies (last visited June 23, 2015).

open data portal. Many, but not all, are undergoing rapid changes in data management due to the adoption of new information technology. Almost all govern at least some data that is understood to be either sensitive or confidential, though the characteristics of the data subjects and the attributes of those datasets differ considerably. This list includes the departments that receive the greatest demand for public disclosure requests, but also some that experience very few. They rely on a wide variety of third party contractors for information-intensive services.

Importantly, however, departments were selected to represent the variety of technologies and enriched information flows that are the hallmark of smart cities. For this purpose, we based selection on a rationale categorizing sensors and data subjects as "stationary" or "mobile." Both a sensor and data subject can be stationary, as is the case with advanced meters with sensors that automatically record electrical or water use in the home or office. The sensitivity of this data is generally a function of its granularity over time. A sensor can be stationary while the subject of the data is mobile. This is the case in the study and provision of transportation services, which track the movements of data subjects. Both the sensor and data subject can be mobile. Video cameras hoisted on police patrol cars or pinned on the lapels of police officers' uniforms are examples. This schema is useful for beginning to think about ways that information technology advances can result in the production of more sensitive data.

With the eight departments selected, in-person and telephone interviews were conducted with departmental personnel in various roles associated with the push, pull, and spill of municipal data.

a) The Department of Information Technology

Shortly after President Obama signed the 2009 Memorandum on Transparency and Open Government,[54] the start-up firm Socrata approached the Department of Information Technology about purchasing its services to support open data. After about a year of conversation, Seattle contracted with Socrata and began the process of selecting and examining datasets for release to an open data portal.[55]

In considering the publication of data, the Department of Information Technology uses a classification system with four levels:

> **Public Information**: Public information can be or currently is released to the public. It does not need protection from

54. Transparency and Open Government, 74 Fed. Reg. 4685 (Jan. 26, 2009), https://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment.

55. Interview, Department of Information Technology personnel, Seattle, Wash. (Jan. 21, 2015).

> unauthorized disclosure, but does need integrity and availability protection controls. This would include general public information, published reference documents (within copyright restrictions), open source materials, approved promotional information and press releases.
>
> **Sensitive Information**: Sensitive information may not be specifically protected from disclosure by law and is for official use only. Sensitive information is generally not released to the public unless specifically requested. Although most all of this information is subject to disclosure laws because of the City's status as a public entity, it still requires careful management and protection to ensure the integrity and obligations of the City's business operations and compliance requirements. It also includes data associated with internal email systems and City User account activity information.
>
> **Confidential Information**: Confidential information is information that is specifically protected in all or in part from disclosure under the State of Washington Public Disclosure Laws. This could include certain personally identifiable information or vendor trade secrets.
>
> **Confidential Information Requiring Special Handling**: Confidential information is specifically protected from disclosure by law and subject to strict handling requirements dictated by statutes, regulations, or legal agreements. Serious consequences could arise from unauthorized disclosure, such as threats to critical infrastructure, increased systems vulnerability and health and safety, or legal sanctions. Departments handling this category of information must demonstrate compliance with applicable statutes, regulatory requirements and legal agreements. Information in this category could include patient health records or student school records.[56]

Note that the first level pertains to data the City considers applicable for posting as open data (push). The second pertains to data that is subject to disclosure by request (pull). The last two levels pertain to confidential data for which City staff have "a legal reason to refuse public disclosure."[57]

On the incentives for releasing data, department personnel suggest that they try to save costs on public disclosure requests. The message that pushing data to an online portal may result in more efficient public

56. E-mail Communication, Department of Information Technology Personnel (Jun. 29, 2015).

57. Interview, Department of Information Technology Personnel, Seattle, Wash. (Jan. 21, 2015).

disclosure is reinforced by the PRA, which notes, "The internet provides for instant access to public records at a significantly reduced cost to the agency and the public. Agencies are encouraged to make commonly requested records available on agency web sites."[58] Another rationale for municipal open data is the prospect of promoting economic or business growth in the city after the Great Recession. Importantly, department personnel also express hope that public open data has been anonymized properly. As they say, "how do you make a race car go faster? You give it better brakes."[59]

b) The Department of Planning and Development

One of the early and active participants in the open data portal was the Department of Planning and Development.[60] Most city datasets that concern infrastructure do not pertain to critical infrastructure. Among the datasets made public by the Department are Geographic Information System (GIS) files that show plans, land use, zoning, critical areas, topography, vicinity to park property, landmarks, planning and permits. All permits for work done on private property are posted to the open data portal. Department personnel describe the postings as "complete," and they can potentially include location, the property owner's identity, and the work performed.

The Department of Planning and Development, like all departments contributing open data, is thought to be the "owner" of the data, and it is up to their discretion whether to participate. The rationale behind Planning and Development's decision to participate is common to many departments that publicize data. Departments consider "the business case": is this data subject to repeated public disclosure requests? Would the preemptive preparation and release of the data through the open data portal save time and resources when compared to responding to public disclosure requests?[61]

c) Finance and Administrative Services

In the first analysis of sensitive data for release to the open data portal, the Department of Information Technology worked with Finance and Administrative Services to assess the risk of making business license data publicly accessible. As explained in their risk analysis:

58. WASH. REV. CODE § 42.56.520, finding 2010 c 69 (2010).

59. Interview, Department of Information Technology Personnel, Seattle, Wash. (Jan. 21, 2015).

60. Interview, Department of Planning and Development Personnel, Seattle, Wash. (Jan. 14, 2015).

61. *Id.*

> The Department of Finance and Administrative Services has developed a process for evaluating datasets against eight principles of open data and a risk analysis profile associated with publishing the data. The risk analysis defines who the final decision maker should be, and who will decide whether or not to publish the dataset.[62]

The principles the Departments referred to are the "8 Principles of Open Government Data," formulated during a 2007 meeting convened by Tim O'Reilly, of O'Reilly Media, and Carl Malamud, of Public.Resource.Org, with sponsorship from the Sunlight Foundation, Google, and Yahoo.[63] The principles formulated by this group assert that open government data should be:

> 1. **Complete:** All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.
>
> 2. **Primary:** Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
>
> 3. **Timely:** Data is made available as quickly as necessary to preserve the value of the data.
>
> 4. **Accessible:** Data is available to the widest range of users for the widest range of purposes.
>
> 5. **Machine Processable:** Data is reasonably structured to allow automated processing.
>
> 6. **Non-discriminatory:** Data is available to anyone, with no requirement of registration.
>
> 7. **Non-proprietary:** Data is available in a format over which no entity has exclusive control.
>
> 8. **License-free:** Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.[64]

62. City of Seattle Department of Finance and Administrative Services, Open Data Candidate Requirements and Risk Evaluation—Business License Data 3 (May 6, 2010), http://dropbox.ashlock.us/opengov/seattle/Open%20Data%20Candidate%20Requirements%20and%20Risk%20Evaluation%20V1%209.docx [hereinafter City of Seattle, Open Data Candidate Requirements].

63. *Open Government Data Principles*, PUBLIC.RESOURCE.ORG (Dec. 8, 2007), https://public.resource.org/8_principles.html.

64. *Id.*

The Departments also added that customer service personnel responsible for constituent requests should be notified.[65] Seattle's risk analysis compared each data type in the business license dataset to each of these eight principles. Analysis proceeded field by field, noting which were to be excluded from release because they contained data for internal use only, of a personal nature, or data generated by the system (i.e., data that is only of use to those who operate the business registration system). For example, analysis of the data under the first of the eight principles revealed several fields that contained sensitive data, which should be excluded from release.[66]

The final recommendations focused on the potential legal risk if a data type were released. The Departments recommended publishing part of the dataset,[67] that is, publishing the dataset without mailing addresses and [personal] regulatory information. The analysis recommended that a subset of the data be extracted each month, and prepared for output to the open data portal.

The decision to publish the data was influenced by the perceived risks inherent in publication. Low-risk data could be published as is, while high-risk data required "too much data clean up" prior to publication.[68] Medium-risk datasets required exclusion of only certain fields. The business license dataset risk analysis concluded with the statement: "The risk for this dataset is rated at Medium, therefore the final approver for publishing this dataset to data.seattle.gov will be the [Finance and Administrative Services] director."[69]

While this example illustrates the reasoning and approach Seattle has taken toward releasing datasets on Socrata's platform, Financial and Administrative Services Department personnel note that the effort required to release secure data has escalated significantly.[70] The department is currently working in coordination with several other cities in the Puget Sound region on an initiative to convert all business and occupation (B&O) tax data to an online portal for processing payments and providing results to queries for tax information. While not open data in the same sense as the data pushed to the Socrata platform, this initiative also proposes to reduce

65. City of Seattle, Open Data Candidate Requirements, *supra* note 62, at 17.
66. *Id.* at 9.
67. *Id.* at 19.
68. *See, e.g.*, *id.* at 9.
69. *Id.* at 18.
70. Interview, Department of Planning and Development personnel, Seattle, Wash. (Jan. 14, 2015).

costs to taxpayers by allowing secure, online payment and retrieval of tax information.

d) Seattle City Light

Seattle City Light is Seattle's publically owned electric power utility company. Currently, most of Seattle's residences are still outfitted with mechanical or relatively simple digital meters for reading and recording the rate of electricity consumption.[71] Seattle City Light employees take readings at the customer's residence or business location. This method delivers no more than six points of data per year, in sync with the utility's bi-monthly billing cycle.[72] However, technology in this sector has advanced rapidly, and Seattle's meter system is changing.

Seattle City Light has implemented three programs on a path toward smart metering. In 2008, the utility tried a pilot program with 457 meters that relied on cellular technology to provide daily, one-way, communication (from the customer's site to the utility).[73] Another estimated 6,000 meters, in places the utility describes as "hard to reach," are using radio frequency technology to signal usage to the utility.[74] For several years, the utility has also operated a program for customers who manage mid- to large-sized properties, providing continuous two-way communication through meters hooked up to phone lines. Referred to as Seattle Meter Watch, the program is part of a larger industry-led initiative, known as the Green Button Initiative. Since 2012, the Green Button initiative has been a White House-led effort to allow consumers to access detailed data about their electricity usage, and take advantage of online tools for saving money by managing their use. Seattle was the first utility in the nation to be certified under this initiative.[75]

As part of the utility's six-year Strategic Plan, Seattle City Light has begun to scale up the installation of advanced meters. Unlike the city's mechanical meters, which are simply read to produce one aggregated measure of electrical use per household or business address every two months, the meters available on the market today allow the option of using sensors to disaggregate overall electricity consumption in order to discern

71. Interview, Seattle City Light Personnel, Seattle, Wash. (Mar. 29, 2015).
72. *Id.*
73. *Id.* The Pilot Project Summary and Conclusions are on file with authors.
74. *Id.*
75. *Seattle City Light First Utility Certified for Green Button Data*, SEATTLE.GOV, http://powerlines.seattle.gov/2014/06/20/seattle-city-light-first-utility-certified-for-green-button-data.

the use of identifiable electronic appliances.[76] This type of sensor gives users and utilities the option of viewing the consequences of appliance use in terms of electrical demand in real-time.

Beyond allowing users to respond to and manage demand, Seattle City Light personnel also describe the potential benefits of this new technology in terms of the ability to more precisely discern where electricity is flowing, to re-route electricity based on this information, to improve the management of voltage issues and problems in the system, and to ensure a smooth flow of electricity.[77] This will also allow the utility to identify more precisely where in the system people may be tapping electricity illegally. Of course, as all of this data becomes more detailed, reporting electrical consumption over time or by appliance, it carries a greater potential to compromise the privacy and security of the home and workplace.

e) Department of Transportation

Transportation assets are expensive to build, operate, and maintain, and until recently, transportation departments have also had to spend inordinate amounts of money, time, and labor to simply collect data to estimate how much we use the various components of our transportation networks. The integration of GPS technology in smart devices on our person or in our cars has fundamentally transformed this problem for the Department of Transportation from one of costly and time-consuming data collection, to one of concern about the privacy implications of collecting and using personalized data. For example, the City has contracted the services of Parkeon to operate pay stations that accept credit card payments for parking,[78] and has recently added the services of Pay by Phone, a mobile payment vendor. In these cases the vendors develop databases that contain vehicle information and the identities of parking permit purchasers. The vendor attempts to anonymize the data by removing a subset of fields, and feeds the resulting dataset back to the department.

In regard to travel behavior, we found two opposing approaches to data collection underway in the Department. One unit within the Department contracts with the fitness software company Strava to provide data describing the movements of individuals who have opted in to the use of their running and cycling app.[79] Another unit in the Department has been using, through the vendor Acyclica, Bluetooth and Wi-Fi readers installed

76. For an explanation of this process, see UWTV, *UW Four Peaks -- Shwetak Patel*, YOUTUBE, https://www.youtube.com/watch?v=nnzzTFs0O2g.

77. Telephone Interview, Seattle City Light Personnel (Apr. 7, 2015).

78. *Id.*

79. *Id.*

in public places that automatically read and record the Media Access Control (MAC) addresses of multiple devices—i.e., smartphones, laptops, and automobile computers—to track the movement of individuals across the city. A MAC address is a serial number assigned to a computing device, typically during the manufacturing process, to make that device uniquely identifiable from all other network devices in the world. When turned on, personal computing devices constantly send their MAC addresses in signals that perform an electronic handshake with Bluetooth and Wi-Fi routers. In this case, Acyclica has been granted permission from the City to install readers that "sniff" and send the unique MAC identifier of personal devices to the servers of the firm. The firm, in turn, sends the data it collects on personal travel behavior to the Department.[80] Though people have no obvious way of knowing that their movements are tracked by Acyclica's devices, the firm operates a web-based portal that allows anyone with a MAC address to retrieve the travel behavior data specific to that device.[81]

f) Police Department

With respect to open data, the Seattle Police Department is a self-described "manufacturer of data for the public."[82] In terms of the demand for data, people have always expressed an interest in police activities, listening to police scanners, and requesting incident reports and data from 911 calls. The Department has adopted multiple technologies with implications for the generation of big data: they have a cloud-based service that captures citizens' online reporting, they deploy smart phones, they have computers onboard vehicles, they generate in-car video and body camera video, and are proposing to develop a data analytics platform with multiple applications. The Police Department typically receives three times more public record disclosure requests than any other department in the city.[83] In the first quarter of 2015, the number of requests rose by 400%, to an estimated 2,500.[84]

Personnel in the Seattle Police Department note that the rising increase in demand for public disclosure has coincided with the digitization of files and the advent of video recording devices mounted on the dashboards of

80. Interview, Seattle Department of Transportation Personnel, Seattle, Wash. (Mar. 10, 2015).

81. *See Analyzer User Guide*, ACYCLICA, https://acyclica.com/support/documentation (last visited March 12, 2015). The web portal is available at https://acyclica.com/products/acyclica-analyzer.

82. Interview, Seattle Police Department Personnel, Seattle, Wash. (Jan. 14, 2015).

83. Interview, Seattle Police Department Personnel, Seattle, Wash. (Mar. 5, 2015).

84. *Id.*

patrol cars and worn on the bodies of officers.[85] Gradual shifts over time have allowed public disclosure requests to become anonymous and free of charge. Individuals in the department explained that people making public disclosure requests used to have to provide a phone number to call, so that people would be notified when the documents were ready, or could be called to clarify the request. As one interviewee explained:

> [T]he department has moved from paper to electronic, so the people think it should be accessible data, they think that a report should be available right away, even though there are protocols. The types of records [now include] body cams, in-car video, 911 calls, audio statements in the field, photos, officers receive video, text messaging, emails, web browsing. People expect to be able to access this information as much as they want in real time.[86]

Departmental personnel explained how demands rise "on the back end" with the number of public disclosure requests.[87] The department receives about 125 requests per week. The department employs seven people full-time to respond to public disclosure requests, plus additional attorneys, paralegals, and people dedicated to 911 and video requests.[88] Each request to the department generates a series of actions and corresponding logs. Detectives assigned to the relevant case participate in the process, helping review requested information for civilian safety, privacy, officer safety and for compliance with numerous other policies and regulations that pertain to police records. The personnel involved are "very careful and conscious of the fact that we are dealing with victims and the most vulnerable and not on their best day." As they explain, "we want victims to continue to cooperate with the department, all weighting this with trying to be as open as we can." People are given the data they have the authority to receive (e.g., victims receive different data than the media). When data is not released, officers are required to explain the reasons in an exemption log.

The Police Department is struggling with the demands created by the sheer volume of both footage and requests. Specifically, the Department must wrestle with privacy concerns stemming from the fact that body camera video contains recordings of persons other than the police officer. In its most recent move, as part of the recently initiated program using body worn video cameras, the department has launched its own YouTube

85. *Id.*

86. *Id.*

87. Work "on the back end" consists of the tasks that Department personnel must carry out in order to satisfy a public records request.

88. *Id.*

channel.[89] Besides posting raw video clips that have been processed for public disclosure, the department is blurring video content and deleting audio (to "redact" the identity of persons in the video) that has not been through the process, and posting these feeds to YouTube to facilitate public disclosure, with dates, times and incident numbers so that interested parties can see what is available and make more specific requests.

g) Parks and Recreation

Seattle Parks and Recreation maintains twenty-six community centers and organizes hundreds of volunteers to provide community services and events. Those events are attended by thousands of children and adults registered in their databases each year.[90] The Department takes a conservative approach to public disclosure requests. Personnel have been successful in redacting the information describing the people who volunteer to run and attend their programs and, under the law, the City has the discretion to redact considerable amounts of information pertaining to juveniles.

Perhaps as a result of working predominantly with youth and at-risk populations, such as special needs children, Department personnel expressed the need to be careful when releasing information for public consumption.[91] The Department has sensitive information about employees, volunteers, and adults and youth registered for programs. They are aware that the use of personal information, when distributed through either open data portals or in response to public disclosure requests, can give people the information they would need to be able to harass someone, stalk someone, seek revenge, and commit various crimes. Personnel described fights between individuals, a person stalking a volunteer, and community groups pitted against one another over a controversial park project, as examples of circumstances that have precipitated public disclosure requests for personal information. Personnel described their success disclosing incident reports to requestors, while redacting the information that could be used to contact the other party.

h) Fire Department

The Seattle Fire Department manages large amounts of data, but has not yet gravitated to new information technologies to the degree that Seattle City Light, the Department of Transportation, and the Police Department

89. *SPD BodyWornVideo*, YOUTUBE, https://www.youtube.com/channel/UCcdSPRNt1HmzkTL9aSDfKuA (last visited July 22, 2015).

90. Interview, Seattle Parks and Recreation Personnel, Seattle, Wash. (Mar. 5, 2015).

91. *Id.*

have. Unlike other departments, the Fire Department provides emergency medical services, and controls the release of medical data in accordance with The Health Insurance Portability and Accountability Act of 1996 (HIPAA) and related rules and regulations governing personal health information.

Approximately 80–90% of Fire Department responses to calls are medically related.[92] In these situations Fire Department personnel produce paper and carbon copy medical reports that they input into special HIPAA-compliant scanning devices. About 200 two-page medical reports have to be entered each day.[93] Before it is stored in Department databases, data is shared and reviewed by the Department, the station that responded to the call, and with University of Washington doctors working with the Fire Department. It reportedly takes about ninety days before these records enter the Department databases. Department personnel suspect that the movement to digitize this process is not likely to change the demand for public disclosure of these records because requestors have to provide proof of identification, such as a scanned copy of a driver's license, to receive a copy of a report.

The Fire Department also stores sensitive data that does not pertain to HIPAA. And, like other departments, it receives requests that appear "frivolous."[94] Interviewees explained that a person could make a targeted use of the law to inundate the Department with requests. Even though a request appears frivolous, "you are legally required to respond . . . but we can't possibly respond." The PRA requires a response within five days of every request. The fine for missing this window, can reach as much $100 per page, per day. "It's the only hard deadline and [someone] could try to get you to trip up and you have to hit respond to those. Some of them could be months' worth of work. [Someone could] then send a message to the council threatening to sue and say you are not in compliance with the PRA."[95]

The Fire Department, like other departments, is experiencing pressure to release data in the form of public disclosure requests. However, accustomed to maintaining medical and other sensitive data on paper and specialized electronic systems, this department realizes many such requests may not be justifiable.

3. Analysis

The one common approach departments have in regards to open data is the desire to reduce the financial cost of public disclosure. If pushing data

92. Telephone Interview, Seattle Fire Department Personnel (Mar. 19, 2015).
93. *Id.*
94. *Id.*
95. *Id.*

to the open data portal, a YouTube channel, or a more sophisticated portal such as the Green Button initiative, promises to reduce the cost of responding to public disclosure requests, then departments generally aim to do so.

Departments differ widely, however, in their pace and degree of adoption of smart technologies, and thus they differ in terms of the challenges they face in preserving privacy and social justice when data is pulled for public disclosure from city files. Departmental personnel appeared interested in serving the public interest and fostering transparency. Many also share concerns that the PRA can be, or perhaps already is being used for, self-interested, wasteful, or harmful purposes. The timing of the growth of such requests coincides with the transition from paper to digital records, from charging a nominal fee to copy records to providing them at no charge, and from named to anonymous requests. The piecemeal exemptions to public disclosure that have accrued in the PRA show that some departments have tried to solve the problem through the State Legislature. The PRA includes a lengthy list of data exempt from release, categorizing exemptions based on specifically named attributes (e.g., name, address, telephone number) in the data, subjects represented by the data, and public programs or other contexts that motivated the public collection and disclosure of the data.[96] Other departments have taken a slower

96. The Public Records Act lists types of data exempt from public disclosure and, in doing so, either names specific attributes or uses the broader term "personally identifying information" to specify the data that are to be exempt. For example, in a section pertaining to public utilities and transportation information, exemptions include:

> addresses, telephone numbers, electronic contact information, and customer-specific utility usage and billing information in increments less than a billing cycle of the customers of a public utility contained in the records or lists held by the public utility of which they are customers, except that this information may be released to the division of child support or the agency or firm providing child support enforcement for another state under Title IV-D of the federal social security act, for the establishment, enforcement, or modification of a support order.

WASH. REV. CODE § 42.56.330(2) (2014). Further on, in the same section, exemptions include:

> The personally identifying information of persons who acquire and use transponders or other technology to facilitate payment of tolls. This information may be disclosed in aggregate form as long as the data does not contain any personally identifying information. For these purposes aggregate data may include the census tract of the account holder as long as any individual personally identifying information is not released. Personally identifying information may be released to law enforcement agencies only for toll enforcement purposes. Personally identifying

approach to adopting technology, concerned about the very same implications. A few have been more deliberative in their service of public disclosure requests, taking a more proactive stance of exempting personal information from public disclosure requests.

Interviewees' conceptions of the market for municipal data varied. When favoring the commercial application of open data, interviewees' conceptions of the firm appeared to be aligned with small startups and newly created firms. The idea of pushing data to an open platform for commercial use is not universally embraced, however. Many interviewees questioned the idea that it is possible to favor the interests of some firms, such as small startups, over others, when data made open is open to all. Those concerned with the differential treatment of firms seemed to have a broader view of the market for municipal data, including large, well-apportioned organizations. Only the Police Department expressed awareness of the way data brokers use publicly disclosed data—an issue raised because of the uses of profiles in criminal investigations. Contractual relationships between the city and firms cloud these issues. Finance and Administrative Services, for example, raised the issue of the unintended spilling of data by the vendors under contract to the city to create online data portals. Seattle City Light will face the same issues in designing portals for advanced metering data.

C. TECHNICAL ASSESSMENT: OPEN DATA ANALYSIS

This section explains the technical analyses we conducted on the City of Seattle's current municipal open data. At issue is the question of how the city may evaluate, prior to release, the potential for a dataset to compromise privacy.

1. The Problem of Cumulative Risk of Re-Identification

From our initial interviews we learned that most datasets released by the City of Seattle on the open data portal had received some scrutiny with regard to potential privacy harms. However, the practices in place only modeled the risk of data releases for each dataset in isolation.

As various scholars have found, otherwise innocuous datasets can be joined together in ways that result in re-identification and breaches of privacy. This simple fact, evidenced by the accomplishments and practices of firms that have amassed detailed dossiers on millions of people, is reason

> information may be released to law enforcement agencies for other purposes only if the request is accompanied by a court order.

§ 42.56.330(7).

to question the ability of a municipality to release any one dataset about persons while preserving the anonymity of those persons.[97]

Public policy reflects the idea that the potential harm caused by releases of personal information is a function of what the combination of two or more pieces of information may reveal about an individual. This is expressed in various state laws by the way in which they approach Personally Identifiable Information (PII),[98] typically defined as the combination of two or more attributes for the purpose of protecting individuals' privacy, identity and personal safety.[99] The City's policies and regulatory framework for governing the release of data generally follow this line of reasoning. As illustrated by its release of business license data, the City of Seattle correctly and appropriately uses this criterion to manage the issue of potential privacy harm in their analysis of each dataset prior to publication. However, this is an analysis of a dataset in isolation.

The fact that multiple datasets can potentially be joined together using matching information in common fields threatens the validity of any risk assessment that has been limited to a single set of data. All that an actor would have to do to invalidate the claim that the release of any one dataset is risk-free is to join it across common fields with identical or similar data.

---

97. *See* Ohm, *supra* note 11; Narayanan & Shmatikov, *supra* note 11, at 24–26; Solon Barocas & Helen Nissenbaum, *Big Data's End Run around Anonymity and Consent*, *in* PRIVACY BIG DATA, AND THE PUBLIC GOOD, 44–75 (Julia Lane et al. eds., 2014).

98. *Security Breach Notification Chart*, PERKINS COIE, https://www.perkinscoie.com/en/news-insights/security-breach-notification-chart.html (last visited July 21, 2015) (providing a full list of state definitions of PII, current as of June 2015).

99. *See* NIST, GUIDE TO PROTECTING THE CONFIDENTIALITY OF PERSONALLY IDENTIFIABLE INFORMATION (PII), http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf. The NIST Guide defines PII to include:

> [A]ny information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information.

Id. See also Narayanan & Schmatikov, *supra* note 11, at 24. Narayanan and Schmatikov note:

> PII is surprisingly difficult to define. One legal context is provided by breach-notification laws. California Senate Bill 1386 is a representative example: its definition of personal information includes Social Security numbers, driver's license numbers, financial accounts, but not, for example, email addresses or telephone numbers. These laws were enacted in response to security breaches involving customer data that could enable identity theft.

Id.

The resulting merged dataset would not have to be a successful join of every record in order to be used to re-identify individuals, or to associate persons with attributes that threaten to compromise privacy or safety. In other words, cities looking to release public data responsibly face the need to develop their capacity to assess the privacy posture of collections of datasets more globally, encompassing the impact that additional releases may have in combination with existing corpuses of publicly, and perhaps privately available data.

2. A Proposed Method of Ex Ante Evaluation

Our research includes an analysis of the tabular data already released and publicly available at Seattle.gov. The research design presented here models the methods that could be used to assess the privacy of collections of datasets before they are released from municipalities.[100]

Someone wishing to identify potential privacy-violating joins must first take the step of identifying what joins are possible. Traditional database joins involve simply combining records from one table with another based on a known shared field. Our aim, however, is to discern the maximum possible extent of joins. So, in contrast to traditional approaches, the joins we are contemplating combine information, which may not be perfectly matched, or may be nominally classified as different. The purpose is to produce the greatest possible degree of connections across datasets that have been published separately. For example, fields with differing data types, or combinations of fields such as latitude and longitude can be joined across datasets with a field called "address" if sufficiently overlapping information is compared.

A second step is to then assess identified joins for their potential harms to privacy. To accomplish this, some care must be taken to correctly categorize and classify the types of information in the datasets. The analysis depends on an understanding of the harms made possible through the association of different attributes, as they are found in the published datasets and joined using the methods described above. Rules and regulations governing personally identifiable information offer limited guidance;[101]

100. Anyone in the City interested in evaluating an additional dataset prior to release would add that dataset to the corpus of existing public data and repeat the analysis. It is important to note, however, that our analysis was limited in time and resources. It represents a starting point for further research.

101. *See* Narayanan & Schmatikov, *supra* note 11, at 25 ("What is 'reasonable'? This is left open to interpretation by case law. We are not aware of any court decisions that define identifiability in the context of HIPAA.").

empirical cases of re-identification are more likely to inform this part of the exercise.

These two steps are encapsulated in Rob Kitchin's definitions of indexical and attribute data. Indexical data is important because it enables attributes to be linked, and often is the data that can be used to identify the subject of the attribute.[102] Unique identifiers such as passport numbers, account numbers, MAC addresses, order and shipping numbers, and manufacturing serial numbers are examples of indexical data, as well as names, addresses, and zip codes. What people and firms are joining together with the use of indexical data are attributes that describe the subjects of the data. As Kitchin notes, "Attribute data are data that represent aspects of a phenomenon, but are not indexical in nature. For example, with respect to a person the indexical data might be a fingerprint or DNA sequence, with associated attribute data being age, sex, height, weight, eye colour, blood group, and so on."[103] The vast bulk of data in storage are attribute data, and because the attributes that may be sensitive in terms of privacy or social justice are associated with various indexical fields, this association places sensitive data at risk.

The expansion of indexical fields gives rise to new and more expansive datasets, along with rising hazards to privacy and social justice. In addition to these factors, the adoption of advanced technologies further thickens the flow of information, with more opportunity to join or enrich existing datasets with potentially compromising information. Kitchin mentions how the ingenuity and economic drive of people and firms to find more and more ways to join data has resulted in the expansion of fields considered useful for indexing.[104] Thus the threat of re-identification with the release of data is a moving target. As more variables become useful for indexing, more publicly available datasets may be used to join datasets in previously unimagined ways.

One way to operationalize the first step—determining which joins are possible—is to turn collections of tabular datasets into network graphs that illustrate a variety of strategies for identifying potential joins between multiple datasets. This approach casts individual tables (i.e., each a dataset) as nodes in a network, connected by lines as identified by a specific join identification strategy (e.g., joining tables on the basis of specific indexical fields, such as location in space, as identified through latitude and longitude). If each separate table were joined on one indexical variable,

102. *See* ROB KITCHIN, THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES AND THEIR CONSEQUENCES 8 (2014).

103. *Id.*

104. *Id.*

showing tables as nodes and indexical field data on the lines connecting nodes to one another, one could see within the scope of a single diagram the possibility for joining multiple datasets. With a diagram showing the potential to join multiple datasets along one or more indexical fields, determining the possibility of connecting an attribute in one table to an attribute in another table could then become a network pathfinding operation. The network of datasets resulting from this approach would be amenable to the full-range of network analytical methods.[105] New datasets under consideration for release could be added to the network, and the changes in network topology studied with precision.

The second step—the assessment of the potential for harm from any one specific join—is likely to remain somewhat of a human intelligence task. This approach segments individual attributes into a continuum of privacy and social justice risk. Combining this continuum with a network dataset could allow the programmatic identification of instances where connections between low-risk attributes (e.g., describing the built environment) and high-risk attributes (e.g., describing persons in the built environment) result in potential information leaks.

3. Potential Join Strategies

We have envisaged several join identification strategies, all of which have different characteristics, advantages, and disadvantages with respect to quality of results, false positive or negative rates, processing time, and computing resources.

Some of these strategies work at the schema level (i.e., across field names or column headings, in the case of tabular data), and compare the names of individual fields (e.g., latitude, longitude, address). These strategies may be especially useful for inferring links between datasets that are held by a city and datasets that may not be wholly obtainable by a city (i.e., held by a third party). For example, one could infer a potential join where two tables share an "address" column. Other strategies extend the schema comparison approach by using natural language processing to identify conceptually related terms, inferring matches between fields such as "location" and "postal address."

Other strategies that are more exhaustive operate at the level of the data itself. These include the attempt to join, through exact matching, all fields in all datasets. This is computationally expensive, but answers concretely the question of where deterministic joins are possible. Other variants of this

105. An example of an analytical method that could be applied is Dijkstra's shortest path algorithm. *See* E. W. Dijkstra, *A Note on Two Problems in Connexion with Graphs*, 1 NUMERISCHE MATHEMATIK 269 (1959).

strategy include spatial joins, for example, that make geometric comparisons of the spatial attributes within tables.

Many more join identification strategies are likely to be employed by data brokers, or other would-be users of these datasets. Future work might identify additional strategies or integrate ensembles of strategies for identifying potential joins, such as using natural language processing techniques to perform meaning-based comparisons of all fields in all databases.

4. Analysis and Results

We implemented several join identification strategies, and used them to perform an initial analysis of the datasets that were publicly available from the City of Seattle's open data portal, as of April 1, 2015. At that time, there were 235 datasets on the Socrata open data portal from the City of Seattle. The strategies we employed include:

- Exact match of field name
- Tokenized match of field name components[106]
- Levenshtein distance match of field name[107]
- Natural language processing match of field name (i.e., Wordnet)[108]
- Exhaustive exact match of column contents

106. Technopedia offers the following definition of "Tokenization":

> Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. The tokens become the input for another process like parsing and text mining.

Tokenization, TECHOPEDIA, http://www.techopedia.com/definition/13698/tokenization (last visited July 23, 2015).

107. The Levenshtein Distance can be defined as "[t]he smallest number of insertions, deletions, and substitutions required to change one string or tree into another." *Levenshtein Distance*, NIST, https://xlinux.nist.gov/dads/HTML/Levenshtein.html (last updated June 22, 2015); *see also Levenshtein*, PHP MANUAL, http://php.net/manual/en/function.levenshtein.php (last visited July 23, 2015) ("The Levenshtein distance is defined as the minimal number of characters you have to replace, insert or delete to transform str1 into str2.").

108. *The Stanford Wordnet Project*, http://ai.stanford.edu/~rion/swn/ (last accessed July 23, 2015) ("By applying a learning algorithm to parsed text, we have developed methods that can automatically identify the concepts in the text and the relations between them."); *see also* Snow et al., Learning Syntactic Patterns for Automatic Hypernym Discovery (2004 Conference on Advances in Neural Information Processing Systems), http://ai.stanford.edu/~rion/papers/hypernym_nips05.pdf.

- Partial latitude and longitude geometric match of geospatial column contents

Relatedly, we have partial results of an ordering of the individual fields found within Seattle's open datasets. The number of datasets with tabular data that could be analyzed (i.e., contained field names and field contents) was 204. The City offices contributing to the corpus of open data included: City Budget Office; Department of Human Services; Department of Neighborhoods; Department of Planning and Development; Seattle Fire Department; Office of the City Clerk; Seattle Police Department; Office of the Mayor; Seattle City Attorney's Office; Department of Information Technology; Department of Transportation; Finance and Administrative Services; Seattle Public Utilities; and the Seattle City Council.

The datasets contained a wide variety of information, such as building permits, electrical permits, land use permits, code violations, surveys of residents' use of information technology, traffic counts, announcements of learning programs and events, commute trip reduction surveys, police department incident reports, active business licenses, 911 call logs, housing emergency responses, logs of police in-car video, grants and funding, adopted budgets, and neighborhood matching grant reports. Many were inventories of infrastructure assets, such as assets listed for auction, cultural spaces, road weather information systems, trails, street parking signs, and neighborhood maps. Of note are several datasets on the Socrata portal that are produced as part of a performance dashboard for municipal services.[109] Performance dashboard datasets include, for example, pothole complaints and repairs, streetlights data, conservation data, planted trees, first arriving engines in emergency response, police reported collisions, bus ridership, city building energy use data, pea-patch garden registrants, residential burglaries, motor vehicle theft, and civil rights performance data.

a) Joins Using Exact and Flexible Matching Strategies

As one would expect, exact matching strategies (i.e., exact matches of field names, or column headings) for these datasets appear to result in many false-negatives, whereas more flexible matching strategies appear to result in many more false-positives. For the purpose of demonstrating potential flaws in vetting datasets for publication, flexible strategies are important to use so as to not overlook valid matches; eliminating false positives manually was the price for complete coverage.

Results from our schema-based join identification strategies suggest a great deal of connectivity between datasets on Seattle's open data portal.

109. To explore these datasets, and others, see *Performance Seattle*, SEATTLE.GOV, https://performance.seattle.gov.

The total number of field names in the corpus of 204 datasets was 3,859, and the number of unique field names (a product of exact match of field name) was 1,981. Tokenized, the field names in the corpus of datasets produced 6,061 parsed names. Among these were many duplicates. Eliminating duplicates left 1,828 parsed field names. The Wordnet comparison of parsed field names returned thirty-one pairs with 100% match, and another 230 pairs with a 50% match.[110] For example, forty-six fields are named "address." Given the ubiquity of certain terms such as address, as well as other common fields, the number of connectable tables results in a network graph that expresses the possibility of joining nearly all tables in the set—forming one comprehensive table out of 204. This validates the premise that it is possible to recombine data in ways that violate the current model for vetting publication of datasets (i.e., assessing datasets in isolation).

Results from our content-based join identification strategies were also promising. We performed a many-to-many comparison (i.e., an exhaustive comparison of data entries in all cells), using exact matches only, across all fields of all datasets. This resulted in a large number of irrelevant matches for common objects (e.g., numbers, "true/false," "yes/no"), and very few exact matches for data in cells. This result was expected, since the published datasets do not constrain or normalize data in fields. For example, reliance on exact matches produces results that suggest "302 N Baker Street" is not an exact match to "302 N Baker St." This supports the notion that using broader, more flexible strategies for finding matches and weeding out false positives is a useful approach.

After the exhaustive join on exact matches of field contents, the next likely research step was to either use more flexible joining strategies with the entire corpus of data, or more targeted joins on the basis of potential privacy harm. We opted to implement the latter, through one smaller but significant strategy for joins, with the purpose of illustrating some of the unusual qualities of local government data.

b) The Special Relationship Between Municipalities and Spatial Data

The more we studied the open datasets, the more it appeared to us that spatial data is highly represented among Seattle's municipal open datasets. We mentioned the commonality of "address" but it is worth noting that nearly all of the datasets included spatial data of one kind or another (i.e.,

110. Results show how closely Wordnet's system believes they are related to one another. The parsed field names included in this analysis were all nouns. All other parts of speech were excluded.

latitude, longitude, block, location, mailing, shape, zip code, acres, area, and shape files).

There is a logical rationale for this observation. If, as employees of departments had suggested in interviews, efforts to de-identify datasets prior to publication primarily involved the removal of names, telephone numbers, and email addresses, while retaining street address (sometimes aggregated to the nearest 100 block), zip code, or another similar spatial identifier, then spatial data would be more likely to be retained in the datasets made public. Also, considering that cities are primarily interested in data regarding activities within the spatial boundaries of their jurisdiction, and meaningful determinations of demand, supply, and quality of services often pertain to the delivery of services across the spatial extent of the jurisdiction, spatial data is likely to be a key variable in municipal data.

However, spatial data can also be the means to identify individual home and business owners and occupants. Residents are readily identified when their name is associated in any publicly available dataset with these properties. For example, the City of Seattle includes the names of persons on building permit applications in open datasets, and King County (which includes the spatial extent of Seattle) maintains a publicly available dataset that includes the names of the owners, addresses, and assessed value of the properties.

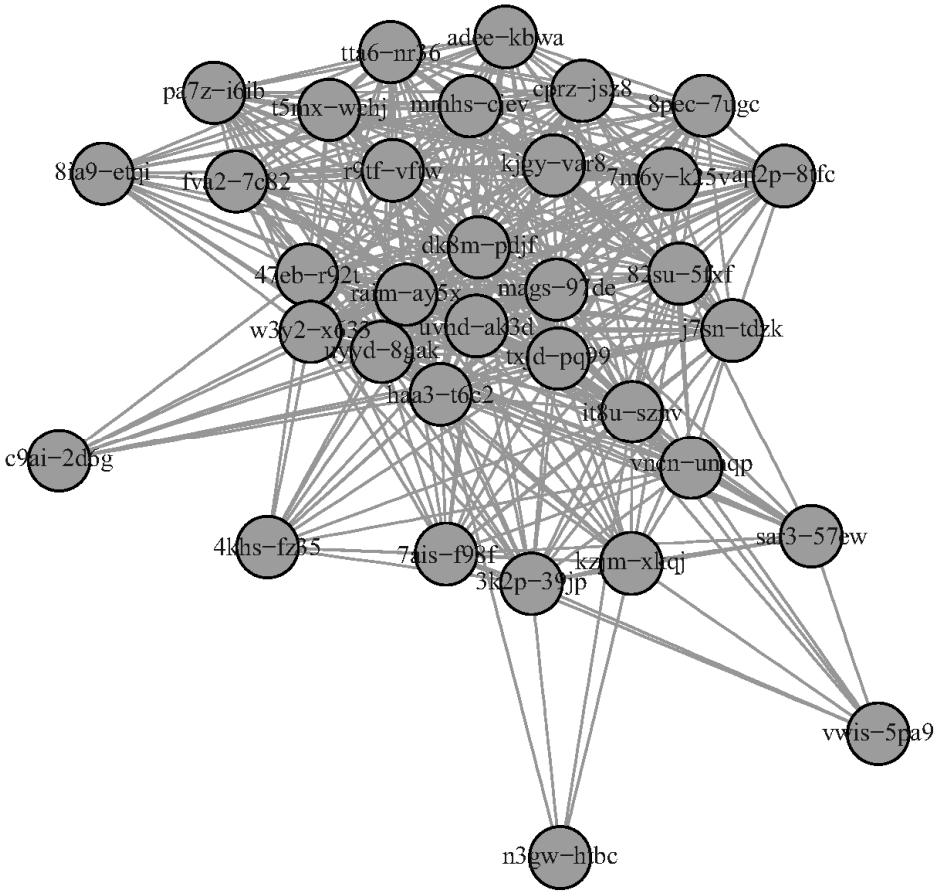

Figure 1: Results of 5-Meter Spatial Join of Latitude and Longitude Column Contents.[111]

On this basis, we conducted a simple spatial join of datasets sharing the field names of latitude and longitude. For this procedure, we drew a circle, 5 meters in diameter around each point in space identified in columns with the heading latitude and longitude (both of which were present in 34 of the 204 tabular datasets available). If the point from one dataset was found within the circle of a point from another dataset, this constituted a join between the two datasets.[112] Joins between two datasets, measured in this

111. Datasets in Figure 1 are represented by circles with alphanumeric identifiers. Datasets are linked to one another in the network graph when six or more location matches, in a 5-meter radius of one another, occur between the datasets. Data collected from all Tabular datasets on the City of Seattle's Open Data Portal, as of April 1, 2015.

112. Analysis was carried out using PostGIS, with an overall program logic instrumented in a combo of Python and Bash. Overview of steps in the analysis:

way, are highly likely to be referring to the same parcel or piece of property. The results are shown in Figure 1.

In the figure, nodes correspond to datasets, and are labeled with the alphanumeric identifiers of datasets used on the Socrata platform. The lines connecting the nodes indicate matches between datasets. Links were removed when the number of matches was less than six, thus all lines indicate more than six matches between datasets. From this visualization one can assume that nearly all tables in this sample of tables (n = 33) will have spatial matches.

The meaningfulness of the match depends on the context of the locations matched. Manual inspection of field names and titles of the sample datasets suggests that the spatial locations matched are perhaps public facilities (e.g., community centers hosting multiple types of events, locations of sensors for data collection such as bicycle and other traffic counts) but also private facilities (e.g., locations undergoing repeated building inspections and permitting procedures, locations identified in multiple events such as 911 calls for police and fire). In this research agenda, the next step would be to conduct more flexible comparisons where, for example, latitude and longitude are geocoded and compared to street addresses or other forms of location information.

c) Attributes on a Continuum of Personalization

In terms of the potential for privacy harm, a very limited scan of attributes amongst datasets, both within and outside municipal open data for Seattle, produced a rather rich set of information for the purpose of profiling individuals. Limited only to three datasets in Seattle and a fourth in King County, these attributes suggest how weaknesses in the ability to

1. Convert lat/lon text strings into WGS84 Geometries (a reference datum used by Socrata).

2. Create new empty geometry field.

3. Translate points into NAD83(HARN) Washington State Plane N format, meter units.

4. Create 5 meter buffer around points. This value was chosen somewhat arbitrarily to allow matches of points that differ only by the floating point precision of the lat/lon. This distance was generous enough to smooth over any minor discrepancies in parcel size, but conservative enough that any identified matches would pretty much be a stones throw from each other.

5. Construct spatial indexes using GiST strategy.

6. Identify matches based on the condition of intersection between any two circular buffers (ST_Intersects function).

7. Return count of matches.

effectively de-identify individuals through the elimination of indexical fields and the aggregation of data across space could result in serious consequences in terms of privacy and social equity.

Table 1: Attributes from Four Open Data Sets on a Continuum of Personalization.[113]

	Datasets				
Fields	**Property Value**	**Tech. User Survey**	**Business License**	**Building Permits**	**Potential Privacy Concern**
Name	●		●	●	Persons
Address/Location	●			●	Persons
Phone Number			●		Persons
Age		●			Persons
Gender		●			Persons
Income		●			Persons
Home Value	●				Persons
Zip Code		●			Persons
Sexual Orientation		●			Groups
Race		●			Groups
Level of Education		●			Groups
Language		●			Groups
Number in Household		●			Groups
Employment		●			Groups
Unpermitted Activity				●	Unknown
Internet Use		●			Unknown
Uses of Cable		●			Unknown
Incident Type/Descrip.			●		Unknown
Permitted Activity				●	Unknown
Value of Alteration				●	Unknown
Permit Type				●	Unknown

The City of Seattle datasets represented in Table 1 include permitting data from the Department of Planning and Development, Business License Data from Financial and Administrative Services, and the Department of Information Technology's survey of resident uses of information technology (n = 2900 residents surveyed). King County's public dataset showing property ownership and tax assessment is also included. Note the ability to

113. Some fields of Table 1 contain data that may be used to identify persons or infer the identity of persons. Some fields contain data that may be used to categorize persons into racial, social, or economic groups. Government data contains many additional fields of data with as yet unknown implications for privacy. Fields that may be used to identify (e.g., name, address) or infer the identity of persons (e.g., age, gender, zip code) are indexical, and can be used to join these data into one universal set to form dossiers on individuals, groups of individuals, households, and neighborhoods.

join the property value, business license, and permitting databases using the names of the property and business owners. This one act brings together name and contact information, such as address and phone number.[114] While there is no obvious overlap of fields between the technology user survey, and other datasets, it is worth noting that one of the more popular and widely used indexical fields for re-identification is zip code. With the plethora of demographic fields provided in the survey dataset, it is not difficult to imagine a data broker or similar type of firm using zip code to join and re-identify survey respondents. At the very least, the privacy implicating and highly differentiated fields in the survey could make this dataset a desirable target for commercial interests seeking to re-identify subjects and enrich their existing dossiers on city residents.

d) One Simple Example of a Profile

Finally, to demonstrate the kind of personal profile which can be gathered today from open data published by the City of Seattle, we chose a single location and produced joins from eight Seattle open datasets. The information gathered from these datasets revealed:

1. Property owner's full name (multiple spellings)
2. Multiple major building projects, most with associated code violations related to follow-up and/or inspections
3. Junk storage violations
4. Vacant building-related issues
5. A fire in the main structure

There is enough information in any one of these datasets to join this profile with the King County dataset that shows the assessed value of property, which may be used as a proxy for wealth or income. The property is among those in the city that have received the lowest possible valuation.

There is distress involved in some of the revealed incidents as well as loss of personal property and net worth, all tied to dates, times and a specific person's name. The level of information revealed from the combination of these eight open data sets—all indexed using spatial location—is more than most individuals would be comfortable with.

5. Open Data Assessment in Sum

These technical assessments suggest the extent to which the release of multiple, seemingly benign municipal open datasets holds the potential to compromise privacy, or pose threats to social justice. The City of Seattle,

114. This emphasizes the importance of excluding licenses for businesses located in residences from open data.

however, like many cities in the U.S., governs many more datasets than those currently available as open data. Many of those datasets are produced, processed, copied, and stored in the information systems of firms under contract with the City.

D. LEGAL ASSESSMENT: VENDOR CONTRACTS

The preceding section describes risk as a function of technical processes, demonstrating how data that is "safe" in isolation may yield more private details than anticipated when combined or correlated. In this section, we describe risk of another sort: the risk associated with turning over the processing and storage of resident data to third party vendors. Cities use vendors extensively. And vendors have different capabilities and incentives than a municipal government; they may be more or less capable of keeping data secure, and are not likely to be as responsive to residents as their city government. As our qualitative analysis makes clear, stakeholders will ultimately hold cities responsible as custodians and expect them to uphold constituent values.

The relationship between the City of Seattle and its vendors is described in its contracts. We therefore undertook an analysis of a carefully selected sampling of contracts between the City and its vendors. The goal of this research was to determine whether vendors with access to City data—including data about employees and citizens—were contractually obligated to engage in best practices around privacy and security, thus preventing the unintended spilling of data. We found that some were, and others were not. This does not necessarily mean that any vendor engages in bad behavior, only that they do not make commitments that help foreclose the possibility. On the basis of this work, we later recommend that the City generate a standard contract including privacy and security language to use as a starting point for any future outsourcing of data processing, gathering, or storage.

Among the insights we gleaned from our focus group sessions were that residents did not tend to differentiate between the specific constructs of open government or public records requests and the city's role in general as a custodian of resident data. The city collects, stores, processes, and in some instances shares information. Although we have developed a taxonomy of push, pull, and spill in this paper, the picture for residents seems rather less differentiated.

In general, we found that relatively few vendor contracts made guarantees around the privacy or security of resident or employee data, and that the contracts that did make such guarantees did not use anything like the same language. There was no "smoking gun," in the form of a highly irresponsible provision, but there were places where due diligence might

have recommended changes to allay stakeholder fears and concerns. The findings that follow form the basis of our recommendation, *infra* Section IV.G, that the City develop a standard vendor agreement that incorporates baseline or default provisions regarding how information is accessed, shared, and secured.

Residents want to feel as though cities are using information wisely to their benefit across the board. Cities do not collect, process, or store information on their own. Like all major enterprises, they work with partners. Accordingly, the circle of trust regarding municipal data is wider than just a city itself—it includes their providers. Cities entrust resident data to providers for a variety of purposes, including storage, analysis, and connectivity. For example, the City of Seattle Police Department works with Evidence.com—a subsidiary of Taser—to store video from police lapel cameras. Seattle employees work with Verizon and Motorola to communicate. As noted previously, the City's existing open data portal is managed by Socrata.

The primary means by which cities can maintain its trust with residents in light of these partnerships is by getting these providers to agree to a comparable level of responsibility and data hygiene. Indeed, the city's relationships with vendors are governed by terms of service, privacy policies, and other service agreements.

We undertook to examine these documents in an effort to assess whether they respect privacy and security by their terms. Our method involved selecting eighteen particularly important master agreements (plus sub-documentation) from five departments. We based this selection on the in-depth interviews we conducted with employees across the City. An attorney in private practice analyzed the documents according to parameters set by a member of our team with deep experience in privacy law, specifically including privacy policies and terms of service. That team member then reviewed and synthesized the findings for presentation here.

1. Privacy

We first looked for language addressing what if any rights the subjects of data being processed by the City's partners may have. In the consumer privacy context, such rights generally include understanding what information has been collected and why, how it is secured, with whom it is shared, and so on. A good benchmark is the set of obligations imposed on websites under California's privacy notice law.[115]

115. *See* CAL. BUS. & PROF. CODE § 22575 (West 2014). *See also* CALIFORNIA ATTORNEY GENERAL, MAKING YOUR PRIVACY PRACTICES PUBLIC:

The picture on privacy was mixed. Whereas some providers specifically reference the ability of data subjects to access their data (e.g., Paybyphone, Volgistics, and Microsoft), many others made no reference to privacy or data subjects at all (e.g., Kubra, FileLocal, and MacroCCS).[116] Some agreements assumed a relationship with the data subject: PayByPhone agreed to "provide an easy to use customer account management website."[117] Other agreements seemed to assume that the City would remain the point of contact for data subjects: Microsoft, which hosts and processes a variety of City data, committed *not* to respond to data subject requests absent the City's prior written consent or a legal obligation.[118] There was next to no language obligating vendors to notify data subjects of anything, except in the case of a data breach as discussed in the next section. And long-term retention was, if mentioned, framed as a benefit.

A variety of contracts (e.g., those with CopLogic, Hewitt, and Affirma) addressed the privacy-related concept of "confidential information." Confidential information does not always intersect with the sensitive information of data subjects.[119] For example, the Motorola agreement defines it as "any information that is . . . marked, designated, or identified at the time of disclosure to [sic] as being confidential."[120] However, confidential information can so intersect. CopLogic, a software IT company that services the City's online police reporting system, defines confidential information to include certain "City employee information" such as Social Security numbers or email addresses.[121] Confidential information can also include the vendor's own "ideas, concepts, know-how or techniques," i.e.,

---

RECOMMENDATIONS ON DEVELOPING A MEANINGFUL PRIVACY POLICY (May 2014), https://oag.ca.gov/sites/all/files/agweb/pdfs/cybersecurity/making_your_privacy_practices_public.pdf.

116. Kubra, FileLocal, and MacroCCS jointly service the Washington State Business License and Tax Portal Agency, an online portal to pay for business licenses and taxes for several Washington cities including Seattle.

117. PayByPhone Technologies, Inc. Vendor Contract #2992, § 10 "Ownership and Privacy of End User Information," at 4 (2015) (on file with authors).

118. *See, e.g.*, Microsoft Enterprise Agreement Amendment CTM01E68910, § 9 "Office 365 Security Terms," Subsection (A) Privacy, at 11 (2013) (on file with authors).

119. For two important discussions of the relationship between privacy and confidentiality, see Neil M. Richards & Daniel J. Solove, *Privacy's Other Path: Recovering the Law of Confidentiality*, 96 GEO. L.J. 123 (2007), and Woodrow Hartzog, *Reviving Implied Confidentiality*, 89 IND. L.J. 763 (2014).

120. Motorola Solutions, Inc. Blanket Contract 2592, § 32, subsec. 8, at 14–15 (2011) (on file with authors).

121. Coplogic, Inc. Blanket Contract 2708, § 35.2.1, at 21 (2010) (on file with authors).

information proprietary to that business.[122] Where information is designated confidential it may be subject to special protections by agreement, including the prospect of an audit of the vendor to ensure they are processing the information correctly.

Two agreements discussed internal measures to ensure that only the vendor employees who need access to City data would have it—in general, a best practice in consumer privacy. Microsoft committed that "Microsoft personnel will not use, process, or disclose customer data without authorization," and further that "Microsoft personnel are obligated to maintain the confidentiality of any customer data and this obligation continues even after their engagement ends."[123] Volgistics, too, provided that "Volgistics customer service employees will have access to customer data as needed for the purpose of answering customer support inquiries," and also that "Volgistics accounting staff can only see part of your credit card information."[124] No other contract we sampled limited internal access.

Quite a few agreements mentioned how long information would be retained—a typical subject of privacy policies in the commercial context. Retention terms varied, with longer retention generally framed as a selling point. For example, Socrata, which manages the City's open data portal, advised it would retain City records for six years after the expiration or termination of the agreement.[125] Socrata also provides that it will keep the data at the same geographic location unless the City authorizes a new location in writing. Other contracts provided for the return of the data. For example, Truven, a health analytics company, committed to "provide to the City all City-owned data, property and deliverable . . . in the format originally sent to the Vendor by the City or its Data Sources."[126]

Other agreements discussed the conditions under which City data would ever be shared with a third party. For the most part, the relevant language committed the vendor to hold its subcontractors to the same obligations the vendor has to the City. Language such as Oracle's is common: "Any subcontract made by Vendor shall incorporate by reference

122. Affirma Consulting, Agreement Number CRU 2013-002, § 22, subsec. G, at 11 (2013) (on file with authors).

123. *See, e.g.*, Microsoft Enterprise Agreement Amendment CTM01E68910, § 9 "Office 365 Security Terms," subsec. (A)(e), at 11 (2013) (on file with authors).

124. Volgistics is a company that offers software-based coordination of volunteers, of which the City has many.

125. Socrata, Inc. Blanket Contract 3406, § 27 "Review of Vendor Records," at 24 (2014) (on file with authors).

126. Truven, Vendor Contract 3150, § 41.7.5 "Termination," at 22 (2013).

all the terms of this Contract . . . ."[127] Confidential information, however defined, sometimes enjoyed special protection against disclosure.

Several vendor agreements at least contemplated the possibility of sharing with data with third parties. The Acyclica contract reserved the right for the parties to renegotiate data ownership, "specifically with respect to reselling of data,"[128] whereas Truven required the City to *opt out* of sharing its information with Truven's MarketScan program and, in doing so, give up the "MarketScan contribution discount."[129] We were unable to determine whether the City decided to participate in MarketScan, and we imagine the data would only be shared in the aggregate in any event.

A noteworthy feature of many of the contracts was the treatment of privacy and security; many contracts did not explicitly address privacy concerns by name even though they did so for security. Privacy and security are both important abstractions governing the use of data but are conceptually distinct enough to warrant separate analysis.

2. Security

One of the main concerns of stakeholders—in general, and specifically in our study—is the adequacy of security around data. We are all aware of major breaches affecting even the most sophisticated institutions. Security is one of the venerated Fair Information Practice Principles (FIPPs), which the FTC and others use as a lodestar for privacy policy.[130] A statement of security practices is required for websites operating in California, as alluded to above, and most states impose obligations on data custodians to notify data subjects and the relevant authorities of a breach.[131]

The agreements we sampled and reviewed fared better on security than privacy. Ten out of eighteen specifically reference the adequacy of data security. Several called for security audits or else required vendors to provide documentation of their security policies. Claims of security varied in specificity. For instance, Parkeon simply states it will take "an appropriate

127. Oracle America, Inc. Blanket Contract 3025, § 13b, at 4 (2013).

128. Acyclica Attachment to the Western Systems Purchase Order, § 2.6.1, at 2 (on file with authors).

129. Truven, Vendor Contract 3150, exhibit B § 13(g), at 6 (2013) (on file with authors).

130. FED. TRADE COMM'N, PRIVACY ONLINE: FAIR INFORMATION PRACTICES IN THE ELECTRONIC MARKETPLACE (2000), https://www.ftc.gov/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission.

131. Forty-seven states have laws on the books governing disclosure of data breaches. For a current list see *Security Breach Notification Laws*, NAT'L CONFERENCE OF STATE LEGISLATURES, http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx.

standard of due care,"[132] whereas others offered specific benchmarks. Motorola stated it would treat the city's data as if it were their own, internal data.[133] PayByPhone pegged its standard to the robust Payment Card Industry Data Security Standard.[134] And CopLogic offered an attestation that a security auditor had tested its system for "common security vulnerabilities."[135]

Several companies dealt specifically with the important issue of encryption, i.e., storing or communicating information in ways that would ordinarily be unintelligible if accessed or intercepted by an unintended party.[136] Acyclica, a company that collects and processes traffic data, promised that the City's data would be "encrypted to fully eliminate the possibility of identifying individuals or vehicles."[137] The health analytics firm Truven specified 128-bit Secure Socket Layer (SSL) encryption of some data.[138] Volgistics also uses SSL for data in transit and storage.[139] Finally, Microsoft uses encryption on data and media that is sent on public networks or leaves its facilities.[140] Acyclica, Truven, and Volgistics also refer to the use of de-identification techniques separate from encryption.

Many states, including Washington, obligate companies that experience data breaches to notify consumers and the authorities within a specified time period.[141] Regardless, parties are free to delineate additional, legally

132. Parkeon, Inc. Vendor Contract 1163, Attachment 1 § 5, at 7 (2004) (on file with authors).

133. Motorola Solutions, Inc. Blanket Contract 2592, exhibit A "Data Information Security Services," at 5 (2011) (on file with authors).

134. PayByPhone Technologies, Inc. Vendor Contract 2992, § 13 Security, "Privacy and Compliance," at 5 (2015) (on file with authors).

135. CopLogic, Inc. Blanket Contract 2708 § 16 "Security," at 12 (2010) (on file with authors).

136. We presume many other vendors make routine use of encryption and simply do not mention it.

137. Acyclica Attachment to the Western Systems Purchase Order, § 2.5.1, at 2. This language is probably a little too strong. It may be possible for sophisticated parties to identify people or objects even if encrypted, for instance, by breaking the encryption.

138. Truven, Vendor Contract 3150, exhibit B § 15 "Data Communication," at 6 (2013) (on file with authors).

139. Volgistics Online Form Security and Privacy Policies, "Security Policies," at 2 (2015) (on file with authors).

140. EA Amendment CMT01E68910, § 9 "Office 365 Security Terms," § (D)(a)(v) 4.A., at 14 (2013) (on file with authors). Microsoft encrypts Customer Data that is transmitted over public networks; B. Microsoft restricts access to Customer Data in media leaving its facilities (e.g., through encryption).

141. *See* WASH. REV. CODE § 19.255.010(1). Section 19.255.010(1) states:

> Any person or business that conducts business in this state and that owns or licenses computerized data that includes personal information shall

consistent terms in the event of a security breach and often do so. In the documents we analyzed, we noted that a few vendors committed to notifying the City "immediately" (Socrata) or within one business day (Parkeon).[142]

While state laws may obligate companies to disclose breaches, they do not purport to delineate legal responsibility in the event of a breach.[143] We found that specific vendors attempted to contractually absolve themselves of liability should a breach occur. This could occur generally through an arbitration agreement (e.g., Tokusaku) or vendors could absolve liability quite specifically in the event of a breach. For example, Socrata disclaims *all* damages for loss of data, "whether or not resulting from acts of God, communications failure, theft, destruction or unauthorized access to Socrata's records, programs, or services."[144] In contrast, still other vendors (e.g., Hewitt and Microsoft), provide for credit monitoring or other "direct damages" in the event of a breach. The City itself could be held accountable consistent with sovereign immunity.[145]

3. Analysis

The agreements we reviewed were so-called "enterprise" agreements, i.e., made between sophisticated parties. It would not necessarily be fair to judge agreements between cities and firms against consumer privacy policies or terms of use. Thus, we might not expect the agreements to exactly track the Fair Information Practice Principles of notice, access, choice, and security, or to adhere to the strictures of the California Online Privacy Protection Act requiring every website to identify what data it collects and

> disclose any breach of the security of the system following discovery or notification of the breach in the security of the data to any resident of this state whose unencrypted personal information was, or is reasonably believed to have been, acquired by an unauthorized person.

Id.

142. Socrata, Inc. Blanket Contract 3406, subsec. 5.2.8, at 16 (2014) (on file with authors); Parkeon Vendor Contract 1163, Attachment 1, sec. 5 "Security Standards," at 7 (2004) (on file with authors).

143. *See, e.g.*, WASH. REV. CODE § 19.255.010(1).

144. Socrata, Inc. Blanket Contract 3406, subsec. 17, at 21 (2014) (on file with authors).

145. *See* Kelso v. Tacoma, 390 P.2d 2 (Wash. 1964) (holding that the State of Washington has waived sovereign immunity in tort cases and municipal sovereign immunity); *see also* Locke v. City of Seattle, 172 P.3d 705 (2007). *But see* Cummins v. Lewis County, 133 P.3d 458 (Wash. 2006) (holding that the public duty doctrine still applies to the State of Washington). For a discussion of government liability in Washington see Michael Tardif & Rob McKenna, *Washington State's 45-Year Experiment in Government Liability*, 29 SEATTLE U. L. REV. 1 (2005).

how it is used and safeguarded,[146] even as we employ these standards as benchmarks of best practice.

More so than an individual consumer, however, the City is in a position to dictate the terms on which it will transact. Many of those terms—such as adequate security—should apply in all of the City's dealings around resident or City data. What we most clearly observed in the vendor contracts was a lack of standardization. The city reserves very disparate rights against its various vendors, and receives a wide range of positive guarantees. Privacy basics—such as notification requirements, security standards (including encryption), and internal safeguards against unauthorized access—were not specifically delineated in many instances. Companies like Volgistics and Microsoft made extensive mention of privacy and security, laying out exact terms. But other companies made almost no mention of these.

This reflects the status of cities as market makers, not market takers. Law is not the only modality of regulation. Another is markets: cities can and will drive business decisions because they are major potential customers. An insistence that municipal vendors in the data space agree to basic commitments around privacy and security can make city and citizen data more secure all over the country by raising the market bar.

IV. RECOMMENDATIONS

The Article thus far has described the expectations around, and inner workings of, Seattle's open government initiative and other data processes. A final section outlines some tentative recommendations on the basis of what the team has learned. Though researched for the City of Seattle, the practical nature of the seven recommendations shown in this section could be considered valuable to any municipality seeking public trust, privacy, and social justice on the road to open data.

A. INVENTORY DATA ASSETS

Our first recommendation involves creating a complete inventory of datasets, the fields within those datasets, and metadata explaining how the information was collected, its purpose and use for the municipality, and any other relevant descriptors concerning the proper management and disposition of the data.

While much of this Article has focused on the contents of datasets, the topic of metadata should not be ignored. Metadata can provide the municipal organizations charged with governing data release with

146. CAL. BUS. & PROF. CODE §§ 22575–22579 (West 2014).

information critical to understanding and hopefully acting within the municipal and decidedly public context for the data.[147]

A common standard for metadata is the Dublin Core, a list of categories useful for storing and classifying data. The fifteen fields that comprise the core include: title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights.[148] Amongst these categories are many fields for metadata that are potentially valuable for storing, among other things, records that explain the purpose of collecting the data on the part of the responsible department or office, the public uses of the data, a description of the anticipated public benefits of those uses, the classification of the data (e.g., sensitive, critical infrastructure), the nature of the subjects, the sensitivities of the data, restrictions on releases, requirements for aggregation prior to release, suggested qualifications for note in exemption logs in reply to public disclosure requests, a list of the third parties allowed access to the data, the allowable uses or restrictions on use of the data by those third parties, required security measures, applicable regulations, and a note explaining the ex ante and ex post analyses of risk to privacy and social justice conducted in relation to the distribution of the data.

Metadata includes field names. As our technical analysis highlights, municipalities and their related government offices (i.e., counties, special districts, states) should develop and share a data dictionary—a standardized nomenclature for data fields and entries. This tool can provide multiple efficiencies. It can assist departments and the public in interpreting and using municipal data. Departments will find it easier to locate and identify existing information. It can also reduce the chance that work would be unnecessarily duplicated, as would occur if someone found it difficult to find or properly interpret the datasets that already exist.

A more exact and shared naming convention can also reduce the time and effort needed to determine the risk of harm in releasing datasets to the public. In the case of our research, several of our technical strategies were designed to simply deal with the fact that no shared lexicon currently exists for the field names used by municipal departments. "Address" is just as likely to appear as "ADDR," "Street Address," and "Location," and the difference creates unnecessary hurdles for ex ante analysis of risk of release. Any effort

147. For a definition of metadata, see KITCHIN, *supra* note 102, at 8.

148. *See About Us*, DUBLIN CORE METADATA INITIATIVE, http://dublincore.org (last visited Sept. 1, 2015) ("The Dublin Core Metadata Initiative (DCMI) supports shared innovation in metadata design and best practices across a broad range of purposes and business models."); *Dublin Core Metadata Element Set*, DUBLIN CORE METADATA INITIATIVE (June 14, 2012), http://dublincore.org/documents/dces.

that has to be spent to interpret the existing data is effort that could be saved and spent elsewhere.

B. REQUIRE EACH UNIT TO DEVELOP AND SUBMIT DATA POLICIES

For cities trying to thread the needle of protection for private and social information while enjoying the ability to make other sets of data available to the public, operating as a federated system has its benefits and its drawbacks. Departments in a federated system will have a diversity of strategies that have evolved to implement the policies they have each created and tackle the problems they have each encountered. Revealing the possibility to one department that they may emulate a practice in another may be just the thing to assist departments. In Seattle, for example, some departments appeared to be more comfortable than others sorting meaningful from frivolous examples of public disclosure requests, and denying requests with an explanation filed in their exemption log.

For Seattle, with newly adopted privacy principles, this is an opportune time to learn about the variety of policies departments have already been exercising that, whether they realized it or not, have had the effect of preserving or compromising privacy and social equity. The Department of Information Technology and the Mayor's Office are intent on delivering a citywide privacy policy. The successful implementation of such a policy will depend on the ability of people in these departments to discern the degree to which each department is already delivering practices that preserve privacy and social equity, and to focus attention where it is needed to assist departments that may feel overwhelmed by the shift in priorities.

Consider, in this light, the contrast in notice and consent provided to the residents of Seattle from the Department of Transportation's enlistment of the services of Strava and Acyclica. One need not observe the presence of a field name in a dataset to realize that the data can be used to identify persons. As Montjoye et al. have shown, in their analysis of hourly information flow from devices which record and track the movements of people in time and space by keying in to the MAC address of personal devices (similar to those deployed in Seattle), the traces of mobility left by persons across the urban landscape are highly unique.[149] With only four data points observed in a day, 95% of MAC addresses and persons can be identified. Within the spatial scale of a municipality the task of re-identification is further eased by the classification of municipal land use into residential, office, and other forms of commercial space. Only one, or

149. *See generally* Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, SCI. REP. (Mar. 25, 2013), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3607247/pdf/srep01376.pdf.

perhaps two data points would be needed to identify most individuals: the location at time of day when statistically likely to be in residence, and the location at time of day when likely to be at school or work.

With these facts in mind, notice and consent would seem to be among the prudent cautionary measures necessary for preserving public trust in the privacy-preserving efforts of the Department. Strava's application does not capture the data flow of the entire population, and as an opt-in program the data has limitations, yet it is data that participants agree to provide and it has proven useful to the Department for the study of travel behavior. Acyclica's data covers more of the population and this fact is due to the lack of notice, choice, and related attention to privacy that accompanied the installation and contractual arrangements for Wi-Fi and Bluetooth sniffers in the public spaces around Seattle. The City can create and test new avenues for notice, consent, and choice. People can opt-out of the program if they are aware of it and capable of following the instructions to do so. The City can also adopt more restrictive policies for permitting the distribution of devices for surveillance in public space.

The next step for the City is to ask how important is the public use for which this data is collected, and who should make this determination? If the public use is deemed valuable enough to the taxpayer (including all ancillary costs envisioned to make the data secure), the next question to ask is how relevant this data is—in its entirety—to the public uses for which it is collected. One can question the need for a sample of this size, the frequency of the collection, the granularity and choice of spatial collection, and of course, retention and distribution of the data. If used, for example, for traffic operations on congested arterial streets, and such use is sanctioned by the public or elected representatives, then the obvious condition that should follow is the limitation of the spatial extent of collection. There is no need for traffic operations to include the monitoring and evaluation of travel behavior in the residential zones of the city, where the ease of personally identifying individuals on the basis of time and location is most likely. Like the black-out dates that airlines have employed to prevent the use of discount travel during peak periods, municipalities should adopt black-out zones, to prevent the use of personally identifying surveillance technologies.

What these two cases suggest is also the extent to which a federated system lends itself to an ad hoc approach to problems that are holistic in nature, such as the problem of analyzing the potential privacy and social equity harms involved in data releases. For this, a governance structure is needed.

C. ESTABLISH NESTED GOVERNANCE STRUCTURE

Municipalities need structures to more effectively govern the releases of data, via push, pull, and spill. They need governance structures that operate on more than one level, that emulate the need to coordinate and provide some hierarchy to the complex decisions that municipalities must make through the release of data.

A nested governance structure could help municipalities develop citywide policies and avoid ad hoc decision-making. Such a structure could involve oversight from a municipal decision-making body analogous to an Institutional Review Board (IRB), which are convened to review proposed academic research involving human subjects. At the department level, such a structure would include clear guidance about the types of activities that would be exempt from review by the municipal IRB. The activities of interest would span the life cycle of data, to collection, use, retention, deletion, as well as release. Activities that are not exempt would be elevated for review by the IRB.

Emphasizing the importance of informed, meaningful consent, Barocas and Nissenbaum explain that notice and consent are most effectively refined through the services of such a review board.[150] In their explanation, they borrow from the literature on human subject research in medicine, applying these basic insights to the broader case of notice and consent for privacy.[151] They acknowledge that patient interactions take place against a backdrop of trust, and that consent or waiver should be interpreted narrowly. Quoting O'Neill and Manson, they explain that obligations and expectations of medical service providers are not discarded when patients consent. Consent is requested of subjects in limited ways, for limited times and very specific purposes.[152] In consenting to an appendectomy, one does not consent to other incisions, or to incisions by persons other than the relevant surgeon. Furthermore, consent is not required for expected behaviors; it is required for behaviors that depart from what is expected. The burden is on the researcher or clinician to, in giving notice, "describe clearly the violations of norms, standards, and expectations for which a waiver is being asked."[153] In applying these insights to the more general problem of privacy amidst big data, the authors suggest, "[a] burden is upon the collector and user of data to explain why a subject has good reason to consent, even if consenting to

150. *See* Barocas & Nissenbaum, *supra* note 97, at 64.
151. *Id.* at 44–75.
152. *Id.* at 64–65.
153. *Id.* at 65.

data practices that lie outside the norm. That, or there should be excellent reasons why social and contextual ends are served by these practices."[154]

In the case of Seattle, we have sought to illustrate the contextual circumstances that surround the municipal rush to big data and open data. Several departments are adopting technologies that collect rich datasets about the people living and working in Seattle. Once collected, the data can be subject to public disclosure request, and may be considered for release to an open data portal. All of these activities can occur in ways that pay scant attention to the potential effect on privacy or social justice from releases of data. For example, the "8 Principles of Open Government Data," used to structure the review and release of business license data by the Departments of Information Technology and Financial and Administrative Services, were designed for the purposes of promoting the release of data. In this system of reviewing and releasing data, there is no equivalent guidance in practice to safeguard privacy and social justice.

The process of data review and release is devoid of the contextual and subject-oriented privacy protection that Barocas and Nissenbaum define. Practices to safeguard privacy and social justice are, in the current process, reduced to the evaluation of individual fields within isolated datasets. Given this, it is no wonder that public trust in the privacy-preserving actions of municipalities remains suspect. We suggest the adoption of a municipal IRB, tasked with protecting privacy and social justice, with the authority to veto and condition the collection, use, and release of data, and the interdisciplinary capability and experience to evaluate the public interest in such decisions. Given the countervailing interests of open data and privacy, it is worth mentioning that these aims should not be the responsibility of the same person or division within a city department.

IRBs, however, are not needed in every case of review, and the Department of Information Technology may seek to produce a list of datasets and their fields that may be handled through administrative review within the department that owns the data, or exempted from review altogether. Municipal IRBs should be called into service only when the data subjects are employees of the city, residents, or workers. The IRB can be asked to review requests from departments for public release of data to portals or online platforms and any accompanying supportive analysis, such as an analysis of the nexus between the collection of the data, its public uses, the interests of the taxpayer, and privacy and social justice implications. The City should also consider using the IRB to evaluate public disclosure requests that pose privacy or social justice problems, for which there are no

154. *Id.* at 67.

clear exemptions in the PRA. This should result in recommendations rendered on a case-by-case basis, yet informed by a body of knowledge of preceding cases and their outcomes, as well as ongoing research in the rapidly moving field of re-identification.

D. ESTABLISH AND DISSEMINATE EX ANTE PROTOCOLS FOR PUSH, PULL, AND SPILL

Cities should plan for the fact that departments may want to release data by pushing it out to public portals when they should not or that departments may inadequately act or invest to prevent the pull or spill of data. One effective way to do this is to establish and disseminate protocols for investigating datasets, in order to educate departments about how to preserve privacy and social equity by curbing or curtailing certain types of releases.

Our suggestions stem from our study of how multiple databases may be joined after they have been published. Possibly the simplest approach a city could take in a protocol to evaluate releases ex ante would be to programmatically perform the same kinds of join strategies which our research team did—and perhaps a few others that we did not have time to develop. The join strategies would illustrate the overall joins made possible with other public datasets (and private ones if available) if the proposed new data were to be published. This method would result in two useful artifacts:

1. The resulting joined dataset, which could highlight newly harmful combinations of data made possible with the introduction of new data to the existing corpus of publically available data.
2. A network map that shows precisely which fields would be used to accomplish joins resulting in privacy harm.

The same method could be used to discover and eliminate existing indexical fields, which cause the greatest degree of correlation across the continuum of privacy related attributes in existing datasets. By adopting this practice, and relying on as many existing datasets as possible, the City of Seattle can reduce the likelihood of, and thus manage the risk associated with, the joining of independent datasets in ways which may cause privacy harm.

E. CONDUCT PUBLIC RECORDS AUDIT AND TRAINING

We recommend based on the above that cities engage in audits and training exercises whereby municipalities compare the text of state and federal public records acts with what individual departments are doing on the ground. In the case of Seattle, the City has protocols in place, by

department, on how to respond to PRA requests. However, it is important for *all* employees—not just those with responsibility for responding to outside requests—to understand the law and the City's interpretation of the law. This will help reduce uncertainty and fear around the prospect of abusive pulls or spills of employee data.

In our engagements with City employees, we noticed variation in the understanding and application of public records requests. First, as noted, not all departments adopted the same posture toward a request for information. Parks and Recreation, which deals mostly with children and families, adopted a relatively restrictive stance.[155] The Police Department had to come up with entirely novel procedures to accommodate massive requests for information in the form of video recordings, and defaulted toward sharing everything (with some modifications for privacy).

We also noticed that employees articulated fears about abusive behaviors that should not have been possible under the text of the PRA. The act provides an exception, for instance, for personal information about an employee.[156] Nevertheless, employees worried that other employees or the public would gain access to information for the purposes of relationships, bias, or embarrassment. When the PRA exception for employee personal information was pointed out in an interview, the room erupted in laughter, as if to suggest the exception would not be honored.[157]

This is not to say that any city should ignore the role of context—it may be a good thing that departments do not all react identically to a request for information. However, there should be some standardization. In particular, all employees involved in responding to public records requests should know the exceptions and the reasons behind them, and generally be able to fall back on a clearly articulated policy.

F. EXPLORE CONDITIONED ACCESS OF MUNICIPAL DATA

We recommend that cities explore vehicles by which to make certain data available under specific conditions. This is a fairly common practice. Companies, of course, routinely condition access to information on signing a nondisclosure agreement. In the public sector, more than twenty states condition access to voter databases on noncommercial use.[158] Federal

155. Interview, Seattle Parks and Recreation Personnel, Seattle, Wash. (Mar. 5, 2015).

156. WASH. REV. CODE § 42.56.230(3) (2014) ("The following personal information is exempt from public inspection and copying under this chapter: . . . Personal information in files maintained for employees, appointees, or elected officials of any public agency to the extent that disclosure would violate their right to privacy.").

157. Focus Group, City Employees, in Seattle, WA (Mar. 9, 2015).

158. *See, e.g.*, WASH. REV. CODE § 29A.08.720(2). That section directs:

election law has similar provisions. As cities open up more and more data, they should consider whether one or more use restrictions would be appropriate.

In our focus groups, several citizens and most privacy advocates expressed concern over the prospect that the City would push data for transparency reasons that would instead be used for commercial or political purposes that were disadvantageous to consumers and citizens. Examples included lenders writing off neighborhoods with respect to offers of credit and politicians ignoring complaints from districts with low political participation.[159] There is ample evidence that municipal open data is a major source for data brokers of all kinds.[160] One opportunity might be to follow the example of some states and federal agencies around political data and condition access to certain data sets on noncommercial or nondiscriminatory use. A government might do this when, for instance, citizens may be less likely to participate in a given, beneficial activity such as voting, donating, or volunteering because they fear it will lead them to be targeted for marketing or otherwise cause them to face adverse commercial consequences.

Another example might be conditioning access on the obligation to update the information periodically. The issue here is that commercial entities may copy databases that then become outdated, either because of a mistake (false lien) or because of an update (juvenile record expunged). Meanwhile, although the City now has the correct version, companies and others may be making decisions on the basis of a copy in the hands of a data broker. Presently nothing, apart from industry best practice, obligates these data brokers to keep their databases up to date.

It should be noted that there are a number of pitfalls with this approach. The first is that once data has been released, it is hard to follow. The City

> The county auditor or secretary of state shall promptly furnish current lists of registered voters in his or her possession, at actual reproduction cost, to any person requesting such information. The lists shall not be used for the purpose of mailing or delivering any advertisement or offer for any property, establishment, organization, product, or service or for the purpose of mailing or delivering any solicitation for money, services, or anything of value. However, the lists and labels may be used for any political purpose.

Id. For a summary of state-by-state codes on conditions pertaining to voter list access, see *Voter data use terms and conditions*, NATION BUILDER, http://nationbuilder.com/voterdata; *see also* Kim Zetter, *For Sale: The American Voter*, WIRED (Dec. 11, 2003), http://archive.wired.com/politics/security/news/2003/12/61543?currentPage=all.

159. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015).

160. *See* FED. TRADE COMM'N, *supra* note 9.

might attach rules to its own data but it would have to think through what happens downstream. Imagine, for instance, a condition that commercial users of political data must certify that they will periodically update that data. What if a noncommercial user—a political accountability non-profit—downloads and reposts the data without restrictions? The City would have to look for examples—for instance, in intellectual property licensing—for language that follows the data.

The second is that recent Supreme Court precedent limits the sorts of restrictions that governments can place on uses of data. In *Sorrell v. IMS Health*, the Court invalidated Vermont's attempt to restrict pharmaceutical companies ability to use doctors' prescribing history for marketing purposes—a process called "detailing."[161] The Court found Vermont's attempt to prevent such targeting to be an unconstitutional restriction on these companies' speech.

Note that Vermont did not merely condition access to prescription information on using it for a noncommercial purpose. It singled out particular speakers to silence. According to the Court, "Vermont's law enacts content- and speaker-based restrictions on the sale disclosure, and use of prescriber-identifying information."[162] Specifically, the Court found that "the statute disfavors specific speakers, namely pharmaceutical manufacturers."[163] Thus, the Court concluded that the law ran afoul of constitutional prescriptions of discriminating against viewpoints. Had the state instead kept the data itself and released it only on the condition that it not be used for commercial purposes, the Court might not have taken issue.

In general, there may be situations wherein the City wants *some* types of commercial activities—such as the development of a helpful app by a for-profit start up—but would like to avoid others—such as profiling for marketing. These sorts of restrictions are not likely to survive constitutional scrutiny in light of *Sorrell* and other precedent.[164]

G. DEVELOP STANDARD VENDOR AGREEMENT

We further recommend that the City of Seattle—and others, as well—create a standard vendor agreement to use as a baseline in all future contracting around City data. This agreement would lay out in clear and simple language the obligations that the vendor takes on by virtue of its custody over City data. These include:

161. Sorrell v. IMS Health, Inc., 131 S. Ct. 2653, 2672 (2011).
162. *Id.* at 2663.
163. *Id.*
164. *See id.*; *see also* Discovery Networks v. City of Cincinnati, 507 U.S. 410, 424 (1993) (holding that governments may not ban speech merely on the basis that it is commercial).

- maintaining the confidentiality of data subjects;
- restricting access to those within the organization that need it;
- documenting basic digital and physical security;
- specific notification provisions in the event of a security breach;
- specific delineation of responsibility and liability in the event of a security breach; and
- obligations not to share data in any format absent the express consent of the City and/or the data subject, or by required operation of law.

The suggestion is not that the City would use the exact same agreement in each instance. We recognize that department needs will vary on the basis of the task. Moreover, there may be circumstances when the City or a vendor will need to insist on differing terms. Rather, we recommend the development of a baseline reference document such that any departure would have to be specifically justified.

Models for such contracts already exist. For example, Microsoft has a master service agreement around privacy and data security as part of its own vendor toolkit.[165] Moreover, there were specific contracts—in particular, those of Volgistics and Microsoft—that contained much of the recommended language already. And contracts can and do refer to pre-established standards of security such as PCI—which some vendors already mention—and Internal Organization for Standardization and International Electrotechnical Commission 27001 ("ISO 27001") certification. Ultimately drafting a model agreement may be a task best suited to corporate counsel.

An ancillary, though important, benefit of a standardized vendor agreement would be the effect on the overall market for municipal data. Mid to large-size cities such as Seattle with big information needs and access to considerable resources have the potential to be *market-makers*, i.e., to drive the market toward best practices in privacy and security. Our review of vendor contracts suggests that, with exceptions, the market remains immature in this respect. By insisting on a high bar, the City could not only help justify the trust of stakeholders but improve the overall data ecosystem. We would hope that the City would share any materials it developed with other municipalities.

165. *See Supplier Privacy Toolkit*, MICROSOFT, http://www.microsoft.com/about/companyinformation/procurement/toolkit/en/us/requirements.aspx (last visited July 21, 2015).

V. FUTURE WORK

This research was motivated by three central questions: does the City of Seattle's open data initiative increase the public trust in city government; what kind of legal framing could the City use to capture the benefits of open data while addressing legitimate privacy concerns; and what other kinds of harms could arise from government release of data? This Article is a first step, and much work remains to be done.

This case study points toward promising future work in the area of open data research for municipalities and other related governmental entities. Among the research questions raised, we highlight the following:

Municipalities exist to represent and serve the public, and their departments and offices generally share a keen interest in providing benefits to the taxpayer, in the form of efficiencies as well as public goods. If open data does indeed provide taxpayers with an efficient vehicle for transparency and accountability, then there is no reason to question the validity of the movement to open data. And yet, the activities the City recorded in data collection and released for public and perhaps commercial uses were just as likely to focus on residents as they were the government. What public good is served when the names of people given building code violations are made public? What service is improved by publicizing the names of people applying to participate in pea-patch gardening projects? What is the public benefit of tracking the movements of people through their devices across the neighborhoods of the city? In a more striking case, consider police body-worn video. The shocking videos of shootings that raise public attention toward the activities of police capture, often in full view, the officer as well as the suspect. We are shocked in witnessing, during the course of the video, how a suspect becomes a victim. When the body worn video (recorded on cameras that face forward from the chest or shoulder of the officer) provide little more than moving pictures of the residents of the city, one has to ask whether this technology genuinely serves the purposes of transparency and accountability. If the electric eye is observing only one of these parties, what purposes does this fulfill?

If we presume that the rationale for data collection and use is valid, then the question of efficiency comes into focus. As Aaron Wildavsky has noted, efficiency does not tell you where to go, only that you should arrive there with the least effort.[166] On the grounds of efficiency one could question whether the use of advanced information technology—with sensors that detect, discern, and develop thick flows of information in real-time—delivers on its promise of efficiency to the taxpayer. What is the empirical

166. AARON B. WILDAVSKY, SPEAKING TRUTH TO POWER 131 (1989).

evidence that big municipal data collection followed by big data releases (or big exemptions from releases by State Legislatures), pushed, pulled or spilled, make the City more efficient? When public representatives adopt open data policies, releasing data to the wild, we shift the rules of the game by making private information public. What are the full economic consequences, and how are they distributed amongst the public (who are often the subjects of the data), commercial firms (who often request access to the data about public subjects) and the municipality (whose aim it is to represent the public interest)? What are the distributional consequences—does release heighten or relieve the public of its oft-laden position at the lower end of information asymmetry?

The push, pull, and spill of data from municipalities can predispose the general public and public employees to harms of privacy and social equity. With what legal framework might cities be capable of remedying these harms, and navigating the contested space of data control and release? Much in the case of Seattle may hinge on legal frameworks established by the selective intervention of special interests (public and private) in the adoption of exemptions to the Washington State PRA at the state level, in addition to various privacy-facing federal acts, such as HIPAA. Selective intervention in the rules of the game of state disclosure law suggest that the existing legal framework for balancing privacy and open data is somehow flawed, and this doubt is redoubled through empirically powerful examinations of the inability to use existing legal frameworks—predicated on achieving anonymity by replacing or redacting PII—to protect information that people prefer to keep private. What legal remedies exist, and if they were more widespread, would they be sufficient? What remedies should exist, and how will we know when they are effective?

We've said a lot here; clearly, there is more to be said on the subject.