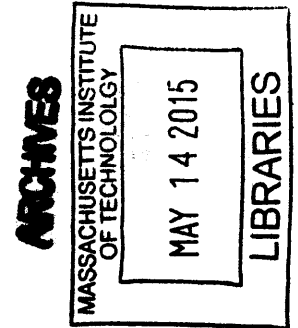# Open Government Data Intermediaries:
Mediating Data to Drive Changes in the Built Environment

by

Suruchi Dumpawar

B.Tech., Visvesvaraya National Institute of Technology, Nagpur (2006)
PGCPD, National Institute of Design, Ahmedabad (2008)

SUBMITTED TO THE
PROGRAM IN COMPARATIVE MEDIA STUDIES/WRITING
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN COMPARATIVE MEDIA STUDIES
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2015

Signature of Author: _Signature redacted_

Comparative Media Studies/Writing
May 8, 2015

Certified by: _Signature redacted_

James Paradis
Robert M. Metcalfe Professor, Comparative Media Studies/Writing
Thesis Supervisor

Accepted by: _Signature redacted_

T.L. Taylor
Associate Professor, Comparative Media Studies
Director of Graduate Studies, CMS

Open Government Data Intermediaries:

Mediating Data to Drive Changes in the Built Environment

by

Suruchi Dumpawar

Submitted to the Department of Comparative Media Studies

on May 8, 2015, in partial fulfillment of the requirements for the degree of

Master of Science in Comparative Media Studies

Abstract

In recent years open data initiatives, which make government data publicly available in a machine-readable format for reuse and redistribution, have proliferated, driven by the launch of open-data government initiatives such as *data.gov* and *data.gov.uk*. Research on open data has focused on its potential for governance, its implications on transparency, accountability, and service delivery, and its limitations and barriers of use. However, less attention has been focused on the practices of data intermediaries–an emerging configuration of actors that plays an essential role in facilitating the use and reuse of data by aggregating open government data and enhancing it through a range of data practices.

This thesis will assess the data practices of open government data intermediaries from three perspectives. First, it will trace the development of open government data initiatives to contend that at a moment when open data policy is seeing global diffusion with the potential of increasing social, political, and economic impact, there is a crucial need to assess the practices of intermediaries to understand how open government data is put to use. Second, it will develop a framework to analyze the role of open government data intermediaries by proposing a definition for "the data intermediary function" constituted by a range of technical, civic, representational, and critical data practices. Third, it will assess the data practices of two open government data intermediaries, *596 Acres* and *Transparent Chennai*, who as urban actors facilitate the conversion of open government data into actionable information for communities to effect changes in the built environment. In describing and assessing the tools, practices, and methods developed by open data intermediaries this thesis will explore the potential and limitations of data intermediaries, and offer recommendations that might inform future open government data initiatives that seek to mediate open government data to facilitate changes in the built environment.

Thesis Supervisor: James Paradis
Title: Robert M. Metcalfe Professor, Comparative Media Studies | Writing

Contents

List of Figures

Acknowledgements

Introduction

Open government data initiatives, which make government data publicly available in a machine-readable format for reuse and redistribution, have proliferated in recent years beginning with the launch of initiatives such as *data.gov* in the United States in 2009 and *data.gov.uk* in the United Kingdom in 2010.[1] In 2013, the G8 countries released the "Open Data Charter," pledging to follow a set of principles to make government data "open by default," and "usable by all," and in 2014 the G20 countries featured open government data as an important focus area, recognizing its potential to tackle corruption.[2] Needless to say, open government data as a policy agenda has spread worldwide, and at present more than a three hundred local and national open government data initiatives, with varying levels of capacities and impact, are operational in countries across the world.[3] These initiatives vary significantly in scale and scope, and involve a range of actors, including local and national governments, non-profits and for-profits, and community-based and civically oriented organizations.[4]

Research on open data has focused on its potential for governance, its implications on transparency, accountability, civic engagement, and service delivery, and its limitations and barriers of use. In *The Responsive City*, Susan Crawford and Stephen Goldsmith, provide a compelling narrative of the influence of open government data on the ways in

---

[1] See Chapter One for a more detailed overview of the growing institutional impact of open government data initiatives.

[2] Tim Davies, "Open Data Barometer Global Report – Second Edition," World Wide Web Foundation and Open Data in Developing Countries (2015): 6, accessed April 13, 2015, http://barometer.opendataresearch.org/.

[3] Ibid. The Open Data Barometer report divides open government data initiatives into four clusters namely: high capacity, emerging &advancing, capacity constrained, and one-sided.

[4] Tim Davies, "Open Data: Infrastructures and Ecosystems," Open Data Research (2011): 1, accessed March 18, 2015, http://ssrn.com/abstract=2549734.

which city governments function. They contend that with the use of open government data city governments are able to: support city managers with data that facilitates better decision-making, forge connections between different city departments for greater collaboration, and predict problems by evaluating historical data and thus optimize service delivery. According to Crawford and Goldsmith, open government data is set to have an impact on urban governance, and as a result city governments could become more responsive to citizen needs.[5] In a similar vein, Anthony Townsend discusses the changing nature of urban and planning and design in the context of emerging technologies, including the ubiquity of data. He discusses the emerging configuration of actors, including civic hackers and technologists, that use open government data to develop applications and services for civic and social good.[6] As the discourse on the wide-ranging impacts of open government data continues to propagate, there have been some efforts to understand its economic potential. According to a report by *Mckinsey*, a global consulting firm, open government data has the potential to unlock $3 trillion across seven sectors globally.[7]

As a result of these discourses, the prominent narrative surrounding open government data is that it has an immense potential to have social, civic, and economic impact. But as research on open government data evolves, more and more attention is being focused on the gap between the expected impacts and potential of open government data,

---

[5] Stephen Goldsmith and Susan Crawford, *The Responsive City: Engaging Communities Through Data-smart Governance* (San Francisco: John Wiley & Sons, 2014), 17-37.
[6] Anthony M. Townsend, *Smart Cities: Big data, Civic hackers, and the Quest for a New Utopia* (New York: WW Norton & Company, 2013), 241.
[7] Michael Chui et al., "Open data: Unlocking Innovation and Performance with Liquid Information," McKinsey & Co. (2013): 2, accessed February 12, 2015, http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information.

and the reality of its current use. I will briefly tease out two strands of critiques here: the utility and usability of open government data, and the implicit biases and privileges in datasets.[8]

Tim Davies has argued that open government data initiatives are primarily driven by the objective of making existing public sector data available in technically and legally reusable form, which he thinks is a very limiting conceptualization of open government data.[9] According to Davies, to begin to assess the impact open government data initiatives we need to consider not only the opening up of government datasets but also focus attention on data practices that support and encourage the use of open data. In order to do so, he contends that research on open government data needs to focus on the use-centric view of open government data rather than dataset-centric view, which fetishes the release of datasets as an end-all.[10]

Furthermore, while the act of making data open can be considered benign, it would be incorrect to assume that open government data in itself is politically neutral. Well-documented cases over the years, such as the undercounting of black and Hispanic households in the United States census, have revealed that government datasets are embedded with privileges and biases, and the dissemination of these datasets and its subsequent use should acknowledge and address the potential social and political nature of these datasets.[11]

---

[8] For a more detailed discussion of these critiques see Chapter Two.
[9] Tim Davies, "Open Data: Infrastructures and Ecosystems,"1.
[10] Ibid.
[11] Jeffrey Alan Johnson, "From Open Data to Information Justice," accessed October, 2014, *Ethics and Information Technology* 16, no. 4 (2014): 265.

To understand how these limitations could potentially be addressed and inform the shape of future open government data initiatives, it is critical to focus on current data practices that engage with open government data and make it intelligible and accessible to communities. To do so, I believe that it is important to understand the data practices of open government data intermediaries—an emerging configuration of actors that plays an essential role in facilitating the use and reuse of data by aggregating open government data and enhancing it through a range of data practices, including but not limited to visualization, analysis, validation and so on. The term intermediaries has found use in discussions of open government data to refer to a wide range of actors that act as immediate users of open government data and facilitate in the process of "making sense of, and creating value out of" open government data.[12] This thesis will assess the data practices of open government data intermediaries from three perspectives. First, it will trace the development of open government data initiatives to argue that at a moment when open data policy is experiencing global diffusion with the potential of increasing social, political, and economic impact, there is a crucial need to assess the practices of intermediaries to understand how open government data is being put to use. Second, it will develop a framework to analyze the role of open government data intermediaries by proposing a definition for "the data intermediary function" constituted by a range of technical, civic, representational, and critical data practices. Third, it will assess the data practices of two open government data intermediaries, *596 Acres* and *Transparent Chennai*, who as urban actors facilitate the

---

[12] Barbara Ubaldi, " Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives," (working Paper on Public Governance, No. 22, Organisation for Economic Co-operation and Development, 2013): 27.

conversion of open government data into actionable information for communities to effect changes in the built environment. In describing and assessing the tools, practices, and methods developed by open data intermediaries, this thesis will explore the potential and limitations of data intermediaries, and offer recommendations that might inform future open government data initiatives that seek to mediate open government data to facilitate changes in the built environment.

Research Questions

The primary research questions that this thesis will address are:

1. What are the limitations of open government data and the potential of data intermediaries to address these limitations?

2. What are the practices through which open government data is converted into information that is useful for communities?

3. What characteristics of open government data make it useful to effect changes in the built environment?

4. What are the challenges faced by open government data intermediaries in relation to quality and sustained use of open government data?

To begin, I will propose a definition for "the data intermediary function" as constituted by a set of technical, civic, representational, and critical practices that will help examine the tools, methods, and practices developed by open government data intermediaries. Based on this definition, I will examine two open data intermediaries. The first case, *596 Acres,* is a New York based community land access advocacy group that

supports local communities in gaining access to vacant public land in their neighborhoods for community projects.[13] The second case, *Transparent Chennai*, is an organization based in Chennai, India, that brings to attention "citizen needs" by collecting, creating, and disseminating data on various civic issues.[14] I examine the tools, methods, and organizational documents created by these data intermediaries, and conduct semi-structured interviews with their founders in order to understand how data intermediaries address the limitations of government data to make it more useful for communities.

These cases were selected because both illustrate an approach to supporting community use of open government data to effect changes in the built environment. In particular, I focus on data practices that provide context to data, and highlight characteristics such as granularity of data, which make it suitable for monitoring and assessing changes in the built environment. In doing so, I also highlight the differences in people's literacies to use open government data in varied contexts.

Through the cases I attempt to consider how data intermediaries engage in the process of converting incomplete, obsolete, or otherwise flawed government datasets into information that is more accurate and community-focused. I argue that the two data intermediaries exhibit an understanding of the limitations and shortcomings of open government data and strive to address these through their data intermediary practices. But, since these organizations operate in very different open government data contexts they respond to challenges peculiar to their context. For instance, *596 Acres* is able to use and aggregate government data that is already available in a machine-readable format through

---

[13] "596 Acres: About," 596 Acres, accessed March 01, 2015, http://596acres.org/en/about/.
[14] "About," Transparent Chennai, accessed March 01, 2015, http://www.transparentchennai.com/about/.

the New York City online data portal, by contrast *Transparent Chennai* has to file *Right to Information* requests and often navigate cumbersome bureaucratic processes to get access to municipal datasets in the Indian context.

Through these focused case studies, rooted in the practices of open government data use, this thesis hopes to contribute to the strand of research that is beginning to emerge with particular attention to the ways in which open government data is finding application in different contexts and diverse settings.[15]

The research is organized into five chapters.

Chapter One describes the recent proliferation of open government data initiatives and frames the discussion on open government data from three perspectives. First, it defines open government data and traces its growth in the last decade that has led to the diffusion of open government data initiatives around the world. I argue that even though open government data initiatives are gaining traction across the world, we have a very limited understanding of how open government data operates in different contexts and diverse settings. Second, Chapter One explores some of the primary motivations and conditions behind open government data initiatives. I argue that these motivations have created a compelling argument for the social, economic, and political potential of open government data. Third, I describe the resources involved in making government data accessible in an open format and its subsequent use and reuse. Through this discussion I seek to emphasize that operating on government data is a resource intensive process, and hence the use of open

[15] "Research Project: Open Data Intermediaries," Open Data Research Network, accessed April 26, 2015, http://www.opendataresearch.org/project/2014/intermediaries.

government data is driven by a combination of actors, namely open government data intermediaries.

Chapter Two attempts to describe open government data intermediaries, including their practices and functions. This chapter is organized into four sections. First, it explores some prominent critiques of open government data—its political and social nature, and its availability and use. This chapter also considers the implications of making government data open without considering potential privileges and biases in datasets. I argue that we need to understand how existing data intermediary practices should critically engage with government data and develop ways to address these challenges and limitations. In Chapter Two, I also argue that intermediaries should not only consider the practices that serve a technical, representational, and civic function, but also take into account data practices that serve a critical function. A number of historical precedents demonstrate a tradition of analyzing, interrogating, and contextualizing government data. I argue that these examples serve as early histories of the data intermediary function and offer valuable precedents for data practices of contemporary data intermediaries. Finally, in Chapter Two I propose a definition for the data intermediary function constituted by a set of data practices. This definition of the data intermediary function serves as a framework to compare the data practices of *596 Acres* and *Transparent Chennai* in Chapter Five.

Chapter Three describes the data collection, correction, and dissemination practices employed by *596 Acres* in building *Living Lots NYC*. I argue that the data practices of *596 Acres* demonstrate an ability to critically engage with data through a combination of virtual

and real "grondtruthing" practices.[16] While the case study highlights how *Living Lots NYC*, the online land access tool developed by *596 Acres*, supports communities interested in intervening in their built environment, it also brings to attention the importance of community outreach methods to facilitate the use of the tool, highlighting the need for funding for not only creation of tools but support for their long term use.

Chapter Four describes the data practices of *Transparent Chennai*, an organization that aggregates and disseminates data about civic issues in Chennai, India, in collecting, augmenting and disseminating public sanitation data. In this chapter I argue that the data practices of *Transparent Chennai* demonstrate an ability to aggregate, analyze and authenticate government data as well as an ability to create data through participatory methods that encourage and support neighborhood level advocacy with the potential to have a constructive impact on the built environment. In my analysis, I also explore the implications of using different data practices for collection and dissemination of data and how the work of the data intermediary may affect who gets to participate in these practices and how. This analysis raises some important considerations concerning potential bias in data intermediary practices.

Chapter Five provides concluding remarks and discusses the scope for further work I examine the similarities in the data practices of the two intermediaries in an effort to highlight the practices that prove to be effective in diverse settings. I emphasize that even though they operate in very different contexts, some of the challenges that *596 Acres* and *Transparent Chennai* face are similar. To conclude, I discuss some of these challenges in

---

[16] For a description of "groundtruthing" see Chapter Three.

greater detail, and offer recommendations that might inform future initiatives that seek to

mediate open government data to facilitate changes in the built environment.

Open Government Data: Growing Global Influence and Impact

In the current technological milieu, our everyday interactions are increasingly mediated by information sensing devices that produce, analyze, and store data. While data generation and dissemination have traditionally been laborious and expensive processes, newer technologies that are inherently data-driven automate the creation and dissemination of unprecedented amounts of data. The *Huffington Post* reports that "the digital universe is expected to double every two years and will multiply 10-fold between now and 2020 – from 4.4 trillion gigabytes to 44 trillion gigabytes."[17] As a data deluge surrounds us, terms such as 'big data' become salient in the popular discourse with myriad claims about the transformative potential of datasets. These developments offer new opportunities and applications in diverse areas. Applications range from—but are not limited to–epidemiologists creating warning systems by capturing and analyzing data on spread of diseases, police predicting crimes by identifying patterns in historical criminal data, governments using demographic and environmental data to streamline emergency disaster response, and so on.[18] It is important to note however that these benefits are accompanied by the considerable cost to curate, manage, store, and analyze large quantities of

[17] Olivier Dumon, "Innovations in Science: Managing the 'Billions and Billions' of the Data Deluge," *The Huffington Post*, accessed November 27, 2014, http://www.huffingtonpost.com/olivier-dumon/innovations-in-science-ma_b_5960306.html.
[18] Elaine Grant, "The Promise of Big Data," *Harvard T.H. Chan School of Public Health*, accessed April 20, 2015, http://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/.

data.[19] The question then becomes whether the potential benefits outweigh the significant costs and resources that need to be invested to realize the value of data.[20]

With these developments in the background, this chapter describes the recent proliferation of open government data initiatives and frames the discussion on open government data from three perspectives. First, it defines open government data and traces the growth in the demand for open government data in the last decade that has led to the diffusion of open data initiatives around the world. I argue that even though open government data initiatives are gaining traction across the world, we have a very limited understanding of how open government data operates in different contexts and diverse settings. Second, this chapter explores some of the primary motivations and conditions behind open government data initiatives. I argue that these motivations have created a compelling argument for social, economic, and political potential of open government data. Third, I describe the resources involved in making government data accessible in an open format and its subsequent use and reuse. Through this discussion I seek to emphasize that operating on open government data is a resource intensive process, and hence the use of open government data is driven by a configuration of actors, namely open government data intermediaries. To develop a fine grained understanding of the use of open government data requires an understanding of the complex data practices that open government data intermediaries are currently engaged in.

---

[19] Eric Savitz, "The Big Cost Of Big Data," *Forbes,* accessed April 20, 2015, http://www.forbes.com/sites/ciocentral/2012/04/16/the-big-cost-of-big-data/.
[20] Tariq Malik, "TeradataVoice: Big Data In Government's Service To Citizens And State," *Forbes*, accessed April 20, 2015, http://www.forbes.com/sites/teradata/2015/01/30/big-data-in-governments-service-to-citizens-and-state/.

1.1 Defining Open Government Data

While novel technologies that are inherently data-driven, such as networked embedded electronics, automate data production, there has been a corresponding effort to digitize, link, and make existing datasets openly available. This practice of making previously obscure or privately held data, publicly available in a reusable and redistributable format is termed as open data.[21]

The term *open* has become a prevalent descriptor in various rights to access movements such as open source, open content, and open science. *Open* in each of the above contexts, while related, refers to specific ways in which content is made available for reuse, redistribution, and retention. In recent years, a number of organizations have developed sets of characteristics or guidelines that define open data. *OpenDefinition.org* defines open data and content as data that "can be freely used, modified, and shared, by anyone for any purpose," emphasizing its redistributable, remixable and shareable nature. Recognizing its sensibilities specific to the United States, *opengovdata.org* sets out eight principles that bring to attention the completeness and granularity of datasets.[22] These principles highlight the quality and the nature of the data, along with its reusable and redistributable nature.[23] While these definitions point to the redistributable nature of open data, *Sunlight Foundation* notes the importance of appropriate formats that facilitate technical reuse and redistribution, and metadata that contribute to

---

[21] Rob Kitchin, *The data revolution: Big data, Open data, Data Infrastructures and their Consequences* (London: Sage, 2014) 48.

[22] The eight principles that government data need to comply with to be considered open: Complete, primary, timely, accessible, machine Processable, non-Discriminatory, non-proprietary, license-free. One of the early posts on the 'Open Government blog' of the United States government refers to the eight principles delineated by the *Opengovdata.org* while referring to open government data."8 Principles of Open Government Data," Public Resource, accessed December 8, 2014, https://public.resource.org/8_principles.html.

[23] Ibid.

interoperability and public discovery.[24] Along with the above characteristics, it also reasons that data creation processes should be made public in order to provide valuable perspective on data creation and quality; and opportunities should be created for public feedback about quality, format, and selection of datasets.[25]

The broad idea behind open data is to make data freely available in an electronic format over the web. In the context of public data, the objective of open government data is to make publicly acquired data at the national, regional, and local level available in a format that is suitable for reuse and redistribution.[26] This notion of open government data is driven by the belief that government data should be a publicly owned good. So, open government data is public data that is:

1. **Freely accessible**: Proactively published and available for free.

2. **Reusable and Redistributable**: Published with minimal legal restrictions on the reuse and redistribution of data.

3. **Machine-readable**: Published in file formats that are suitable for sorting, extraction, and linking, and so on.[27]

Open government data can be seen as an extension of the "Right to Information (RTI)" legislation in different countries that advocates for the right to access government information. More than 70 countries have such legislations in place with some countries seeing more

---

[24]"Open Data Policy Guidelines," Sunlight Foundation, accessed December 8, 2014, http://sunlightfoundation.com/opendataguidelines/.
[25] Ibid.
[26] Michael Gurstein, "Open Data: Empowering the Empowered or Effective Data Use for Everyone?" *First Monday* 16, no. 2 (2011), accessed December 18, 2014, http://firstmonday.org/ojs/index.php/fm/article/view/3316.
[27] These characteristics are derived from the definition of Open Government Data (OGD) used in the Open Data Barometer Global Report - Second Edition. Tim Davies, "Background and Introduction," Open Data Borometer, accessed April 19, 2015, http://barometer.opendataresearch.org/report/summary/the_barometer.html.

satisfactory adoption than others. [28] In many countries these legislations are either not properly

executed or are not adequate. However, it is important to tease out the distinctions between open

data and RTI laws, also known as the "Freedom of Information laws (FOIA) " in some countries.

RTI states that citizens have the right to request access to full or partial records of government

agencies, and in many countries this right is enforceable in a court of law. Even so, while some

countries had historically made government data freely available to the public on request, others

(for example the United Kingdom) charged a fee for the release of government data.[29] However,

a critical distinction between open data and RTI laws is that whereas RTI laws facilitate access to

government information on request, open data makes that information openly available and

accessible by default.[30] Consider the multi-step process required to get access to government data

outlined by Jonathan Feldman, the Chief Information Officer of the City of Asheville, North

Carolina. In order to get access to data from a government department the following must take

place: 1) a request is submitted to the government department, 2) the request is forwarded to the

Public Information Office (PIO), 3) IT or the government department must find the requested

data, 4) the legal department of the city must review the data to check for legal issues, and 5) IT

must release the data to PIO, who ultimately releases it to the requester. [31] This description

demonstrates that getting access to government data through mechanisms such as RTI and FOIA

---

[28] David Banisar, "Freedom of information around the world: A Global Survey of Access to Government
Information Laws," Privacy International (2006): 1, accessed April 21, 2015.
http://freedominfo.org/documents/global_survey2006.pdf.
[29] Kitchin, 54.
[30] "Executive Order -- Making Open and Machine Readable the New Default for Government Information," The
White House, accessed December 7, 2014, http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-
making-open-and-machine-readable-new-default-government-.
[31] Jonathan Feldman, "Asheville's Open Data Journey: Pragmatics, Policy, and Participation," in *Beyond
Transparency: Open Data and the Future of Civic Innovation*, edited by Brett Goldstein. (San Francisco: Code for
America Press, 2013), 51-62.

is a multi-step process, involving multiple government departments and their time and resources. Indeed one of the perceived benefits of open government data is the reduced day-to-day transaction cost of answering public requests.

It is also important to note that while RTI and FOIA provides selected access to public records, open government data provides access to government data at much larger scale.[32] This scale changes the way the data is understood, and also enables applications that make connections across multiple datasets and databases. For instance, the city of Louisville in collaboration with a medical health provider (Propeller Health) was able to link the data of inhaler use with air quality data to assess the need for interventions.[33] As such applications of open government data continue to emerge there has been increase in the demand for open government data across the world. Indeed the number of open government data initiatives has increased from two in 2009 to over 300 in 2013.[34] The following section explores the conditions and discourses that drove the demand for open government data.

1.2 Growth in Demand for Open Government Data

The popular use of the term open data is most often linked to the launch of open government data websites such as the *data.gov* in the 2009 and *data.gov.uk* in 2010. However, demands for making public data freely available have been gaining prominence since the late

---

[32] Joshua Tauberer, "Open Government, Big Data, and Mediators," in *Open Government Data: The Book,* accessed April 24, 2015, https://opengovdata.io/2014/open-government-big-data-mediators/.
[33] Tanvi Misra, "How New Orleans, Louisville, and San Francisco Are Using Open Data to Solve Problems," CityLab, accessed April 24, 2015, http://www.citylab.com/cityfixer/2015/04/3-cities-using-open-data-in-creative-ways-to-solve-problems/391035/.
[34]"The GovLab Index: Open Data, (Updated and Expanded)," The Governance Lab, accessed April 25, 2015, http://thegovlab.org/govlab-index-open-data-updated/.

2000s.[35] In March 2006, *The Guardian's* technology section featured an article called "Give us

Back our Crown Jewels," arguing for free release of data collected by government agencies

using public funds.[36] Launching the campaign "Free our Data," the article reasoned that

restricting access to national government data stifles innovation and limits the economic value

generated by government data.[37] The article does not use the term open data to refer to freely

available and accessible public data. Scholars have written about *The Guardian's* "Free our

Data" campaign as one of the drivers for open data in the United Kingdom.[38]

In the United States, thirty open government advocates, including open source activists

such as Lawrence Lessig and Aaron Swartz, came together in Sebastapool, California in

December 2007 for a meeting hosted by Tim O'Reilly of O'Reilly Media and Carl Malamud of

Public.Resource.Org. The goal of the meeting was to "devise a list of principles for open

government… to publish these principles and perhaps even get candidates in the upcoming U.S.

elections to adopt them."[39] Drawing a parallel between the open source movement and open data

movement, Lawrence Lessig recommended at the end of the meeting that just as the open source

movement set out a definition of open source software, the open source meeting should outline a

definition and fundamental characteristics of open government data.[40]

[35] Kitchin, 49.
[36] Charles Arthur and Michael Cross, "Give Us Back Our Crown Jewels," *The Guardian,* accessed December 4, 2014, http://www.theguardian.com/technology/2006/mar/09/education.epublic.
[37] Ibid.
[38] Noor Huijboom, and Tijs Van den Broek, "Open data: An International Comparison of Strategies," *European journal of ePractice* 12, no. 1 (2011): 11.
[39] "Memorandum," Public Resource, accessed December 4, 2014, https://public.resource.org/open_government_meeting.html.
[40] Lawrence Lessig, "Larry Lessig on Open Government Data Principles," YouTube Video, 1:33, posted by David Orban, December 8, 2007, accessed December 4, 2014, https://www.youtube.com/watch?v=AmlzW980i5A.

In 2009 *data.gov*, the United States government's online portal for open datasets was launched followed by the launch of *data.gov.uk* in the United Kingdom in 2010. More recently, the Obama administration passed an executive order that sets open machine-readable data as the new default for government information.[41] Launched as a part of the open government initiative of the Obama administration, *data.gov* uses the term open data to refer to federal, state, and local datasets made available through the portal. Since the launch of *data.gov* in 2009, open government data initiatives, which make government data publicly available in a machine-readable format for reuse and redistribution, have proliferated around the world. International advocacy and research groups such as the Sunlight Foundation, Open Data Research Network, Open Data Institute, and the Open Knowledge Foundation have contributed to this development by setting out standards for open data, developing software tools for data dissemination, and enabling community outreach around open government data.

Tim Berners-lee, the inventor of the *World Wide Web* and the President of the *Open Data Institute* while commenting on these developments in a TED talk entitled *The Year Open Data Went Worldwide* (2010) suggests that "there is an open data movement afoot now in the world."[42] In his widely viewed TED talk he goes on to list some of the prominent government-led and community-led open data initiatives such as *data.gov*, *data.gov.uk*, and *Open Street Map* that are active across the world, and discusses the implications and potential use of this data for communities.

---

[41] "Executive Order -- Making Open and Machine Readable the New Default for Government Information," The White House, accessed December 7, 2014, http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-.
[42] Tim Berners-Lee, "The Year Open Data Went Worldwide," *TED* video, 5:33, February 2010, accessed December 7, 2014, http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide?language=en.

The first edition of Open Data Barometer, a report that evaluates the diffusion of open

government data policy across the world, finds that the concept of open government data has

spread across the globe, and that out of the 86 countries surveyed 55% have developed open

government data initiatives in the last five years.[43] It then goes on to develop a set of indicators

to assess the quality of open data initiatives, the kinds of data being made available, and the

emerging impacts of open data in different countries. The report demonstrates that the practice of

making government data open for commercial and non-commercial use is indeed an emerging

global phenomenon. However, the countries represented in the report exhibit a wide variation in

the implementation of open government data policy at the national and regional level,

highlighting that far from being a uniform phenomenon, open data practices are evolving in

relation to the particular political, social, and technical conditions in different countries. The

report divides the countries surveyed into four clusters (Figure 1):

1. **High-Capacity**: Countries such as the UK, US, Sweden, France, and New Zealand, where

    open government data has received significant political backing and multiple government

    departments have released data at the national and local level.

2. **Emerging and Advancing**: Countries such as Spain, Chile, Brazil, Peru, and India, where

    open government data initiatives are emerging and advancing, and which are exploring

    practices to open data that cater to their populations.

3. **Capacity Constrained**: Countries such as Indonesia, Turkey, Ghana, Rwanda, and

    Vietnam, which face challenges in setting up sustainable open government data initiatives

---

[43] Tim Davies, "Open Data Barometer: 2013 Global Report," World Wide Web Foundation and Open Data Institute, accessed March 13, 2015, http://www. opendataresearch. org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf.

due to constraints in government, private, and public sector capacity (technical as well as financial).

4. **One-sided Initiatives**: Countries such as Malaysia, Kazakhstan, Saudi Arabia, Bahrain, and Qatar, where the release of open government data is one-sided and is not matched by the capacity of private or public sector to engage with and use that data.[44]
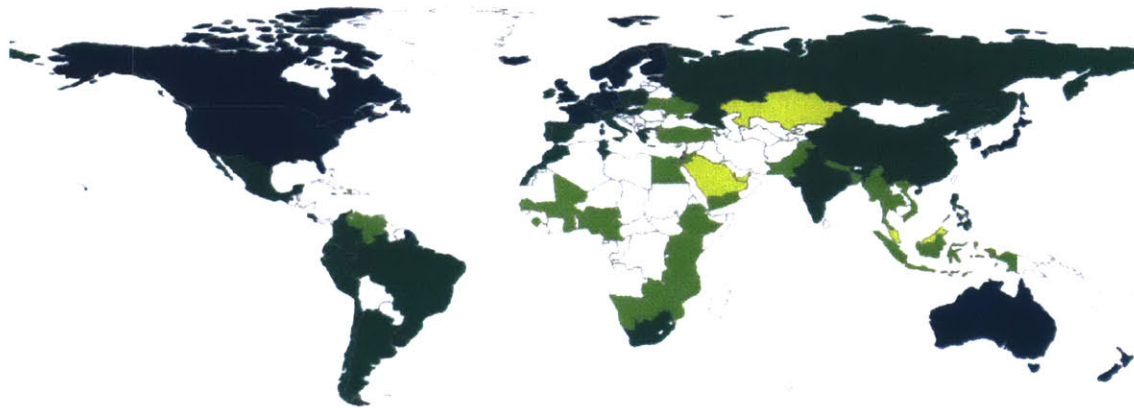
Country clusters based on Open Data Barometer Readiness and Impact questions
High capacity, Emerging & advancing, Capacity constrained, One sided initiatives

Figure 1: The countries in the Open Data Barometer report clustered as High Capacity, Emerging and Advancing, Capacity-Constrained, and One-Sided Initiatives. Source: Open Data Barometer

1.3 Primary Motivations Driving Open Government Data Initiatives

Open government data at a policy level is seeing an increasing diffusion around the world motivated by arguments that see open government data as a facilitator for increased transparency and accountability, greater civic participation, and significant economic and social innovation. Significant open government data initiatives are currently underway around the world that are

---

[44] "Data & Analysis: Clusters," Open Data Barometer, accessed April 18, 2015, http://www.opendatabarometer.org/report/analysis/index.html.

driven by demands of greater transparency and accountability of administrative processes,

accompanied with a discourse of data-centric citizen engagement and innovation. For instance,

in Iceland the investigators of the financial collapse of 2008 blamed oversight and lapses in

governmental decision-making for the breakdown of the country's economy, leading to

increasing demands for greater transparency through open access to public data.[45] Subsequently,

opening up government data came to be regarded as high priority by different branches of

government in Iceland.[46] In countries that have made government spending data public, websites

such as *Wheredoesmymoneygo.org* and *USAspending.gov* have sprouted that show how the

taxpayer's money is spent. In addition to facilitating public oversight, open government data also

propagates the vision of greater public participation in the government.

These idealized characteristics of open data are perceived to have implicit benefits and

impacts that are used by governments to make the case for open data. We may divide the

arguments being made for open data into five broad categories:

1. Open data make administrative processes open to public scrutiny by shedding light

onto the day-to-day operations of various government agencies.[47] Open sharing of public data

promotes a culture of transparency in administrative processes, and is considered to provide a

mechanism for oversight over the performance of government programs and activities. And "in a

well-functioning, democratic society citizens need to know what their government is doing,"

contends Open Knowledge Foundation.[48]

---

[45] Hijalmar Gislason, "Iceland: From the Financial Crisis to Open Data," *Open Knowledge Blog*, accessed April 24, 2015, http://blog.okfn.org/2010/03/26/iceland-from-the-financial-crisis-to-open-data/.
[46] Ibid.
[47] Ibid.
[48] "Open Knowledge: What Is Open?" Open Knowledge, accessed December 7, 2014, https://okfn.org/opendata/.

2. Open data is believed to provide opportunities for the public to engage in governance more directly.[49] Open data creates the occasion for continuous dialogue between citizens and governments. With access to public data citizens can become informed about, and contribute to the decision-making processes of government.[50]

3. Releasing public data makes administrative processes open for internal and external analyses.[51] Internally, organization-wide data can lead to insights into the functioning of different units within an organization, resulting in improved operational efficiencies and greater intra-organization collaboration. Externally, open data can be used to seek feedback and crowd-source solutions. For instance, on challenge.gov many government agencies pose problems and challenges to the public, and open data is considered as one of the enabling factors for such government-public interaction.[52]

4. Open data fosters brand value.[53] This is particularly apparent in the case of city level open government data initiatives. Having an open data portal, suggests that the city is committed to improving the quality of services available to its residents, and positions the city as innovative and forward-thinking.[54]

5. Open data is believed to release the social and economic value of previously obscure or hidden datasets.[55] In many areas, data created and owned by public organizations might be the

---

[49] Kitchin, 48-67.
[50] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk, "Benefits, Adoption Barriers and Myths of Open Data and Open Government," *Information Systems Management* 29, no. 4 (2012): 24.
[51] Kitchin, 48-67.
[52] Gwanhoo Lee and Young Hoon Kwak. "An Open Government Maturity Model for Social Media-based Public Engagement." *Government Information Quarterly* 29, no. 4 (2012): 499.
[53] Kitchin, 48-67.
[54] "How Chicago Is Growing Its Open Data Economy," Socrata, accessed December 9, 2014, http://www.socrata.com/case-study/chicago-growing-open-data-economy/.
[55] Kitchin, 48-67.

only source of data (geospatial, meteorological etc.), and policies that stifle the release of this data can have wide-ranging economic impacts.[56] As more and more organizations come to rely on data for their operation, it is believed that opening up government data facilitates creation of new data-centric services and applications, fostering innovation in the market. Supporting this reasoning, *data.gov* United States' open data portal claims, "open data is fuel for innovators. It has the potential to generate more than $3 trillion a year in additional value"[57]

Rob Kitchin contends that over the past few decades the arguments in support of open data have constituted an influential "discursive regime" that has been efficacious in convincing governments and organizations in opening up their data.[58]

While open government data sees a global diffusion, "we are still to understand how it plays out in national, sub-national, and local community contexts."[59] Researchers have argued that there is a need to study and assess the current use of open government data in different contexts.[60] The following section begins to do so by tracing the practices involved in making government data open and its subsequent use.

---

[56] Rufus Pollock, "The Economics of Public Sector Information" (working Paper in Economics, Faculty of Economics, University of Cambridge, United Kingdom, 2008), accessed January 12, 2015. http:rufuspollock.org/papers/economics_of_psi.pdf.

[57] "Impact," Data.gov, accessed December 8, 2014, http://www.data.gov/impact/.

[58] Kitchin, 48-67.

[59] Tim Davies and Zainab Ashraf Bawa,"The promises and perils of open Government data," *The Journal of Community Informatics* 8, no. 2 (2012), accessed February 12, 2015. http://ci-journal.net/index.php/ciej/article/view/929/926.

[60] In this poster, Morgan Currie proposes to address this gap by creating a typology of usage, and by developing case studies of open government data usage in the context of the United States. Morgan Currie, "Participation in the City: A Typology of Open Government Data Use" (poster presented at the iConference at University of North Texas, Fort Worth, Texas, February 12-15, 2013), Accessed March 12, 2015, http://hdl.handle.net/2142/42062.

1.4 Beyond Access: Open Government Data from Dataset to Use

Governments have long collected data, including demographic, environmental, civic, and crime data, for administrative and managerial purposes. In countries such as the United States where e-government policies have been implemented, these datasets are often collected, stored, and managed using resource planning software systems. Moreover, government departments have internal processes that backup and transfer data from these organizational systems to privately-held department-specific integrated data repositories (also referred to as data warehouses).[61] Brett Goldstein, the former Chief Data Officer of the City of Chicago, describes the process of releasing the crime data of the city through its open data portal:

> Crime data is recorded in the Chicago Police Department's transactional system and then replicated into their data warehouse. Our approach was to fire an ETL (a set of database functions for moving data from one place to another) from an internal utility server to pull data from the police warehouse and load it into the city's data portal via Socrata's API.[62]

This description suggests that even when data is already available in a digital format, in this case from the Chicago Police Department's reporting system, additional processing is involved to translate it into a format suitable for an open data portal. This processing entails translating the agency-specific data into an open format, and adhering to the standards, formats, and schemas of the open data portal. Datasets that have been converted to follow open standards, such as the Chicago crime data, are then aggregated and disseminated through national, regional, and city-level online data portals. *Data.gov*, the US government's open data portal, currently provides access to more than 130,000 datasets from over 50 government agencies categorized by different

---

[61] Brett Goldstein, "Open Data in Chicago: Game On," in *Beyond Transparency: Open Data and the Future of Civic Innovation*, edited by Brett Goldstein. (San Franciso: Code for America Press, 2013), 13-26.
[62] Ibid.,16.

topics such as agriculture, health, and local government.[63] However, it does not host any of the

datasets directly, but rather just collects metadata about datasets from multiple sources, and

makes it accessible and discoverable from one location.[64] Similarly, some open data portals act

primarily as aggregators and disseminators of datasets, with a search functionality that facilitates

access to datasets based on government department, location, and so on. Others provide an

additional ability to visualize, selectively export, and initiate discussions around particular

datasets. So, it would appear that *data.gov* is designed for a more technically advanced user as it

provides datasets to be downloaded as machine-readable files, but it does not facilitate other

data-centric user interactions (Figure 2). The Chicago city open data portal, in addition to

offering datasets for download as machine-readable files, begins to provide additional

functionality such as filtering, visualizing and selectively exporting data, which might prove

useful to a more casual user (Figure 3). For instance, the Chicago crime data is available for

download from *data.gov* in multiple file formats (Figure 2). Whereas the same dataset on the

Chicago city open data portal can be filtered to show the crimes that were reported on a

particular date (Figure 3).

---

[63] "Data.gov," Data.gov, accessed December 4, 2014, http://www.data.gov/.
[64] "Frequently Asked Questions (FAQs)," Data.gov, accessed April 26, 2015, http://www.data.gov/faq.

Figure 2: The Chicago crime data on data.gov, with options to download data in multiple file formats. Source: United States Open Data Portal

City of **Chicago** Data Portal

Home About Help Developers Terms of Use City of Chicago Sign Up Sign In

Crimes - 2001 to present

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each ▸

Manage More Views Filter Visualize Export Discuss Embed About

		Block		IUCR		Primary Type		Description		Location Description
10		097XX S ESCANABA AVE		031A		ROBBERY		ARMED: HANDGUN		RESIDENTIAL YARD (FRONT/BACK)
11		034XX W 62ND ST		0460		BATTERY		SIMPLE		STREET
12		028XX N CLARK ST		0860		THEFT		RETAIL THEFT		SMALL RETAIL STORE
13		035XX W POLK ST		1320		CRIMINAL DAMAGE		TO VEHICLE		STREET
14		0000X W HUBBARD ST		0890		THEFT		FROM BUILDING		TAVERN/LIQUOR STORE
15		082XX S HOUSTON AVE		4625		OTHER OFFENSE		PAROLE VIOLATION		STREET
16		052XX S MARSHFIELD AVE		143A		WEAPONS VIOLATION		UNLAWFUL POSS OF HANDGUN		STREET
17		114XX S DR MARTIN LUTHER		0486		BATTERY		DOMESTIC BATTERY SIMPLE		SIDEWALK
18		034XX W HURON ST		0470		PUBLIC PEACE VIOLATION		RECKLESS CONDUCT		SIDEWALK
19		082XX S HOUSTON AVE		143A		WEAPONS VIOLATION		UNLAWFUL POSS OF HANDGUN		SIDEWALK
20		062XX S LAFLIN ST		2024		NARCOTICS		POSS: HEROIN(WHITE)		ALLEY
21		065XX W BELMONT AVE		0860		THEFT		RETAIL THEFT		CONVENIENCE STORE
22		012XX W TAYLOR ST		0479		BATTERY		AGG: HANDS/FIST/FEET SERIOUS INJUR		BAR OR TAVERN
23		042XX W WILCOX ST		2230		LIQUOR LAW VIOLATION		ILLEGAL CONSUMPTION BY MINOR		SIDEWALK
24		021XX N MASON AVE		1310		CRIMINAL DAMAGE		TO PROPERTY		ALLEY
25		064XX S EBERHART AVE		1320		CRIMINAL DAMAGE		TO VEHICLE		RESIDENTIAL YARD (FRONT/BACK)
26		033XX N CLARK ST		0610		BURGLARY		FORCIBLE ENTRY		RESTAURANT

Totals

Figure 3: The Chicago crime data on the Chicago city open data portal, with options to filter, visualize, and selectively export data. Source: City of Chicago Data Portal

These user interactions, while useful, are still very rudimentary, and converting a dataset into a data visualization or an interface often necessitates complex data processing. In describing the process of developing a visualization of neighborhoods in Washington D.C. using metro ridership data, Joshua Tauberer, a civic technologist, outlines four steps: 1) getting the data 2) cleaning the data 3) transforming it into a visual form 4) disseminating the results.[65] While Tauberer describes in detail the complexities at each step, I briefly describe the process of cleaning the data to give a sense of the intricate data practices involved. Tauberer suggests that even when the government data is available as a machine-readable file, it is seldom in a form and structure that could be directly useful.[66] Government data has to typically go through a process of *normalization,* that gives it structure and a consistent format.[67] This might involve, for instance,

---

[65] Joshua Tauberer, "Visualizing Metro Ridership," in *Open Government Data: The Book*, accessed April 26, 2015, https://opengovdata.io/2014/visualizing-metro-ridership/.
[66] Ibid.
[67] Ibid.

ensuring that all dollar amounts are numbers by getting rid of $ signs in the dataset, and so on – a

process that is often a combination of "automated tricks" and "time consuming manual effort."[68]

Needless to say, both opening up government data and putting it to use is a resource

intensive process. Many national and city-level open government data initiatives are supported

by the government and often fall under the purview of the department of information technology.

In the United States, the *Electronic Government Fund (EGF)* provides funding for *data.gov* and

other transparency related websites such as *USAspending.gov*.[69] According to the Sunlight

Foundation, while planning for a new open government data initiative funding should be

considered for the hiring of new staff, acquiring new software, training of existing personnel, and

maintaining the data portal.[70] However, even when open government data initiatives are

government-supported, they are susceptible to budget cuts. For instance, during the budget

debate in 2011 in the US, the *EGF* was stashed from $34 million to $2 million.[71] More

importantly, as open government data continues to spread across the globe, the emphasis of

funding is on making government data open, and not on understanding the subsequent use of

data. Open data initiatives frame the user to be "everyone," and yet they rely strongly on

technically competent actors such as Joshua Tauberer, who act as important intermediaries in the

subsequent use and reuse of data.[72]

---

[68] Ibid.

[69] Daniel Schuman, "Budget Technopocalypse: Proposed Congressional Budgets Slash Funding for Data Transparency - Sunlight Foundation Blog," accessed April 27, 2015, http://sunlightfoundation.com/blog/2011/03/23/transparency-technopocalypse-proposed-congressional-budgets-slash-funding-for-data-transparency/.

[70] "Open Data Policies and Implementation: Frequently Asked Questions," Sunlight Foundation, accessed April 20, 2015, http://sunlightfoundation.com/policy/opendatafaq/

[71] Schuman, "Budget Technopocalypse."

[72] Fred Saunderson, "Investigating Public Participation in Open Government Data," Open Data Institute, accessed May 6, 2015, https://theodi.org/blog/open-participation-public-participation-and-open-government-data.

In his examination of use of open government data in the United Kingdom, Fred

Saunderson has found that the UK open data portal relies heavily on intermediaries, who turn

open government data into visualizations, interfaces, and other outputs, for data usage.[73]

Similarly, the Open Data Barometer report argues that successful open government data

initiatives need intermediaries to realize the social and economic potential of government

datasets.[74]

At this moment when open government data continues to evolve and spread across the

globe with the potential of social, political, and economic impact, I think that we have a great

opportunity to develop a sharper understanding of open government data use through a deeper of

exploration of the work of intermediaries in diverse open data contexts. In this section, I have

sought to highlight the complex data practices that go into the conversion of a dataset into a

format that is suitable for wider use. Through an initial example, I have begun to outline a few of

these data practices. However, we need a more robust framework to appreciate the complex

practices employed by open government data intermediaries. This I will set out to do in the

following chapter.

---

[73] Ibid.
[74] Davies, "Open Data Barometer Global Report - Second Edition," 6.

Open Government Data Intermediaries: Practices and Definition

In Chapter One, I have noted the emergence of open government data initiatives around the world and the primary motivations that have driven the global diffusion of these initiatives. Even with the growing salience of open government data, we do not have a nuanced understanding of how open government data operates in diverse settings and different contexts. We can gain a better understanding of how open government data is put to use in diverse settings by exploring the practices of an emerging configuration of actors—open government data intermediaries— that play a crucial role in facilitating the use of open government data through a range of data practices. These could be non-profits, for-profits, community-based organizations, developers, civic technologists, and designers, working in diverse settings to make open government data intelligent and accessible. To develop a nuanced understanding of the data practices of intermediaries, I propose that we develop a definition for the "data intermediary function," which is comprised of a set of technical, representational, civic, and critical data practices employed by these actors in mediating open government data.

This chapter is organized into four sections. In the first section "Limitations and Critiques of open government data," I will tease out the prominent critiques of open government data –its deeply political and social nature, and questions of access and use–and discuss the implications of making government data open without taking into consideration these implicit privileges and biases in datasets. I believe that we need a better understanding of existing data practices that critically engage with government data to recognize the ways in which these challenges and

limitations might be addressed. In the second section "Defining Open Data Intermediaries," I explore current discussions on the role of open government data intermediaries in the context of use of open government data. We need to consider the data practices of open government data intermediaries that not only serve a technical, representational, and civic function, but also take into account data practices that serve a critical function. In the third section "Early Social Studies as Histories of Data Intermediaries," I examine a few historical precedents that reveal a tradition of analyzing, interrogating, and contextualizing government data. I suggest that these social studies serve as early examples of the data intermediary function and help us understand the data practices of contemporary data intermediaries. Finally, in the last section, I propose a definition for the data intermediary function constituted by a set of data practices. This definition of the data intermediary function will serve as a framework in my comparison of the data practices of *596 Acres* and *Transparent Chennai* in Chapter Five.

2.1 Limitations and Critiques of Open Government Data

While there is a significant discussion about the positive implications of open government data, a critical discourse about the ways in which open data are made available and put to use is starting to emerge. Danah Boyd and Kate Crawford begin their paper "Six Provocations for Big Data" with a quote from Geoffrey Bowker that highlights the assumption, biases, and limitations inherent in data, calling for a closer analysis of "data as representation" rather than "data as objective fact."[75]

---

[75] The paper offers a critical perspective on the growing discourse around "big data," questioning the biases and assumptions inherent in datasets. Danah Boyd and Kate Crawford, "Six Provocations for Big Data" (paper presented at Oxford Internet Institute's conference, "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society." Oxford, United Kingdom, September 21, 2011): 1, accessed December 13, 2014, http://ssrn.com/abstract=1926431.

We need to open a discourse – where there is no effective discourse now – about the varying temporalities, spatialities and materialities that we might represent in our databases, with a view to designing for maximum flexibility and allowing as possible for an emergent polyphony and polychrony. Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care.[76]

This is the primary argument of *Raw Data is an Oxymoron*, in which Lisa Gitelman and her contributors brings to attention different practices of data manipulation (such as "data scrubbing") that are considered acceptable in academic disciplines.[77] Inherently, data production, aggregation, and dissemination as processes are fraught with subjective decisions. But, quantitative data produced by these processes are often misconstrued as self-evident facts, with a disregard to the multiple levels of manipulations and abstractions that they have gone through. In describing how an environmental phenomenon is translated into meaningful scientific data, David Ribes and Steven J. Johnson unravel a data collection process that is marked by human interference and mediation at every stage, concluding that existence of data depends on a "complex assemblage of people, instruments, and practices dedicated to their production, management, and care."[78] The actors involved in the production, management, and interpretation of data leave traces of their values, assumptions, and worldview in the data they produce.[79] Moreover, data analysis is also an act of interpretation.[80] Emphasizing the interpretive flexibility involved in analyzing what might seem straight-forward historical solar eclipse data consisting of time and place information, Matthew Stanley observes that the researchers working with the data have to exercise considerable subjective judgment because

---

[76] Ibid.
[77] Lisa Gitelman, ed. *Raw data is an oxymoron* (Cambridge: MIT Press, 2013), 70.
[78] David Ribes and Steven J. Jackson. "Data bite man: The work of sustaining a long–term study," in *Raw data is an oxymoron* (Cambridge: MIT Press, 2013): 147.
[79] Jeffrey Alan Johnson, "From Open Data to Information Justice," accessed October, 2014, *Ethics and Information Technology* 16, no. 4 (2014): 265.
[80] Boyd and Crawford, 5.

the data is a "result of complicated, messy processes that are far from standardized."[81]

While practitioners in academic disciplines such as sociology have been attentive to the

constructed nature of data they produce, even accounting for the limitations of their research

methods, the political and social nature of datasets is not critically questioned in most

discussions about open data.[82] For instance, while the Sunlight foundation offers an extensive

set of guidelines that addresses data quality and format, it does not in particular address the

values, assumptions, and biases that might be implicit in government data.[83] But, as Jeffrey

Alan Johnson argues, government data reflect existing social patterns, and in doing so reproduce

patterns of privilege and injustice prevalent in the society.[84] Through well-discussed cases, he

illustrates how data collection, management, and dissemination practices of the government,

result in datasets that over-represent and privilege the interests of certain groups over others. In

the case of the United States census, black and Hispanic households are prone to be

underrepresented in the census count at a higher rate than non-black households. Among other

causes, this irregularity is attributed to the disproportionate barriers in data interactions between

underprivileged groups and state.[85]

The manner in which an open data system is structurally implemented can affect its

outcomes. In the widely discussed case of digitization of land records in Karnataka India, a

government initiative was aimed to bring transparency into the management of land records, but

---

[81] Matthew Stanley, "Where Is That Moon, Anyway? The Problem of Interpreting Historical Solar Eclipse Observations," in *Raw data is an oxymoron* (Cambridge: MIT Press, 2013): 77.
[82] Boyd & Crawford, 5.
[83] "Open Data Policy Guidelines," Sunlight Foundation, accessed December 8, 2014, http://sunlightfoundation.com/opendataguidelines/.
[84] Johnson, 4.
[85] Ibid., 5.

resulted in marginalizing a certain community by underrepresenting its members in the database created. *Bhoomi*, an open data system, was created in such a way that it included only certain types of structured documents for digitization, excluding more informal and unstructured data sources.[86] The land claims of *Dalits,* an underprivileged and disadvantaged group in India, were not represented in the documents chosen for digitization, as they were supported by unstructured and informal data sources.[87] This resulted in the exclusion of land claims of *Dalits* from the open data system altogether. In reference to *Bhoomi*, Kevin Donovan contends that by privileging particular kinds of data the open data initiative facilitated in the process of empowering the empowered, and marginalizing the poor.[88]

Such cases illustrate that government data collection processes are far from being neutral or objective. Advertently or inadvertently, values, assumptions, privileges, and biases are implicit in government datasets; and when such datasets are released openly they disseminate and perhaps even amplify the privileges and biases embedded in them. While the philosophy behind open data might be considered politically benign, open data in itself is not politically neutral.

Moreover, the increasing salience of data in contemporary times raises questions about access and control over data, and the ways in which differential access to open datasets creates new hierarchies and divides. Implicit in many discussions of open data is the idea that, once made open, data is accessible to everyone. While open and equal access to data is the driving

---

[86] Ibid., 6.
[87] Ibid., 7.
[88] Kevin Donovan, "Seeing Like a Slum: Towards Open, Deliberative Development," *Georgetown Journal of International Affairs* 13, no. 1 (2012): 98-99.

force behind many open data initiatives, and although these initiatives facilitate new opportunities of interface between public and government, the general public is often unaware of the kinds of data made available, how the data was generated, and maybe put to use.

Moreover, while the goal of open data initiatives is to make datasets available in standardized and machine-readable formats such as CSV, JSON, and XML, analyzing and interpreting data made available in these formats requires access to appropriate skills and technology.[89] Data scientists, civic hackers, developers, and designers are able to participate in the open data movement as they not only possess the requisite tools to download and process open data, but also the relevant knowledge and skills. However, the interactions of a technology savvy group of people with open data, does not represent the interactions of the general population. So, despite the potential of open data, most people are able to engage with it only minimally. This is because even though the tools that facilitate the analysis of data have proliferated, the skills required to download, process, and analyze open data are still the domain of a privileged few.

Much like the digital divide, the constraints of access and use of data create a data divide, limiting the use of open data to those with skills, expertise, and resources to make sense of it. And while open data might be driven by the philosophy of equal access to everyone, the unevenness in access and use of data might result in exacerbating existing divides. Significant

---

[89] CSV, JSON, and XML are examples of popular machine-readable file formats. CSV (Comma Separated Values) is a file format used to save tabular data in plain-text. JSON (JavaScript Object Notation) is a format used to represent data structures. And XML (eXtensible Markup Language) is a format that defines a set of rules for encoding data. "A Primer on Machine Readability for Online Documents and Data," Data.gov, accessed May 9, 2015. https://www.data.gov/developers/blog/primer-machine-readability-online-documents-and-data.

discussion about access to open data has been framed in terms of the implications of unequal

data literacies in an increasing data-centric society. For instance, in the case of the *Bhoomi*

project in India, centralization, formalization, and digitization of land records removed the land

management process away from the village level. [90] This made the management process more

legible and controllable for large corporations and well-to-do farmers who had the capacity to

exploit and engage with the new system, and left already disempowered groups with little power

to influence the process.[91] Solly Benjamin and his colleagues found that one of the material

implications of this data divide was that large players were able to influence the land

management process to acquire large quantities of land.[92]

With the awareness of the implications of unevenness in use of open data, there has been

significant discussion regarding the need for policies that not only make data open, but also

support the development of capacities to analyze and interpret data. Framing the question of

access in terms of data literacies, Erica Deahl argues that in an increasing data-rich society there

is a critical need to build general public's ability to critically engage with and interrogate data,

and to participate in opportunities of "data dialogue" facilitated by greater availability and

openness of data.[93] Framing the question of access in terms of "effective use," Michael Gurstein

distinguishes between the opportunities offered by open data from the actual realization of these

---

[90] Solomon Benjamin, R. Bhuvaneswari, and P. Rajan, "Bhoomi:'E–governance', or, an anti–politics machine necessary to globalize Bangalore?" (working paper, Collaborative for the Advancement of Studies in Urbanism through Mixed Media, January 2007), 33, accessed December 12, 2014, https://casumm.files.wordpress.com/2008/09/bhoomi-e-governance.pdf.

[91] Ibid.
[92] Ibid., 3.
[93] Erica Sachiyo Deahl, "Better the Data You Know: Developing Youth Data Literacy in Schools and Informal Learning Environments," S.M. thesis, Massachusetts Institute of Technology, 2014, 33s.

opportunities.[94] He contends that efforts to make data open should be accompanied with

corresponding efforts that equip and enable the widest possible range of users to translate access

to open data into useful outcomes towards achieving collectively identified goals.[95]

According to Gurstein, one way to enable this translation process is to embed learning

and training resources into open data initiatives that facilitate interpretation, analysis,

visualization, and use of data. However, open data initiatives rarely offer access to widely

accessible learning and training resources that might facilitate the analysis and interpretation of

data. As a suggestion of usability and accessibility, *data.gov* showcases applications,

visualizations, and tools developed using the datasets made available on the website. But, apart

from the "Developers" section on the website, which provides technical information about the

development of *data.gov* and access to APIs, little attention is paid to making resources

available that might afford the general public to learn how to meaningfully use, interpret, and

analyze the datasets.[96]

Most open data initiatives act as delivery mechanisms for the dissemination of

government datasets, framing the citizen as a consumer of government datasets, "rather than

committing to build mechanisms for citizens to engage with the policy process… on the basis of

the data that is made available."[97] However, as Tim Davies has argued the impacts of open

government data initiatives are better realized not through the dissemination of datasets alone,

[94]Michael B. Gurstein, "Open data: Empowering the empowered or effective data use for everyone?" *First Monday* 16, no. 2 (2011), accessed December 18, 2014, http://firstmonday.org/ojs/index.php/fm/article/view/3316.
[95] Ibid.
[96] "Open Source," Data.gov, accessed December 15, 2014, http://www.data.gov/developers/open-source.
[97] Tim Davies, "Can the G8 Open Data Charter Deliver Real Transparency?" accessed December 15, 2014, https://theconversation.com/can-the-g8-open-data-charter-deliver-real-transparency-15434.

but through a "range of technical, social, and political resources" that are mobilized to support

activities around datasets.[98] With the proliferation of open government data, an emerging

configuration of actors—open data intermediaries—is playing a crucial role to support and

facilitate activity around government datasets. The practices of open data intermediaries shed

light on processes that affect the usefulness and reliability of open government data. Moreover,

to understand the ways in which the limitations of open government data are being addressed,

we need to understand existing data practices that critically engage with government data, and

facilitate community access and use of data.

2.2 Defining Open Data Intermediaries

The metaphor of an ecosystem has been used by scholars for open government data to

convey an idea of "interdependent social systems of actors, organizations, material

infrastructures, and symbolic resources that can be created in technology-enabled, information-

intensive social systems."[99] Magalhaes et al. argue that the eco-systemic view of open data

(rather than a citizen-centric view) accounts for the emergence of data intermediaries that serve

as a bridge between data providers and users.[100] And as research on open government data has

evolved, researchers have sought to outline the function of data intermediaries and the myriad

ways in which they work with government data through a description of the range of data

---

[98]Tim Davies, "Open Data: Infrastructures and ecosystems," Open Data Research, (2011): 1, accessed March 18, 2015, http://ssrn.com/abstract=2549734.
[99] Gustavo Magalhaes, Catarina Roseira, and Sharon Strover, "Open Government Data Intermediaries: A Terminology Framework," In *Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance, Association for Computing Machinery*, (2013): 331.
[100] Ibid.

practices that they are engaged in. These definitions differ with regard to the roles they highlight

(technical, civic, social), and the data practices they include or exclude.

Many researchers have used the metaphor of "bridging the gap," while describing the

technical function of data intermediaries to highlight the gap between the availability and

technical skill required to use data. One of the early uses of the term intermediaries in the context

of government data comes from Craig and Sawicki, who use it in the context of organizations

that facilitate access to urban data for community groups.[101] They frame the function of

intermediaries as "information providers," performing a range of functions, ranging from

providing data in an easy to understand format to providing policy advice based on information

extracted from data.[102] However, Rufus Pollock, co-founder of the Open Knowledge Foundation,

argues against this conceptualization of the open data ecosystem as a "one way street," with a

one-directional flow of open government data from the data provider to the data intermediaries

and finally to the end user.[103] Instead, he offers an aspirational model for the open data

ecosystem in which intermediaries not only process open government data, but also release

processed data, and users are not just passive consumers of data but are able to contribute and

correct data.[104]

---

[101] David S. Sawicki and William J. Craig, "The Democratization of Data: Bridging the Gap for Community Groups," *Journal of the American Planning Association*62, no. 4 (1996): 516.
[102] Ibid.
[103] Rufus Pollock, "Building the (Open) Data Ecosystem" Open Knowledge Blog, March 31, 2011, accessed April 22, 2015, http://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem/.
[104] Ibid.

This notion of data intermediaries goes beyond the technical function to emphasize mediation as a process that creates "loops of data" within the open data ecosystem.[105] Embracing this notion of open government data, Sumandro Chattapadhyay suggests that "different scales of granularity and expanse (spatial and/ temporal)," in data supports its diverse uses and the situational and context-specific use of data necessitates a diverse range of actors, including data intermediaries.[106]

Researchers and practitioners working in the context of developing countries highlight a range of practices of employed by data intermediaries, including but not limited to, rectifying and correcting government data, shaping demand for open data, and facilitating communities in creating data, expanding the role of intermediaries to account for their civic and critical function.[107] Open data intermediaries:

1. **Facilitate use and reuse of data:** Intermediaries often go beyond their technical function to support the conversion of data into useful information.

2. **Generate Data:** Intermediaries address gaps in government data by helping communities in generating complementary or corrective data.

3. **Validate and correct government data:** Intermediaries validate government data to reflect the conditions on the ground.

---

[105] Sumandro Chattapadhyay, "Opening Government Data through Mediation: Exploring the Roles, Practices and Strategies of Data Intermediary Organisations in India," Open Data Research, 2014, accessed March 18, 2015. http://ssrn.com/abstract=2549734.
[106] Ibid.
[107] "Ground-up Open Data Intermediaries – Who? Where? How?" (session, Open Knowledge Festival, Berlin, Germany, July 17, 2014), accessed April 22, 2015, https://pad.okfn.org/p/Ground-up_open_data_intermediaries.

4. **Shape demand for government data:** Intermediaries might reveal gaps in government

data and thus shape demand for government data through their work.[108]

In this discussion of data intermediaries, it is also important to consider the work of

researchers who discuss the processes through which open government data is put to use, but

might not necessarily use the term intermediaries. For instance, Tim Davies outlines five distinct

processes through which open government data is put to use in his in-depth analysis of

data.gov.uk:[109]

1. **Data to fact:** Extracting or seeking out facts from open government data.[110]

2. **Data to information:** Representing and interpreting open government data through data

   visualizations, infographics, and reports.[111]

3. **Data to interface:** Creating new ways to interact with and explore government datasets.[112]

4. **Data to data:** Disseminating and sharing data derived from government data.[113]

5. **Data to service:** Using open government data to create services and applications[114]

It is important to note that multiple processes could be involved in a particular "pattern of use" of

open government data.[115] For instance, in order to create an interface ("data to interface") a data

intermediary might also need to visualize and map the data ("data to information"). Moreover,

these processes serve as a useful framework to assess and tease out the data practices of

---

[108] Ibid.
[109] Tim Davies, "Open Data, Democracy and Public Sector Reform: A Look At Open Government Data Use From data. gov. uk" M.Sc Dissertation, Oxford University (2010), 3.
[110] Ibid.
[111] Ibid.
[112] Ibid.
[113] Ibid.
[114] Ibid.
[115] Ibid.

contemporary data intermediaries. For instance, the Sunlight Foundation offers a range of APIs that facilitates access to government data, but with the added ability to filter and sift through government datasets. The Capitol Words API created by the foundation allows user to access the frequency of words and phrases spoken in the capitol. This data is made available by converting congressional records obtained as plain–text from the Government Printing Office into an XML, and subsequently processing the data to split the text into words and phrases.[116] In the process Capitol Words API serves the data intermediary function by allowing access to a converted and processed dataset ("data to data"), and providing a filter functionality to sift through the dataset ("data to data").

In another case, "Adopt-a-Hydrant," developed by Code-for-America for the city of Boston, allows citizens to adopt a hydrant, and bear the responsibility for shoveling it during heavy snow. It uses data on urban infrastructure (location of fire hydrants in the city) and maps it to present fire hydrants in the neighborhood of the user, marking the fire hydrants that have been adopted as green and those that haven't been adopted as red. In this case, "Adopt a Hydrant" serves the intermediary function by mapping and contextualizing urban data ("data to information"), and providing an interface to explore and act on that data ("data to interface" and "data to service").

By focusing on the processes and practices through which open government data is being used, we already begin to develop a fine-grained understanding of the complex data practices of open data intermediaries. Drawing on these discussions on data intermediaries and my broader research, I will propose a definition of the data intermediary function, constituted by a range of

---

[116] "Capitol Words," Capitol Words, accessed April 23, 2015, http://capitolwords.org/?terma=war&termb=peace.

data practices. But, before doing so, I explore a few historical examples that serve as precedents

to contemporary intermediaries. This discussion will inform the definition of the data

intermediary function in the subsequent section.

2.3 Early Social Studies as Histories of Data Intermediaries

While the term data intermediaries, has found use in the context of open government data

only recently, the notion of data intermediaries in itself is not new. There is a rich history of

supplementing, interrogating, and representing government datasets that long predates the open

data movement. Between 1886 and 1893, Charles Booth, an English philanthropist, businessman,

and social researcher, funded a research team to document and map the urban social condition in

London.[117] Unconvinced by a report published by Henry Hyndman in 1885 that claimed that

25% of Londoners lived in extreme poverty, and displeased with the lack of empirical data about

the conditions of poverty in London, Booth undertook an extensive inquiry into the conditions of

the working class population in London.[118] For the study, Booth and his team of researchers

combined extensive ethnographic and survey data gleaned from interviews and observational

notes, and census and other official data obtained from different government departments. The

findings of the survey were published as three editions of *Life and Labour of the People in

London,* with the final edition amounting to seventeen volumes.[119] It is worth noting that having

assisted in the allocation of the Lord Mayor of London's Relief Fund by assessing census returns

in 1884, Booth was well aware of the shortcomings of government census data.[120] By accounting

---

[117] "Charles Booth (1840-1916) - A Biography," Charles Booth Online Archive, accessed December 13, 2014,
http://booth.lse.ac.uk/static/a/2.html.
[118] Ibid.
[119] Ibid.
[120] Ibid.

for, explaining, and reconciling discrepancies in official data in *Life and Labour of the People in London*, he provides a unique insight into his role as an interpreter, critique, and intermediary of government collected data.

Throughout *Life and Labour of the People in London*, Booth notes the use of official data collected from sources such as Census, the School Board Visitors, and the School Board of Guardians.[121] In relating the findings from the study, on multiple occasions Booth compares figures obtained from census data with data gathered by his fellow researchers and himself, often finding the census data to be bare, meager, and outdated."[122] In the section entitled 'Influx of Population,' he notes that while the census data puts the population of foreign-born residents and foreigners in London at 2.75 % and 1.5 % of the population respectively, it is probable that the census data understates the actual number of foreigners and foreign-born residents. He goes on to explain that foreigners, particularly those living in poorer quarters, might find it difficult to fill in the census document, and hence might be underrepresented in the census count.[123]

Booth's concerns about the quality and exactitude of official data obtained from different government departments demonstrate a fine-grained understanding of the data collection practices of the government; and through his work Booth was able to advocate for the need for social surveys to address the gaps and shortcomings in government datasets. Booth's survey of London is considered a pioneering effort in using empirical data to analyze social inequality, and was referenced by other sociological studies undertaken in the late 19th century. Prominent among these are *Hull-House Papers and Maps (1895)* that present the observations of

---

[121] Charles Booth, *Life and Labour of the People of London*, (London: Williams and Norgate, 1889), 178.
[122] At one point Booth notes, "the need for a more frequent census is badly felt," Ibid., 176.
[123] Ibid., 519.

Jane Addams and her colleagues from a predominantly impoverished district in Chicago and *The*

Philadelphia Negro (1899), a study by W.E.B Du Bois that offers an insight into the living

conditions of African Americans in Philadelphia's seventh ward.[124]

Between 1892 and 1920, Hull-House, a social settlement co-founded by Jane Addams in

1889, served as a center for various social inquiries into the conditions of living in Chicago

neighborhoods.[125] In 1893, Florence Kelley, a resident of Hull-House, was chosen to lead the

Chicago arm of the nationwide survey, *A Special Investigation of the Slums of Great Cities*,

undertaken by the United States Department of Labor to understand urban poverty.[126] Florence

Kelley and her fellow researchers conducted an extensive survey of the district surrounding Hull-

House. The data collected for the study was a combination of ethnographic data and data

gathered from administering family and tenement survey schedules.[127] Before the survey data

was submitted to the Commissioner of Labor in Washington, the Hull-House residents made a

copy of data related to birth, nationality, and employment history of individuals.[128] Subsequently,

Kelley and her colleagues transferred this information onto a color-coded map of the district,

showing individual houses and streets, with the color of the house indicating the nationality of

the residents in one case and the wages of the resident family in another.[129] While the data

collected from the study was published as an official government report, the maps and other

observations from the study were also published as *Hull-House Papers and Maps* in 1895. In the

---

[124] Shannon O'Connor, "Methodological Triangulation and the Social Studies of Charles Booth, Jane Addams, and WEB Du Bois," *Sociation Today* 7, no. 1 (2009).
[125] "CSISS Classics - Florence Kelley: Slums of the Great Cities Survey Maps, 1893," accessed December 15, 2014, http://www.csiss.org/classics/content/35.
[126] Ibid.
[127] Jane Addams, Hull House Maps and Papers (Boston: Thomas Crowell & Co.,1895), 7.
[128] Ibid.
[129] Ibid., 8.

"Map Notes and Comments" section of *Hull-House Papers and Maps,* Agnes Sinclair Holbrook

notes:

> The United States Department of Labor states the exact figures as part of the report
> on *The Slum Investigation,* and all statistics related to the subject are already
> published. But the partial presentation offered here is in more graphic and minute
> form; and the view of each home and lot in the charts, suggesting just how members
> of various nationalities are grouped and disposed…may have its real as well its
> picturesque value.[130]

Holbrook's observation acknowledges the report published by the government, and yet expresses

a parallel and considered intent to publish data in a format different from the official publication.

This intent to represent already published government data in a form that might offer a more

nuanced and illustrative understanding of data, resonates with contemporary practices of data

intermediaries, who transform open government data into data visualization based applications in

an effort to make it more widely accessible and intelligible. Early accounts illuminate historical

data practices that interrogate, scrutinize, and supplement government data. Although, the actors

involved in these accounts are not historically framed as data intermediaries, by interpreting and

representing government data for the general public they served an intermediary function in the

data ecosystem at the time. Moreover, their concerns about the challenges of using government

data resonate with contemporary concerns about the limitations of open government data. These

historical examples underscore the critical data practices of intermediaries, which in the

following section I will argue is an important component of the data intermediary function.

2.4 The Data Intermediary Function

Accounts of data practices of intermediaries in contemporary and historical context allow

---

[130] Ibid., 9.

us to develop a working definition for the data intermediary function. By focusing on a range of technical, representational, civic, and critical data practices of data intermediaries we can develop a useful definition of data intermediaries that can help in evaluating their roles in different contexts. I divide the data practices supported by intermediaries into six broad categories:

1. **Practices of Aggregation:** Data intermediaries collect, combine, convert, link and filter data (often from multiple sources and in multiple formats) to glean information, and in doing so exhibit an understanding of database structures and functions.

2. **Practices of Rectification:** Data intermediaries address the inaccuracy, incompleteness, and obsolescence of government datasets by validating, updating, and correcting (often iteratively) government datasets.

3. **Practices of Interpretation:** Data intermediaries augment the intelligibility of government datasets through critical analysis, and in doing so attempt to reveal the implicit biases and assumptions in data.

4. **Practices of Representation:** Data intermediaries visualize and contextualize open government data through a range of representation methods, including maps, infographics, and graphs.

5. **Practices of Dissemination:** Data intermediaries release converted, validated, integrated data in a reusable way, and in doing so contribute to the flows of data in the open government data ecosystem.

6. **Practices of Augmentation:** Data intermediaries address gaps and limitations in open government data by augmenting and annotating data, and creating data (often in collaboration with communities).

7. **Practices of Connection:** Data intermediaries support data-based interactions between individuals and governments.

These practices can be further clustered as technical, representational, civic, and critical data practices. For instance, downloading and linking multiple datasets from an open data portal is a practice of aggregation and serves a technical function. Alternately, validating a dataset against other datasets is a practice of rectification and serves both a technical and a critical function.

This matric or rubric of data can provide a robust framework for assessing the function of data intermediaries. Evaluating the role of intermediaries using this approach can help reveal similarities in data practices of intermediaries, as well as bring to attention disparities in the patterns of use of open government data in diverse settings. For instance, while the practice of aggregation in one context might be a matter of downloading datasets from an open data portal, in another context it might involve establishing contact with city officials for the release of crucial datasets. So, even though both practices involve the collection of data, in the first case the practice mainly serves a technical function, whereas in the second case it largely serves a civic function.

Moreover, this definition differs from the earlier definitions in its emphasis on the critical function that data intermediaries serve. As discussed in the previous section, data collection practices of the government can be inherently political, resulting in datasets embedded with social values, privileges, and biases. Government datasets might implicitly empower certain communities, while disenfranchising others. By "practices of augmentation" that support communities underrepresented in government datasets, data intermediaries supplement open

government data. In doing so, intermediaries can potentially inform policy based on newly gathered data, and help spark discussion about the limitations of data collection practices of the government, and serve a critical function.

Through these initial discussions I have sought to develop a framework for assessing and comparing the data practices of intermediaries. In the following chapters, I explore the applicability of this framework in greater detail by using it to explore the data practices of two intermediaries – *596 Acres* and *Transparent Chennai.*

596 Acres: A Public Land Access Intermediary

In the previous chapter I proposed a definition for the data intermediary function to assess

and evaluate the data practices of two data intermediaries. In this chapter, I will apply the

definition to *596 Acres,* a land access and advocacy organization in New York City established

in 2011, which builds tools that combine multiple data sources, including open government data,

to facilitate community land access. Through online tools and community outreach it makes

information about public vacant land openly available, and makes it easier for people living in a

neighborhood to cluster and organize around particular lots of land. Its team constitutes

attorneys, advocates, urbanists, designers, and developers. Its first online interactive tool was

created in 2011, and mapped and represented pieces of public vacant land in Brooklyn. After

operating the New York City land access online tool as a pilot project from August 2011 to

March 2015, recently *596 Acres* created *Living Lots NYC,* a New York City community land

access tool, which incorporates the best practices identified during the project. The funds for

building *Living Lots NYC* came from the Sunlight Foundation, whereas *596 Acres* derives its

funding from multiple sources including, but not limited to, fund for city of New York,

Awesome foundation's New York City chapter, two crowdfunding campaigns on IOBY, and

individual donors. At the same time, *596 Acres* has developed online interactive tools to support

community land access in Philadelphia, Los Angeles, and New Orleans in collaboration with

other organizations. The data it creates in this process is freely downloadable. Moreover, it has

also made its source code openly available, and supports organizations interested in creating similar online land access tools for other cities through "hands-on mentorships."[131]

3.1 Background

According to founder Paula Segal, *596 Acres* began as a response to the questions that emerged during the Brooklyn Food Conference (2011), which highlighted the need to reform the food system in cities by posing the questions; "Where is the land? Where can we grow?"[132] The name *596 Acres* was inspired by the fact that in 2011, according to MapPLUTO, a dataset released by *The Department of City Planning of New York*, 596 Acres was the total area of vacant public land in Brooklyn.[133] However, Segal observes that the focus of *596 Acres* soon shifted from finding out the total area of vacant public land that could be released for potential community use to understanding how people could be connected to this information locally.[134] So, the question that *596 Acres* sought to answer was, "Where are the [vacant public] lots? Who lives near them? And who wants to change them?"[135] In order to answer these questions, initially, *596 Acres* created posters (Figure 4) of the Brooklyn borough area with all the public vacant lots in the neighborhood represented as green spots. This combined series of data practices translated the data on public vacant land available in spreadsheets from the NYC open data portal into context-specific information, intended to highlight the abundance and the spread

[131] "596 Acres: Maps For Other Cities!" accessed March 31, 2015, http://596acres.org/resources/other-cities/.
[132] "BrianLehrer.tv: This Land Is Your Land, under an Oligarchy" YouTube Video, 56:16, posted by cunytv15, April 23, 2014, accessed March 1, 2015, https://www.youtube.com/watch?v=Vv4m6vNFKmQ.
[133] Ibid.
[134] Ibid.
[135] Ibid.

of public vacant land available in the borough.[136] Printed versions of this poster were posted

outside some public vacant lots, inviting and encouraging interested neighbors to get in touch

with *596 Acres* to understand how to begin organizing around the public vacant lot and to

reclaim it for community use.[137] While these graphic posters spurred responses from neighbors

of certain lots, an online interactive tool that represented data about public vacant land in

Brooklyn and subsequently included data on other boroughs of New York City, soon

accompanied them (Figure 5).[138] According to the "About Us" section on the *596 Acres* website,

these tools "turn city data into information about particular pieces of land and connect people to

one another through simple social networking functions."[139]

---

[136] Mariana Mogilevich, Ben Campkin and Rebecca Ross, "'There's Land If You Want It': How a Hand-Drawn Map Is Transforming Vacant Lots in Brooklyn," *The Guardian*, accessed March 03, 2015, http://www.theguardian.com/cities/2014/dec/12/vacant-lots-brooklyn-596-acres-new-york.
[137] "News for July 2011," 596 Acres, accessed March 03, 2015, http://596acres.org/en/news/2011/07/.
[138] Ibid.
[139] "About Us," 596 Acres, accessed March 24, 2015, http://596acres.org/en/about/about-596-acres/.

Figure 4: "There's Land if You Want It," a poster designed by Julia Samuels for *596 Acres* that represents public vacant lots as green spots on the map of Brooklyn. Source: 596 Acres

Figure 5: The online interactive tool that represents public vacant lots (and some private vacant sites whose owners have made them available for public use) in different boroughs in the New York City. Source: 596 Acres
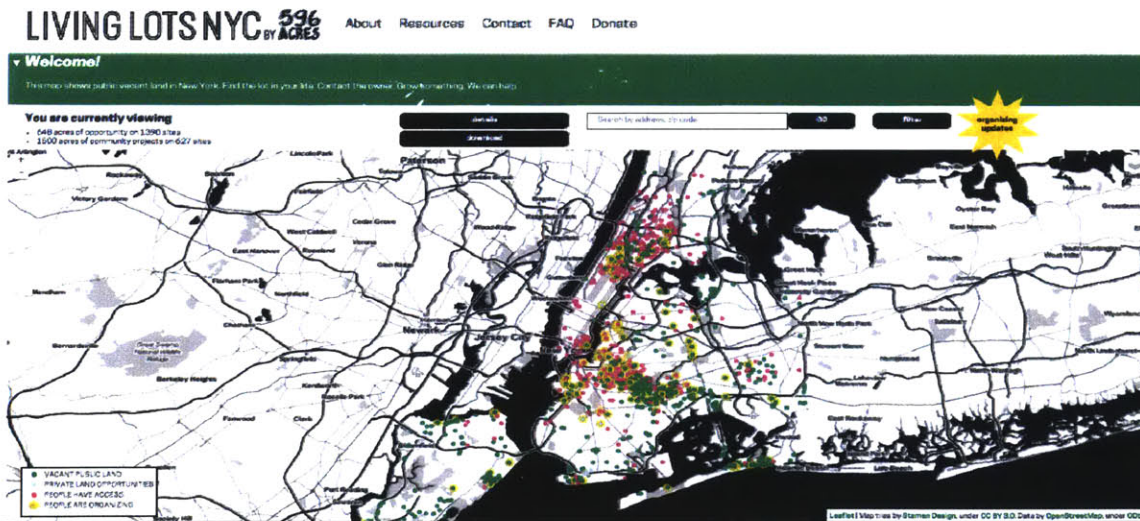
Figure 6: *Living Lots NYC*, the most recent iteration of the online interactive tool developed by *596 Acres*.[140] The green dots represent public vacant land and the light blue dots represent private vacant land. The pink dots represent land where people already have access and dots with a yellow outline represent lots where people are organizing. Source: 596 Acres

---

[140] "Home," Living Lots NYC, accessed March 17, 2015, http://livinglotsnyc.org/#11/40.7301/-73.9895.

3.2 Overview of Tools

The early versions of the online interactive tool developed by *596 Acres* showcased both public and private vacant sites that were available for community use, with subsequent versions highlighting sites on which community groups had established access with the support of *596 Acres* (Figure 5). In between August 2011 and March 2015, *596 Acres* facilitated the conversion of 32 vacant lots in New York City into community spaces.[141] For the purpose of this case study, I will focus on the most recent iteration of the New York community land access online tool, *Living Lots NYC,* and the ways in which it presents data to facilitate communities in organizing around particular pieces of public land (Figure 6). It is important to note, however, that in developing land access tools for cities, *596 Acres* follows a multi-phase data collection process that might differ based on the particular databases available in different cities.[142] Commenting on the need to vary the data collection practices for different cities Segal observes:[143]

> If we take a look at the four tools that we've built the data behind them is different, because it is responsive to the data environment that different cities have. But, really what we are doing is just looking for information wherever we can find it. If it doesn't exist as open data, we'll buy it. If it doesn't exist as data at all, we'll go read the documents and create our own spreadsheets.

In the subsequent section, I focus on the data practices employed in the development of *Living Lots NYC* to illustrate the different functions that the tool serves with regard to transforming open data to make it usable for communities. The account draws from interviews with *596 Acres* founder Paula Z. Segal, analysis of use of the interactive online tool *Living Lots NYC* developed

---

[141] "About," Living Lots NYC, accessed March 17, 2015, http://livinglotsnyc.org/about/about/.
[142] "About Us," 596 Acres, accessed March 24, 2015, http://596acres.org/en/about/about-596-acres/. 596 Acres has developed land access tools for Philadelphia, New Orleans, and Los Angeles in partnership with other organizations.
[143] Paula Z. Segal (Executive and Legal Director of the NYC land access program, *596 Acres*) in discussion with the author, January 2015.

by *596 Acres*, and documentation of publicity materials and newsletters on the *596 Acres* website.[144]

3.3 *596 Acres* and the Data Intermediary Function

To collect data about public vacant land in New York City *596 Acres* followed a multi-phase data collection and correction process, using data made available through the city-specific open government data portal as a starting point (Figure 7). In the process of building its online tools the data collected from city agencies was cross validated with other datasets and subsequently corrected and continually updated. In this section, I will provide an overview of the range of data practices through which *596 Acres* performs the intermediary function with respect to open government data and in the process transforms "city data that can be outdated and incomplete to a living database that is more accurate and community-driven."[145]

3.3.1 Practices of Aggregation: Collecting Government from Multiple Sources

The data collection process for the creation of *Living Lots NYC* began with a process of aggregating data from multiple datasets made available through the New York City open data portal. The New York City open data portal makes available 12000+ citywide datasets from different city agencies in machine readable, reusable, and redistributable format.[146] However, according to Segal, *Housing Preservation & Development (HPD)*, the New York City agency that owns the inventory of public vacant land in New York City, hasn't made that dataset public.

---

[144] Ibid.

[145] "Living Lots NYC Data," Living Lots NYC, accessed March 17, 2015, http://livinglotsnyc.org/about/living-lots-nyc-data/.

[146] "NYC Open Data," NYC Open Data, accessed March 30, 2015, https://nycopendata.socrata.com/.

So, the *596 Acres* team had to rely on data published about *HPD* properties by other city agencies.

To collect data on city-owned lots in New York City, the *596 Acres* team used data from two datasets on the NYC online data portal: *Local Law 48* of 2011 and the *Integrated Property Information System (IPIS)*.[147] The *Local Law 48* and *IPIS* are datasets of city owned and leased properties in New York City published by the *Department of Citywide Administrative Services*.[148] Both datasets have columns that detail the type and the current use of a city owned properties. For instance, in the IPIS dataset the column "RPAD_DESCRIPTION" describes the type of the property and the column "Primary Use Text" describes its primary use (Figure 7).

	BLOCK	LOT	RPAD_DESCRIPTION	Primary Use Text
1		74	20 VACANT LAND	NO USE
2		241	22 VACANT LAND	NO USE
3		241	22 VACANT LAND	NO USE
4		427	14 VACANT LAND	NO USE
5		1,615	59 VACANT LAND	NO USE
6		2,356	72 VACANT LAND	NO USE
7		3,834	54 VACANT LAND	NO USE
8		334	1 VACANT LAND	NO USE
9		334	2 VACANT LAND	NO USE
10		334	4 VACANT LAND	NO USE

Figure 7: A screenshot of the IPIS dataset from the NYC open data portal filtered to show public vacant land currently not in use. Source: NYC Open Data Portal

By combining these two datasets the *596 Acres* team created a list of public vacant lots that were not being used. However, this initial data collected from the city agencies had some inconsistencies. As Segal notes, one of discrepancies was that the city data classified lots that were being used as community gardens as unused land. While this inconsistency might seem

---

[147] Ibid.
[148] "Local Law 48 Of 2011 Report," NYC Open Data, accessed March 30, 2015, https://data.cityofnewyork.us/City-Government/Local-Law-48-Of-2011-Report/2b6x-2bw6.

innocuous at the onset, such inconsistencies have implications for the built environment. This

inconsistency perhaps provides an insight into the perceived value of such community spaces for

city agencies. For instance, recently a request for qualifications by the *Housing Preservation*

and Development (HPD) for developing affordable housing included 17 such community

gardens.[149]

While aggregating data, *596 Acres* marked lots that were being used as community

gardens by using a New York City community gardens database. The *596 Acres* team was able

to remove community gardens from the initial list of public vacant lots by comparing the data

with the 2010 GrowNYC community garden survey, which provided a list of city-owned

properties that were used as community gardens.[150] Thus, through a process of combining and

sourcing multiple datasets the *596 Acres* was able to create a list of public vacant lots in New

York City.

3.3.2 Practices of Rectification: Correcting Irregularities in Government Data

After the initial phase of data aggregation from multiple datasets, the next phase involves

rectifying and correcting the data collected in the initial phase–a process that *596 Acres* refers to

as "groundtruthing." During the "virtual groundtruthing" phase, the *596 Acres* team hired a NYC

property data expert, who used a combination of virtual mapping resources such as Google

Streetview, Satellite imagery, and OasisNYC, a GIS-based online tool that provides information

on every block and lot in New York, to virtually verify the data collected in the initial phase of

---

[149] "News: Gardens on Housing Preservation and Development's (HPD) List of Sites Available for Housing Development," 596 Acres, accessed May 7, 2015, http://596acres.org/news/2015/01/21/gardens-on-housing-preservation-and-developments-hpd-list-of-sites-available-for-housing-development/.
[150] "Living Lots NYC Data," Living Lots NYC, accessed March 17, 2015, http://livinglotsnyc.org/about/living-lots-nyc-data/.

the project.[151] In this phase, the lots, which were found to be in use, were removed from the

database and lots, which were found to be gutterspaces or potentially unusable by communities

because of inaccessibility or size, were appropriately marked in the database and moved onto a

different layer of the map.[152] Thus, by validating the original data not only for accuracy but also

for potential use, the *596 Acres* team was able to remove the lots that were not fit for community

use.[153] This method of validating data proved to be very efficient, and the *596 Acres* team was

able to remove about half of the lots in Brooklyn using this method.[154]

Further, the data was manually compared to the data on recent deed-transfers to remove

the lots that were no longer city-owned, thus making the dataset more up-to-date. Subsequently,

the data was published as an online map on the *596 Acres* website, to initiate the final phase of

"real groundtruthing."[155] In this phase, the data collected thus far was compared to the present

day use of the lot by the *596 Acres* team and the neighbors of a particular lot. To initiate

conversations around particular lots, the *596 Acres* team posted signs, encouraging neighbors to

write back seeking more information about a lot.[156] Moreover, the *596 Acres* team sought photos

and notes about the lots that were published as green dots on the online map from the

communities that lived around them.[157] Each green dot that represents a public vacant lot on the

map when clicked, opens up as a new webpage on the *Living Lots NYC* website, where people

interested in a particular lot can list themselves as *Organizers*, and upload images and post notes

---

[151] Ibid.
[152] Ibid. Gutterspaces are small pieces of land that are result inconsistencies in surveying and zoning errors.
[153] Ibid.
[154] Ibid.
[155] Ibid.
[156] Ibid.
[157] "News: Ground-Truthing!" 596 Acres, accessed March 23, 2015,
http://596acres.org/en/news/2011/12/28/ground-truthing/.

about the current status of the vacant lot (Figure 8).[158] What results is a continual process in which the data posted from organizers of a particular lot is used to update the *Living Lots NYC* database.

On the *Living Lots NYC* community land access tool, there are examples of organizers posting notes about the changing status of a particular vacant lot. For instance, in the case of a vacant lot in Queens an organizer posted in May 2012, "half of the block is occupied by an industrial facility, the other half seems vacant and heavily overgrown."[159] Months later, in October, 2013, another organizer posted about the same lot, "I biked over there last weekend and the entire lot was full of parked cars and trucks. Is it still worth calling [the city agency] or does this mean that it is an error in the data and is not actually a vacant site?" Thus the web page of a lot comes to reflect more closely its condition on the ground. Moreover, at any point, discrepancies and mistakes in existing data can be reported. The "Make a Correction" form under the "Have We Made a Mistake?" section on the web page of a lot can be filled in to update its actual use (Figure 9). So, for instance, once a group of organizers gets access to a public vacant lot for community use and indicates so through the form available on the webpage, the *Living Lots NYC* database is updated to reflect the actual use of the lot as either 'community green space' or 'community garden'.

---

[158] The *Living Lots NYC* online tool lists people who are organizing around a particular lot as *Organizers*, and people who are following the activity on a particular lot as *Watchers*.
[159] "Queens Block 777, Lot 1," Living Lots NYC, accessed March 23, 2015, http://livinglotsnyc.org/lot/4007770001/.

Figure 8: The *Living Lots NYC* online tool, with the *Queens block 777, lot 1* selected. Source: Living Lots NYC

Figure 9: A screen shot of the webpage of the *Queens block 777, Lot 1*, with options to organize and contribute information about the lot. Source: Living Lots NYC

3.3.3 Practices of Augmentation: Supplementing Government Data

Through the comments and images posted by the organizers of a particular lot, the data

about some lots achieves a level granularity and detail that did not exist in the initial data

aggregated from multiple open government datasets. For instance, in the case of the *Queens*

block 1174, lot 134, in response to a query from a member of the *596 Acres* team concerning the

present-day use of the block, an organizer noted "the current use is for vehicle accessway for

neighbors BUT they have the option to use the other side."[160] Subsequently, the organizer posted

a set of images of the lot from different vantage points, in order to illustrate that the lot, which

was currently being used as an accessway could potentially be used as a community green space

(Figure 10).[161]

---

[160] "Queens Block 1174, Lot 134 | Living Lots NYC," accessed March 24, 2015,
http://livinglotsnyc.org/lot/4011740134/.
[161] Ibid.

Figure 10: An image posted by an organizer on the web page of the *Queens block 1174, lot 134* with the caption: "the current use is for vehicle accessway for neighbors BUT they have the option to use the other side." Source: Living Lots NYC

In another example a member of the *596 Acres* team posted a note in the comments section

of the *Siempre Verde Community Garden*, detailing the history of the lot:

> In the mid-1980s, a woman named Cecllia lived at 179 Stanton Street, in a older
> tenement back building. Like many older Latino residents in the neighborhood, she
> was concerned with having three small and desolate empty lots next to her building.
> Over 25 years ago, she started Siempre Verde, a garden which used the two HPD lots
> and had obtained permission to use the privately owned Gottleib lot in between them
> as well. Siempre Verde was listed on 1980s and 1990s garden maps of the
> neighborhood. In the 1990s, other residents of her building planted the two large trees
> that shade these lots today. [162]

These examples illustrate that for community organizers *Living Lots NYC* facilitates not only

engagement and interaction with data derived from government data but also supports the

---

[162] "Siempre Verde," Living Lots NYC, accessed March 23, 2015, http://livinglotsnyc.org/lot/58238/.

contribution of additional data that adds granularity and detail to the data available for a particular lot.[163]

The process of "real groundtruthing" not only results in an additional layer of information for existing public vacant lots but in some instances has also resulted in the addition of new lots to the *Living Lots NYC* database.[164] For instance, an organizer for the *Queens block 777, lot 1*, posted about another vacant lot in the vicinity:

> There is also a piece of (big) land by Costco. I have been looking at for a year now it's amazing. I was thinking because it's so big it would be really cool if it can be turned into a sustainable ecosystem...[165]

But, while in the case of this particular lot, it was subsequently discovered that the land was a part of the La Guardia extension plan and could not be used as a community space, there have been instances of lots being added to the *Living Lots NYC* database through this process.[166] Thus, through a multi-phase data collection and correction process approach the *Living Lots NYC*, comes to reflect not only a more accurate but also a more granular and detailed picture of public vacant land in New York.

3.3.4 Practices of Connection: Mediating Interactions Between Communities and City Agencies

By making each data point clickable and allowing organizers to post comments on the webpages that are sent as updates to other organizers, the *Living Lots NYC* tool facilitates the

---

[163] For a more detailed discussion on granularity of data see Chapter Five.
[164] "Living Lots NYC Data," Living Lots NYC, accessed March 24, 2015, http://livinglotsnyc.org/about/living-lots-nyc-data/.
[165] "Queens Block 777, Lot 1," Living Lots NYC, accessed March 23, 2015, http://livinglotsnyc.org/lot/4007770001/.
[166] "Living Lots NYC Data," Living Lots NYC, accessed March 24, 2015, http://livinglotsnyc.org/about/living-lots-nyc-data/.

formation of communities of interest around pieces of land. There are multiple examples of the

comments section of the online tool facilitating in-person meetings and collaboration between

organizers of a particular lot. The comments section serves as an online space to announce

meetings, volunteer help, distribute tasks, seek collaboration, report developments, and exchange

encouragement. When a comment is posted on the web page of a particular lot all the organizers

and watchers of that lot are automatically notified, keeping everyone interested in a particular lot

updated about its status. While in some cases the online tool has led to the creation of

communities of interest around lots that didn't exist before, in others it has resulted in facilitating

communication between already existing communities.

Moreover, there are instances of the comments section being used to create a shared

vision about the potential use of a public vacant lot. In the case of the *Vinegar Hill Community

Garden*, a member of the *596 Acres* team shared a potential use for the vacant lot from a

neighbor of the lot:

> This lot has been sitting here for some time now...I would like to know if I
> could open this lot for myself and the other providers in this community to
> give the children a safe and learning space outdoor. I was thinking of a
> garden w[h]ere the children can learn about science. There is not much in
> this area for children and I would like to help keep this place organized and
> clean.

In such cases, by facilitating the sharing of potential uses of a particular lots the online tool

becomes a forum for assessing community outlook and forming consensus.

Apart from enabling communication amongst organizers around a particular piece of

land, *Living Lots NYC* also supports communication between organizers, and city agencies and

city officials. The webpage of a lot lists the names of the city council representative, provides

links to all the elected officials for that lot, and suggests other useful contacts. As Segal notes,

"we want to be able to say, here is a point at which you can intervene in a built environment," and providing links to relevant officials facilitates this intervention.[167] On the *Living Lots NYC* website there are multiple examples of organizers trying to figure out the relevant city official to contact and relating their experiences of interacting with a city official about a particular lot.

Through continued interaction on the online tool, communities receive support from each other and also from the members of the *596 Acres* team. For instance, in the case of *Queen Block 1999, Lot 38*, an organizer remarked, "currently in contact with the *Department of Citywide Administrative Services (DCAS)* to get permission for access into this lot."[168] Over the course of the discussion with *(DCAS)*, which is documented in the comments section, the organizers of the lot and the *596 Acres* team found that the lot was in fact a part of a street mapped by the *New York City Department of Transportation (CDOT)*, and subsequently the members of the *596 Acres* team contacted the legal department of *CDOT* to confirm the license permission for the particular piece of land. This example illustrates that finding the appropriate city department to contact in order to get access to a piece of land could be a drawn out process, and could prove to be an initial hurdle for community action.

These examples highlight what Segal observes during the interview about making data contextual so that people could act on it. She says, "we take the data, we take a lot of time translating it into information that is contextual. That it is about this lot or that lot, or this neighborhood or that neighborhood, this program or that agency, and how it impacts the particular property. " *Living Lots NYC* is rich in context-specific data that enables interaction

---

[167] Paula Z. Segal (Executive and Legal Director of the NYC land access program, *596 Acres*) in discussion with the author, January 2015.
[168] "Queens Block 1999, Lot 38," Living Lots NYC, accessed March 24, 2015, http://livinglotsnyc.org/lot/4019990038/.

between people in a neighborhood through the formation communities of interest, and at the time same facilitates their communication with city officials and agencies.

3.4 Conclusions

The data collection, correction, and dissemination practices employed by *596 Acres* in building *Living Lots NYC* demonstrates the ability of its approach in facilitating and supporting communities interested in intervening in their built environment. Through an online tool that combines multiple data inputs, including open government data and inputs from the community, resulting in a dataset that is continually updated by the community, *596 Acres'* approach illustrates how a data intermediary can enable the creation of a dataset that is community-driven.

The data on public vacant land is not made available from the *Housing Preservation and Development*, which in fact maintains this inventory for the New York City, but through its practices of aggregation *596 Acres* is able to create a list of public vacant lots by combining datasets from other city agencies made available on the NYC open data portal. Its practices of rectification combine both online and offline methods. By using virtual mapping resources it is able to eliminate vacant land that is not appropriate for community use, and by involving communities in the rectification process it is able to create a dataset that reflects the conditions on the ground more closely than the government data.

596 Acres' practices of connection illustrate that as a data intermediary it supports community interaction with open government data by making it available in a format that is context-specific and facilitates the formation of communities of interest that seek to organize around particular pieces of land by making each data point on the map actionable.

While the case study brings to attention the characteristics of the roles data might play in driving these changes, it also underscores the importance of supporting communities in the use of data. As Segal notes, "the bulk of the work is actually facilitating what can be done with the data once we have it."[169] Apart from the online tool that supports dissemination of data, *596 Acres* organizes workshops to support communities in the use of its data and online tool. Indeed, one of the challenges that *596 Acres* faces in its current approach is to find financial support for facilitating the use of the tools it develops.[170]

In Chapter Five I will discuss in greater detail the challenges and opportunities of designing for sustained use of data.

---

[169] Paula Z. Segal (Executive and Legal Director of the NYC land access program, *596 Acres*) in discussion with the author, January 2015.
[170] Ibid.

Transparent Chennai: An Urban Civic Data Intermediary

After exploring the data practices of *596 Acres* in the previous chapter, I now focus on a

data intermediary that works in a very different open data context. *Transparent Chennai* is an

organization based in Chennai, in the state of Tamil Nadu, India, that brings to attention "citizen

needs" by collecting, creating, and disseminating data on various civic issues. Housed in the

Center of Development Finance in the *Institute of Management & Research* in Chennai, its work

thus far has focused on, but not been limited to, the areas of urban governance, walkability,

pedestrian infrastructure, sanitation, electoral accountability, and cultural heritage in the city of

Chennai.[171] The organization derived its initial funding from Google's Inform and Empower

initiative and the ICICI Foundation for Inclusive Growth.[172] Its research team constitutes

researchers from varied academic backgrounds including, but not limited to, urban planning,

development studies, social sciences, law and economics. For its mapping initiatives,

Transparent Chennai has often worked with interns and citizen volunteers supervised by

researchers from *Transparent Chennai*. From time to time, it sends out open volunteer calls to

seek participation from citizens through its website and through announcements in local

newspapers, and involves them in its documentation and surveying efforts.[173]

---

[171] "About," Transparent Chennai, accessed March 16, 2015, http://www.transparentchennai.com/about/.
[172] "Only 600 Public Toilets in City," *The Times of India*, accessed December 23, 2014,
http://timesofindia.indiatimes.com/city/chennai/Only-600-public-toilets-in-city/articleshow/6686444.cms
[173]"Calling Volunteers for Walkability Survey!" Transparent Chennai, accessed May 7, 2015,
http://www.transparentchennai.com/calling-volunteers-for-walkability-survey/.

According to *Transparent Chennai's* website, its work "aims to empower residents by providing them useful, easy-to-understand information that can better highlight citizen needs, shed light on government performance, and improve their lives in the city" with the goal of enabling residents to have a "greater voice" in the city-wide planning and monitoring process.[174] In the context of Chennai, the collection of this information often involves working with a complex bureaucratic system to get access to government data, as well as collaborating with communities and other local organizations to create data that addresses the limitations of government data. Administratively, the city of Chennai is divided into 15 zones consisting of 200 wards, and is managed by *Corporation of Chennai (CoC)* one of the oldest municipal bodies in India.[175] It is led by a Mayor, who heads a council of 200 members, with each ward represented by a ward councilor.[176] In a few instances *Transparent Chennai* has worked closely with the *CoC,* leveraging the relationships developed through its continued work in the city to create data about issues of public interest in Chennai. Recently, the collaboration between *Transparent Chennai* and the *CoC* has deepened further, and they have signed a Memorandum of Understanding (MOU) to launch the *Chennai Data Portal,* the first city-level open data portal in India.[177]

As a data intermediary, *Transparent Chennai's* work brings to attention the complexities of creating and accessing data in an environment where the open data policies are still in the process of being implemented. In this chapter, I will focus on *Transparent Chennai's* efforts to

---

[174]"About," Transparent Chennai, accessed March 16, 2015, http://www.transparentchennai.com/about/.
[175] "Welcome to Corporation of Chennai," Corporation of Chennai, accessed May 3, 2015, http://www.chennaicorporation.gov.in/about-chennai-corporation/index.htm.
[176] Ibid.
[177] "Chennai Data Portal," Challenge Post, accessed March 27, 2015, http://opportunityhack2014chennai.challengepost.com/forum_topics/3731-chennai-data-portal.

advocate for more public toilets in Chennai through collection and mapping of public sanitation

data, a research project it initiated in 2009. Through its data collection efforts, it revealed that

Chennai had only 714 public toilets and not 1500, which the city had assured its residents.[178]

Moreover, through spatial mapping of public toilet data it also found that public toilets were

notably lacking in informal settlements and slums, areas where people often do not have access

to private sanitation and where public sanitation amenities perhaps are even more crucial.[179]

Transparent Chennai's research on public sanitation in Chennai received widespread

attention in the media, with notable news publications such as *The Hindu* and *The Times of India*

citing the research to report on the state of public sanitation in Chennai.[180] Through continued

advocacy, supported by public sanitation data that highlighted the need for more public toilets in

the city, *Transparent Chennai* helped "give credibility to citizen demands" for more public

toilets in Chennai (Figure 11). In 2013, the *Corporation of Chennai* indicated that it would install

5000 prefabricated toilets in the city.[181]

---

[178] "Public Toilets and Sanitation," Transparent Chennai, accessed April 1, 2015,
http://www.transparentchennai.com/public-toilets-and-sanitation/.
[179] Ibid.
[180] "Transparent Chennai's Public Toilets Research in the News," Transparent Chennai, accessed April 13, 2015,
http://www.transparentchennai.com/transparent-chennais-public-toilets-research-in-the-news/.
[181] "Transparent Chennai's Research Gives Credibility to Citizens' Demands for Public Toilets in Chennai!"
Transparent Chennai, accessed April 14, 2015, http://www.transparentchennai.com/transparent-chennais-research-
gives-credibility-to-citizens-demands-for-public-toilets-in-chennai/.

Figure 11: A public toilet in Chennai, India. Source: Shreya Krishnan, Volunteer, Transparent Chennai

In this chapter I will evaluate the data intermediary practices employed by *Transparent Chennai* during its research on public sanitation in Chennai. In particular, I will explore the characteristics of government data that it make salient for being used by communities to monitor and effect changes in the built environment. I will also describe how *Transparent Chennai* helped shape demand for open government data by working closely with the *Corporation of Chennai*. The data practices of *Transparent Chennai* demonstrate an ability to analyze and authenticate government data as well as an ability to create data through participatory methods that encourage and support neighborhood level advocacy with a potential to have an impact on

79

the built environment. In my analysis, I also explore the implications of using different data

practices for collection and dissemination of data and how the work of the data intermediary may

affect who gets to participate in these practices and how. This analysis raises some important

considerations concerning potential bias in data intermediary practices.

My account draws on an interview with *Transparent Chennai* founder Nithya V. Raman,

analysis of research papers, reports, and online interactive tools, and documentation of

experiences of researchers on the *Transparent Chennai* blog.[182]

4.1 Background

Transparent Chennai, according to founder Nithya V. Raman, emerged as a response to

the changing urban planning environment in the late 2000s, particularly in India.[183] With the

launch of *Jawaharlal Nehru National Urban Renewal Mission (JNNURM)*, a massive urban

renewal scheme initiated by the government of India in 2005, city governments were

increasingly using the language of data and mapping for better planning.[184] Earlier experience of

working with land rights organizations in India led Raman to feel that the organizations she was

working with needed to be able to use the language of data and mapping to make their claims to

the government. *Transparent Chennai* was founded in 2009 to aggregate, create, and disseminate

data about important urban issues facing the city of Chennai.[185]

Along with requesting and aggregating data from the city through informal and formal

channels (*Right to Information* requests), *Transparent Chennai* also works with local

---

[182] Ibid.

[183] Nithya V. Raman (Director, *Transparent Chennai*) in discussion with the author, March 2015.

[184] Named after the first Prime Minister of India Jawaharlal Nehru, the Jawaharlal Nehru National Urban Renewal Mission (JNNURM) was launched in 2005 to develop Indian cities through systemic investment in their socio-economic infrastructure.

[185] "About," Transparent Chennai, accessed March 16, 2015, http://www.transparentchennai.com/about/.

organizations and communities to produce data using qualitative data collection methods such as survey interviews and paper-based mapping exercises. Through these datasets *Transparent Chennai* has primarily focused on, but not been limited to, mapping civic infrastructure in the city, such as public toilets, bus routes, police stations, and garbage disposals.[186] In addition, it makes available non-sensitive data, collected and created in this process, through an online interactive tool and the "Download" section on its website. Along with disseminating data through its website, *Transparent Chennai* also conducts public events and workshops, and publishes reports in local languages, to disseminate data in communities that might not have the capacity and skills to use online tools, but are deeply affected by the urban issues in question.

4.2 Overview of Tools and Data

Transparent Chennai makes non-sensitive data collected through its efforts available through online tools and downloads. The online interactive map on the *Transparent Chennai's* website represents data collected on municipal infrastructure in Chennai. One can select from multiple layers of data, including transportation, infrastructure, housing and demographic data, to build a customized map. Some of this data is available to be filtered for more detail; for instance, the category of "Infrastructure by Ward" can be filtered to show public toilets and footpath length by ward (Figure 12). This layer provides a ward-wise overview of public toilets in the city, with the darker colors representing wards with higher number of public toilets and vice versa. For instance, white represents the wards that have zero toilets, whereas the darkest color represents wards that have toilets in the range 6-16. This map is a static representation of aggregated public toilet data and besides representing data does not support further user

interactions. A separate "Public Toilet" layer represents the GPS locations of the 49 public toilets surveyed in zone 4 as data points on the map (Figure 13).[187] This map offers some interactivity in that each data point, when clicked, opens up to show the individual profile for each public toilet with additional data. This data includes attributes such as cleanliness, presence of caretaker, and time of operation.
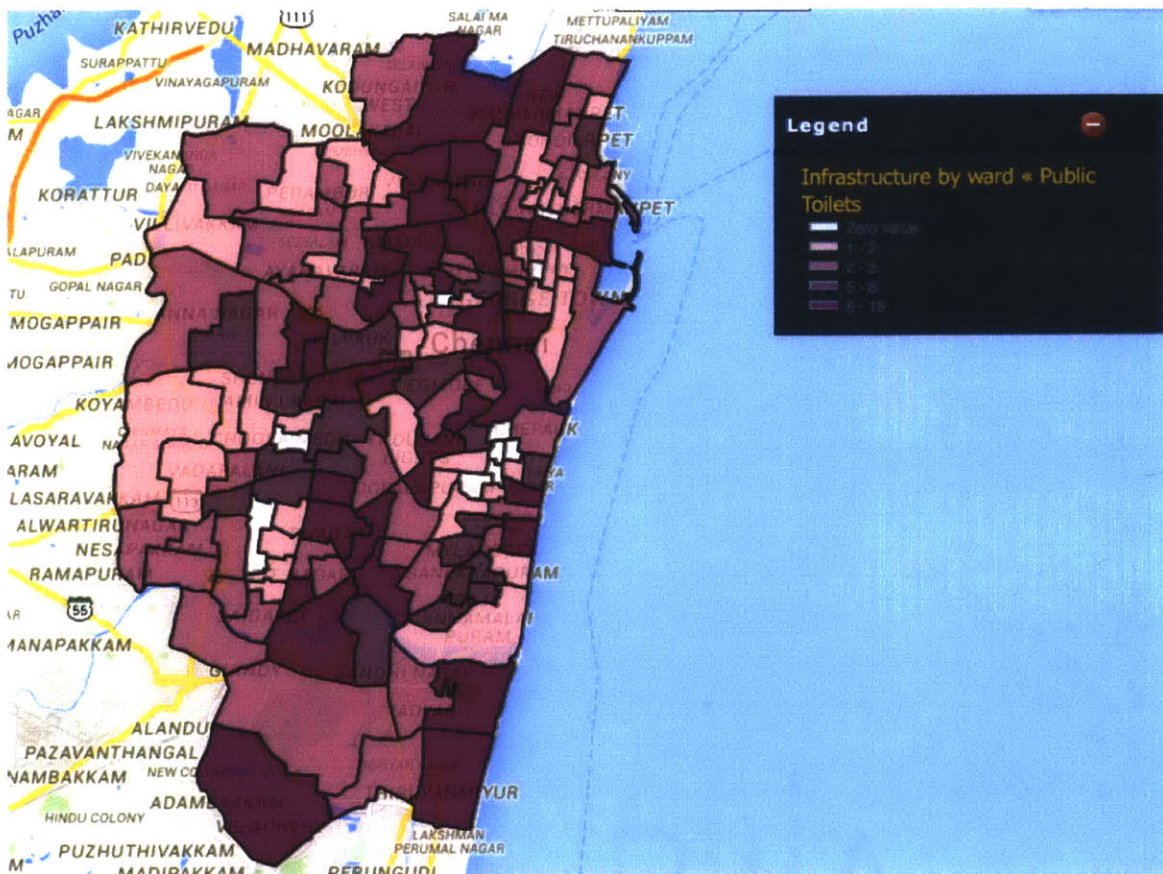
Figure 12: The online map on the Transparent Chennai website, filtered to present the aggregated public toilets data by ward. Source: Transparent Chennai

---

[187] Out of the 10 zones for which it had initial government data, Transparent Chennai selected zone 4 for a more detailed survey, and created an individual profile for the 49 toilets in this zone.

Figure 13: The online map on the Transparent Chennai website, filtered to present the public toilets in zone 4 as individual data points. Source: Transparent Chennai

It is important to note that these two different representational strategies, a ward-wise overview and a toilet-specific representation, inform the user-interaction with data very differently. The map that offers a ward-wise overview of public toilets represents aggregated data, and while useful to compare and contrast the public sanitation conditions in different wards, is not very useful in evaluating the condition of specific toilets (Figure 12). Whereas, the toilet specific representation presents each public toilet as an individual data point on the map, and is therefore more useful in assessing the state of particular toilets (Figure 13). This representational strategy presents data in a disaggregate form, and facilitates communities who live around a

public toilet to assess and evaluate its conditions and perhaps even advocate for change on the

basis of the data available.[188]

4.3 *Transparent Chennai* and the Data Intermediary Function

Transparent Chennai started collecting data on public toilets in 2009, because in a

meeting organized with over 300 workers from the unorganised sector in Chennai, female

workers from varied organizations and occupations expressed a need for more public toilets in

the city.[189] Public toilets are an important civic service for workers employed in the unorganized

sector, particularly women, who often have limited or no access to private sanitation.[190] To make

a case for more public toilets *Transparent Chennai* sought to understand the condition and usage

of existing public toilets.

To collect data on public sanitation in Chennai, *Transparent Chennai* followed a multi-

phase data aggregation, rectification, augmentation and dissemination process.[191] In the process

of collecting data on public sanitation, the *Transparent Chennai* team aggregated a list of public

toilets from multiple zonal offices of the *Corporation of Chennai (CoC)* through formal and

informal channels, including visits to zonal offices and a *Request to Information* application.

However, this initial list of public toilets had multiple discrepancies, such as toilets with

incomplete addresses and toilets that were non-existent on the ground (Figure 14).[192] So, this

initial list of public toilets was updated to reflect the location and situation of public toilets on the

[189] "Public Toilets in Chennai," India Sanitation Portal, accessed April 27, 2015,
http://www.indiasanitationportal.org/1148.
[190] "Public Toilets and Sanitation," Transparent Chennai, accessed April 1, 2015,
http://www.transparentchennai.com/public-toilets-and-sanitation/.

[192] I discuss this in greater detail in the following section

ground through baseline surveys. Subsequently, the data was also updated to include additional

parameters that could be used to gauge the functionality and condition of existing toilets

(Figure 17) In the process of generating this data, the *Transparent Chennai* team also shared this

data with city officials and residents through public meetings and survey reports, supporting

communities in interactions with their elected representatives based on the data and findings

from the survey.

After these surveys, according to Raman, there was a marked improvement in the

conditions of public toilets that were surveyed. Subsequently, in 2013 the *CoC* announced that it

would install pre-fabricated toilets in the city, *Transparent Chennai* worked with junior

engineers of the city to map the location of existing and proposed toilets, and assisted the *CoC* in

collecting geo-spatial data and assessing the feasibility for the proposed toilet locations.[193]

Furthermore, the additional parameters that *Transparent Chennai* added to the government data

for evaluating the conditions of public toilets supported the creation of a performance monitoring

metric that was a part of the tender released by the *CoC*.[194] Having given an introduction to the

different phases involved in aggregation and dissemination of Chennai public sanitation data, in

the following sections I turn to the data intermediary practices employed during each phase of

the process.

4.3.1 Practices of Aggregation: Collecting Government Data from Multiple Sources

While the *Corporation of Chennai* is all set to launch a data dashboard in collaboration

with *Transparent Chennai*, thus far municipal datasets have been available to the public only

---

[193] "Development of Modular Public Conveniences in DBFOT in Chennai," Transparent Chennai, accessed May 4, 2015, http://www.transparentchennai.com/development-of-modular-public-conveniences-in-dbfot-in-chennai/.
[194] Ibid.

through *Right to Information* (RTI) requests.[195] The *RTI* Act passed in the year 2005 requires

government bodies to comply with citizens' requests for information, as well as to make

government data available as computerized records for wider dissemination of public

information. However, getting access to information through *RTI* requests can prove to be a

cumbersome process. The *Transparent Chennai* blog documents the experience of members of

its research team in seeking data from the government, which often involves in-person visits by

the researchers to the offices of *CoC* and filling out extensive forms to seek information.

The *Transparent Chennai* team had to undergo an elaborate process of requesting public

toilets data from the different zonal offices in the city because the *CoC* did not maintain this data

in a consolidated form.[196] This involved visits to ten zonal offices, and in-person meetings with

city government officials for the release of data.[197] Narrating the experience of the member of a

research team in one of the zonal offices, Raman writes: [198]

> Although the Assistant Commissioner had personally directed the request letter
> to the Engineer, the Engineer was not sure whether the list of toilets could be
> given out to a member of the public without explicit approval from the
> Corporation Commissioner, the senior-most bureaucrat in the city. Meryl told
> him that she had collected the same data from seven other zones without any
> problem. The Engineer nodded, and asked her whether she had already obtained
> the data from Zone 9. She said yes, and he then called the Assistant
> Commissioner of Zone 9 to ask whether it would be prudent to give out the
> information. Finally satisfied that a list of public toilets was safe to give out to a
> member of the public, he instructed the Engineering Department to prepare the
> list for her. After waiting another 45 minutes, she left the office with a list of the
> 31 toilets in the Zone and their addresses.

---

[195] The Right To Information (RTI) Act was instituted and implemented by the Government of India in 2005 to replace the Freedom of Information Act (FOIA).
[196] At the time of data collection there were ten zonal offices in Chennai, now there are 15. Nithya Raman, "Collecting data in Chennai City and the limits of openness," *The Journal of Community Informatics* 8, no. 2 (2012), accessed January 18, 2015, http://ci-journal.net/index.php/ciej/article/view/877/908.
[197] Ibid.
[198] Ibid.

Through a long drawn process that involved multiple visits to zonal offices, the *Transparent Chennai* team was eventually able to compile a list of public toilets and their locations. Subsequently, it also filed an *RTI* request to access the information on the citywide data of public toilets. [199] However, this process also involved contacting and following up with the zonal offices, and took longer than the stipulated 30 days for release of information sought through an *RTI* request.[200]

This example illustrates the complexities of getting access to municipal datasets from the *CoC*, and highlights the role of *Transparent Chennai* in releasing information that is difficult to get access to. Hence, one of the critical functions that *Transparent Chennai* serves as a data intermediary in the Indian context is opening up government data that is difficult to access. It also highlights the fact that even when the *RTI* act states that Indian citizen's have a right to government information, the rules of access to government data are not always systemized, and may act as barriers to the citizen's access to government data.[201]

4.3.2 Practices of Rectification: Correcting Irregularities in Government Data

On comparing the data from informal and formal channels, the *Transparent Chennai* team realized that the data on the number of toilets collected from the zonal offices through personal visits and the data collected through the *RTI* request did not match (Figure 14). According to initial data collected from visits to zonal offices, the total number of toilets in Chennai was 572, whereas the data collected through the *RTI* request indicated the total to be 714.

---

[199] The *RTI* act requires that a response be made to a request within 30 days of receipt. Ibid.
[200] Ibid.
[201] Raman notes that there is a "culture of fear" among lower level city officials to release data without the approval of their seniors. Nithya Raman, "Collecting data in Chennai City and the Limits of Openness."

Zone	Initial Data	Rtl Data
1	31	49
2	72	51
3	82	133
4	20	49
5	60	69
6	41	61
7	35	77
8	72	74
9	60	58
10	99	94
Total	**572**	**715**

Figure 14: The aggregated zonal public toilet data, showing the discrepancy between data collected from visits to zonal offices (initial data) and that collected through *RTI* request. Source: Transparent Chennai

In order to verify the data collected, the *Transparent Chennai* team conducted a series of baseline surveys in zone 4 of the city, with volunteers physically mapping and validating the list of public toilets received from the *Corporation of Chennai* (Figure 15).[202] This exercise revealed that the list of public toilets was not comprehensive and inconsistent. Researcher Somya Sethuraman discusses the difficulty of locating the public toilets based on the street names they were located on, listed on the data obtained from the *CoC*:

> With only the street names with us, it was extremely difficult to locate these [public toilets] on the maps…Second, there were some street names which were at multiple locations on the map, making our search all the more difficult. But this was only the beginning. Once located on the map, the same had to be discovered in the field. Third, some of the streets did not even have a mention

---

[202] Ibid.

on any of the maps we referred to. So, we just went blindly to Zone 4 and began our search.[203]

Moreover, many toilets that were listed on the data collected from the Zone office seemed to be non-existent on the ground, and some toilets that the volunteers located on the ground did not exist in the government dataset altogether.[204] While some of the toilets listed in the government were found to have incorrect addresses, others were found to be demolished or to be in a state of disrepair and disuse.[205] Raman states that anecdotal evidence from interviews with local individuals suggests that some zonal level city officials might benefit from keeping non-existent public toilets in government registers because there is an incentive to create "fictional maintenance contracts" for these toilets, which could be a source of income for city officials.[206] Indeed, an article in *The Hindu* reports on attempts of some ward councilors to influence the auctioning process of maintenance contracts for monetary benefit.[207] This discussion also suggests that incorrect or obsolete government data can be used to support corrupt practices.

Discrepancies in government data also raise important questions about the quality of civic data in the city.[208] To understand how the *CoC* collects, stores, and disseminates data, *Transparent Chennai* conducted surveys, in-depth meetings with city officials, and public meetings in an effort to evaluate the existing data practices of *CoC*.[209] Through this research it

---

[203] Somya Sethuramanan, "Transparent Chennai Team Gets Involved in a New Initiative!" *Transparent Chennai Blog*, accessed May 7, 2015, http://www.transparentchennai.com/122/.
[204] "Putting Toilets on the Map," Transparent Chennai, accessed April 1, 2015, http://www.transparentchennai.com/putting-toilets-on-the-map/.
[205] Ibid.
[206] Ibid.
[207] "Toilet Maintenance Contracts Face Problems" *The Hindu*, accessed May 7, 2015, http://www.thehindu.com/todays-paper/tp-national/tp-tamilnadu/toilet-maintenance-contracts-face-problems/article1854011.ece.
[208] Ibid.
[209] Satyarupa Shekhar and Vinaya Padmanabhan, "Still Open Ended," (research poster, Transparent Chennai, Chennai, 2014), accessed March 18, 2015. http://www.opendataresearch.org/project/2013/tc.

identified four factors that contributed to poor quality civic data in Chennai: 1) lack of

appropriate skills and methodology for data collection, 2) data formats that were focused on

infrastructure rather than services, 3) lack of standard formats across departments, 4)

unwillingness to share data.[210]

In case of public sanitation, poorly maintained data demonstrates lack of mechanisms to

assess and evaluate access to public sanitation in the city.[211] Lack of effective data practices

perhaps even results in inadequate planning of public sanitation. This is evidenced by the

findings of the public sanitation study conducted by *Transparent Chennai*, which highlight that

public toilets are not located in poorer areas of the city where they might be needed the most.[212]

For instance, from their survey of public toilets in zone 4 of Chennai, *Transparent Chennai*

found that public toilets were located away from slums, markets, and bus stops, and were found

in areas with little foot traffic.[213] In one particular case, a cluster of slums with a population of

about 30,000 people in zone 4 was found to have no public toilets.[214]

4.3.3 Practices of Augmentation: Supplementing Government Data

The public sanitation data collected from the government offices was a list of public

toilets and their addresses, but it lacked geospatial data, and data on administrative and sectoral

boundaries, which limited its usability.[215] For instance, none of the official data releases included

maps or shapefiles, which is a format typically used for geospatial data, making the data

---

[210] Ibid.

[211] Ibid.

[212] "Transparent Chennai Research Highlights Inequity in Service Provision," Transparent Chennai, accessed April 27, 2015, http://www.transparentchennai.com/transparent-chennai-research-highlights-inequity-in-service-provision/.

[213] Transparent Chennai, "Public Toilets in Chennai," (Research Report, Transparent Chennai, Chennai, India, 2011), 3, accessed March 18, 2015, http://bit.ly/1zVThMH.

[214] Ibid.

[215] Nithya Raman, "Collecting data in Chennai City and the Limits of Openness."

unusable for mapping.[216] A critical limitation of government data is that it lacks attributes that

will enable communities in their demand for services and monitoring changes in the community.

In the context of the public toilet discussion this would mean that the data should reflect the

current condition and use of public toilets, and include attributes that describe lighting,

cleanliness, water, and time of operation.

In the process of validating government data during baseline surveys, the volunteers of

Transparent Chennai also collected data on the existing conditions of public toilets in Chennai

by locating public toilets and speaking with their caretakers to evaluate their use and

maintenance (Figure 15).[217] The volunteers and researchers from *Transparent Chennai* not only

mapped the physical locations of toilets, but also added other parameters such as water, lights,

doors, cleanliness and so on, supplementing the government data with information that could be

used to evaluate the condition and usability of existing toilets. The goal of this exercise was to

augment government data to help identify toilets that are in a state of disrepair and are currently

unusable or have limited use for public sanitation.[218]

[216] Ibid.
[217] "Putting Toilets on the Map," Transparent Chennai, accessed April 1, 2015,
http://www.transparentchennai.com/putting-toilets-on-the-map/.
[218] Ibid.

Figure 15: Volunteers of *Transparent Chennai* collecting data about public toilets. Source: Transparent Chennai

This data was then used to create a ward-wise map of the usable and unusable toilets (Figure 16).

The map represents the unusable toilets as red dots, and the usable toilets as green dots within the

administrative boundaries of the ward 121.
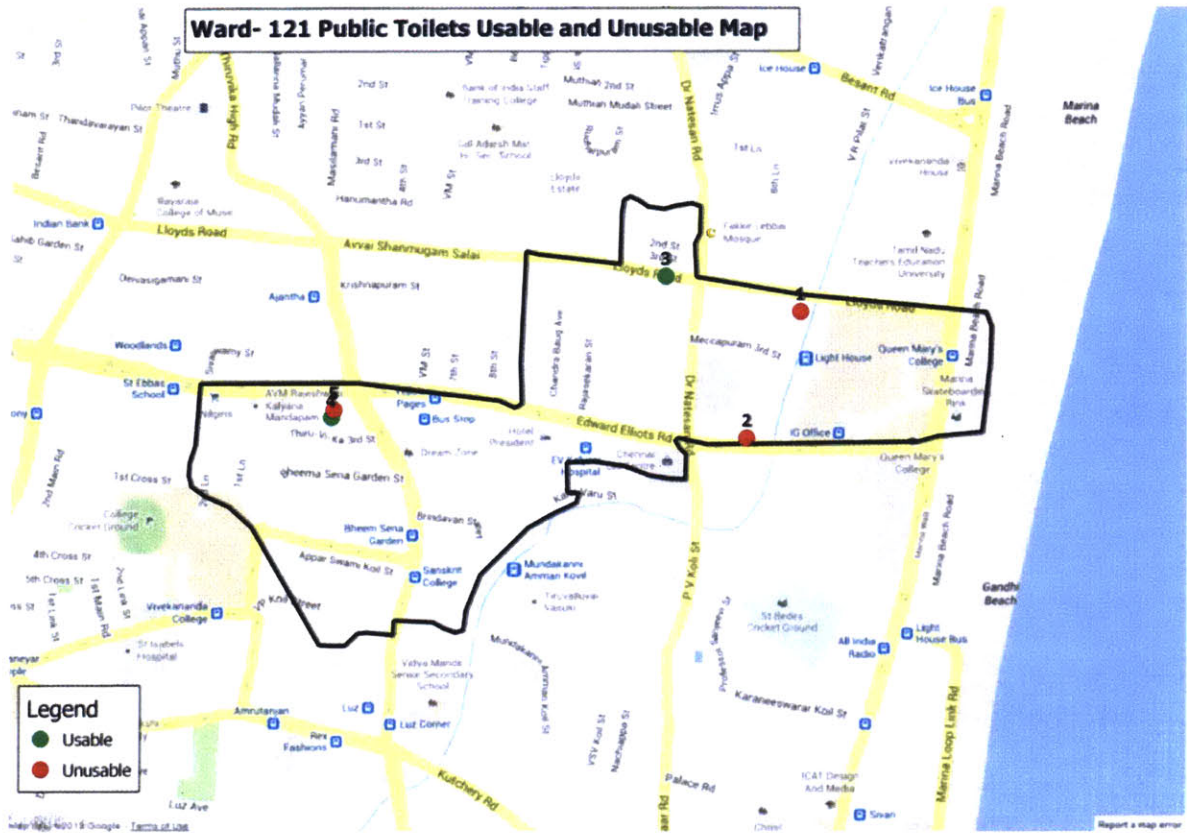
Figure 16: The map of ward 121 showing the usable and unusable toilets. Source: Transparent Chennai

Date of survey	Toilet number	Quality of infrastructure							Quality of service			
		Usable	Unusable due to :					Cleanliness	Free/ Pay-use	Time of operation		Cleaning per week
			Door	Light	Water	Blockage						
12/6/2013	121/1	X	Missing locks on some doors (women)	Some lights missing both women's and men's cubicles	No water in taps/ no collected water	Visible blockages both men and women's side		Dirty	Pay-use	04:00:00 AM	08:30:00 PM	14
12/6/2013	121/2		Closed					--	--	--	--	--
12/6/2013	121/3	Yes							Free	05:00:00 AM	10:00:00 PM	14
12/6/2013	121/4	YES						Dirty	Free	24hrs	24hrs	7
12/6/2013	121/5	X	No doors	Lights in cubicles not working	No water in taps/ no collected	Visible blockages in toilets		Dirty				7

Figure 17: The public toilet data detailing the quality of infrastructure, quality of service, and access conditions of public toilets in ward 121. Source: Transparent Chennai

The additional parameters collected in this process were translated into an excel sheet that enumerated the quality of infrastructure, quality of service, and access for a particular toilet, (Figure 17). "Quality of Infrastructure" presents data collected on the material conditions of the toilets with attributes such as cleanliness, lighting, and so on. "Quality of Service" offers data on time of operation, the usage fee, and the number of cleanings per week, and data on "Access" lists the statistics of use. In addition to listing attributes such as cleanliness and hygiene, this data also links these attributes to the usability of a toilet. For instance, toilet number 1 was found to be unusable because of missing locks and no tap water. With these additional parameters, it becomes possible to assess and evaluate the conditions of public toilets, to understand why they are unusable, and to potentially address the problem, which was not the case with the government data on public toilets. It is important to note here, that as discussed previously, one of the shortcomings of the data practices of the *Corporation of Chennai* was that it uses data formats that focus on infrastructure rather than services. In contrast, by focusing on the usability of public toilets, the data format used by *Transparent Chennai* addresses the limitation of government data format used for public toilets.

In addition, through community mapping exercises, *Transparent Chennai* worked in collaboration with vulnerable communities to help them voice the problems they face in relation to public sanitation (Figure 18). For instance, during the data collection process a community mapping exercise was held in Sivarajapuram, a slum area, during which *Transparent Chennai* in collaboration with a group of participant slum dwellers, created a paper map to identify and document public toilets in the vicinity (Figure 19).

95

Figure 18: A community mapping exercise in progress in Chennai. Source: Vinaya Padmanabhan, Transparent Chennai
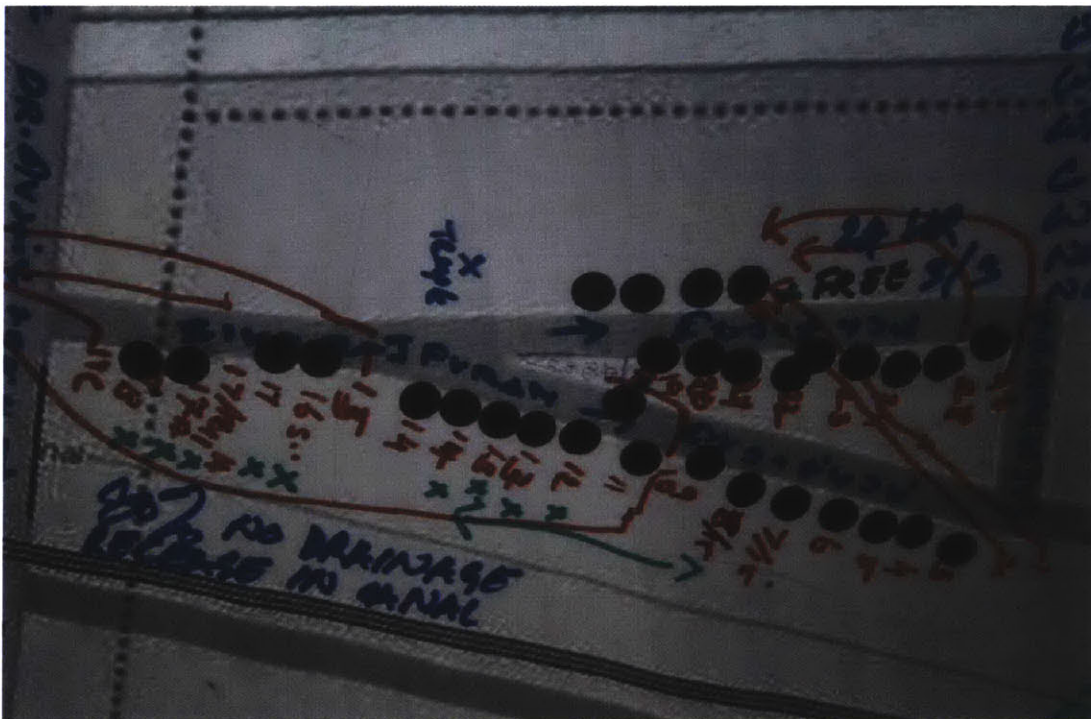
Figure 19: A section of the map created during the community mapping exercise, showing homes (black dots), private toilets (green crosses) and public toilets. Source: Vinaya Padmanabhan,Transparent Chennai

The *Transparent Chennai* blog documents the proceeding of one such session, where the

Transparent Chennai team used paper mapping tools to facilitate community participation in

data collection.[219] Nidhi Subramanyam, a member of the research team, writes:

> Initially, community members were oriented with respect to a base map, and
> asked to create a community base map by marking bus stops, temples, and other
> landmarks they thought were important. They were then asked to identify public
> toilets in the vicinity that they frequented daily. Initial anxiety and inability to
> orient themselves was quickly overcome when a few members, of notable
> mention is Ms. Maarithalai, enthusiastically started marking symbols on the map
> by consulting and debating with each other.[220]

This excerpt suggests that while mapping data has the potential to spark a discussion within

communities, it still requires considerable skill to read and understand maps. Providing local

context helps orient communities in physically mapping data, and the use of primitive tools, such

as paper maps, perhaps facilitates this process. Communities manage to assert an agency and

voice in the discussions regarding issues that affect them through community mapping exercises,

but need facilitation and guidance to overcome the initial barriers in contributing to these

exercises.

Satyarupa Shekhar, a researcher at *Transparent Chennai*, contemplates the question of

barriers to access and the role technology might play in exacerbating these in a blog post that

describes the idea of creating a short messaging service (sms) based data feedback system for

public toilets. She notes:

> First, what about those toilets where the users of the toilets are largely passersby,
> rather than local residents, who have little or no incentive in the upkeep of the
> toilet? Would they take the trouble to text their feedback? There would be cases
> of some toilets figuring in the monitoring system more than others but it would

---

[219] Nidhi Subramanyam, "Community Mapping Exercise at Sivarajapuram," *Transparent Chennai Blog,* accessed April 13, 2015, http://www.transparentchennai.com/community-mapping-exercise-at-sivarajapuram/.
[220] Ibid.

be incorrect to conclude that those are toilets, which are highly dysfunctional or that the others are perfectly functional.[221]

This comment demonstrates an awareness of the questions that participatory data collection is fraught with, such as who participates in the process and why, and the biases implicit in the resulting data collected. Raman also considers the ethics of creating data in collaboration with communities in the Indian context:

> In the Indian context you have to be very careful of not promising too much, and not taking too much time, because the power imbalances between those who collect data and those who are affected are much more.[222]

This analysis raises important considerations about the biases that can be introduced through data intermediary practices, and suggests that data intermediary practices need to account for existing power structures and consider how these might get embedded in the data being collected and disseminated.

4.3.4 Practices of Dissemination: Sharing Data with Communities and City Officials

Transparent Chennai disseminates data through a range of methods including online documentation, research reports and papers, and community outreach. It publishes reports in regional languages and circulates them within so that communities can use them to initiate dialogues with city officials. In addition it prints large copies of maps for communities that helped create them, so that they can use them to advocate for improved service delivery. While it makes data available for download through its website, it is considerate about sharing data publicly, especially in relation to its work with vulnerable communities.[223]

---

[221] "Does Technology Reinforce Real-World Inequities?" Transparent Chennai, accessed May 7, 2015, http://www.transparentchennai.com/does-technology-reinforce-real-world-inequities/.
[222] Nithya V. Raman (Director, *Transparent Chennai*) in discussion with the author, March 2015.
[223] Ibid.

While the online dissemination of data and findings might not necessarily prove to be particularly useful in reaching out to affected communities, it reaches news media outlets, who often cite *Transparent Chennai's* research and data on public toilets in their news articles about the state of public sanitation in Chennai. *The Hindu* emphasizes the lack of maintenance of the public toilets in the city by referencing *Transparent Chennai's* survey of zone 4, where out of the 49 public toilets "only 13 were found to be clean, 27 had no lights within, and approximately 80 per cent of toilets surveyed did not have 24-hour water supply"[224]

---

[224] "Inadequacy and Ill-Maintenance Plague Public Conveniences," *The Hindu*, accessed May 8, 2015, http://www.thehindu.com/news/cities/Chennai/article2455266.ece.

Figure 20: The researchers from the Transparent Chennai team sharing the findings of the public sanitation study with the *Working Women's Forum*, an organization of women in the informal sector. Source: Transparent Chennai

Another important aspect of the dissemination practices of *Transparent Chennai* is that it shares the data and findings from its research with city officials and policymakers. In the case of the public sanitation project, the *Transparent Chennai* team regularly submitted survey reports to the government. In a scenario where government offices are understaffed and lack the necessary resources to conduct regular surveys, access to data that highlights the usability and unusability of sanitation infrastructure helped ward councilors in prioritizing their limited resources for maintenance and service delivery.[225]

---

[225] Arundhati Ramanathan, "Transparent Chennai | A Better Life by the Buckingham Canal," *Livemint*, accessed April 27, 2015, http://www.livemint.com/Leisure/NNwIgnXLnjSxaGN0qLZwOK/Transparent-Chennai--A-better-life-by-the-Buckingham-Canal.html.

4.3.5 Practices of Connection: Mediating Interactions Between Communities and City Agencies

In order to foster connections between city officials and communities, *Transparent Chennai* organized public meetings to disseminate the findings from its research and surveys. In these meetings the needs of communities are often amplified by the findings and data collected by *Transparent Chennai*.[226] After a few such meetings in Kalyanapuram, a slum area in ward 57, Saran, a resident observed:

> The meetings made us believe that our vote was bringing us results. Just because we live in a slum, Transparent Chennai has taught us to not sit quietly and instead claim all that is entitled to us.[227]

This comment suggests that *Transparent Chennai* facilitates greater interaction between communities and elected representatives, and thus supports them in demanding for change from the city government.

The *Transparent Chennai* team has also worked with the city government on multiple occasions and has been able to create a dialogue centered on the need for better quality data in Chennai. In the case of public sanitation data, after the *Corporation of Chennai (CoC)* announced in 2013 that it would install prefabricated toilets in the city, *Transparent Chennai* worked with city officials responsible for planning the location of the new toilets to virtually map out the location of existing and proposed toilets.[228] It also helped them assess the feasibility of the proposed locations by assisting them in collecting geospatial data and in using it with existing data on land availability, and water and sewerage systems.[229] It did so, by creating a

---

[226] Ibid.
[227] Arundahti Ramanathan, "Transparent Chennai | A Better Life by the Buckingham Canal," *Livemint.*
[228] "Mapping and the Municipality," Transparent Chennai, accessed May 6, 2015, http://www.transparentchennai.com/mapping-and-the-municipality/.
[229] Ibid.

mobile application that simplified the process of collecting geospatial data and visually

represented the data on a map using CartoDB.[230] This map of the proposed toilet locations was

subsequently used in the tender issued by *CoC*, written in collaboration with Transparent

Chennai, soliciting proposals for installing public toilets in the city.[231] This project is currently

under implementation. *The Times of India* has reported that while work orders have been issued

to four firms for installing 228 public toilets, only 41 new toilets were installed in Chennai as of

February, 2015.[232]

More recently, *Transparent Chennai* and the *Corporation of Chennai* have signed a

Memorandum of Understanding (MOU) to launch *Chennai Data Portal*, the first city-level open

data dashboard in India.[233] The *Transparent Chennai* team is closely involved in making

recommendations about data practices that the portal needs to adopt, and the technology and

human resources required for this implementation.[234] In addition, the team is also responsible for

understanding and evaluating the existing data practices of the government, and in designing data

collection methodologies to create data "about neglected civic issues" in Chennai.[235] This

suggests that by demonstrating the potential of urban data for planning and service delivery, data

intermediaries and can help articulate the demand for opening up of government data.

---

[230] Cartodb is a web-mapping tool that can be used to represent geospatial data. Ibid.
[231] Ibid.
[232] Christin Mathew Philip,"Residents Say 'Not in My Backyard' to Public Toilets," *The Times of India*, accessed May 8, 2015, http://timesofindia.indiatimes.com/city/chennai/Residents-say-not-in-my-backyard-to-public-toilets/articleshow/46127445.cms.
[233] "Jobs," Transparent Chennai, accessed April 27, 2015, http://www.transparentchennai.com/jobs/.
[234] Ibid.
[235] Ibid.

4.4 Conclusions

In this chapter, I have sought to highlight how the data intermediary practices of

Transparent Chennai respond to the challenges and particularities of the open data context in

Chennai. Through its practices of aggregation *Transparent Chennai* successfully navigates the

complex bureaucratic processes of the city to seek release of public sanitation data. Its practices

of rectification and augmentation emphasize its capacity to correct and enhance government

datasets. At the same time the open data context it operates within intensifies the amount of labor

needed to implement these practices, and limits the scale at which such practices can be

implemented. For instance, while on the one hand, *596 Acres* is able to use virtual mapping

resources such as Google street view and Satellite images for verifying government data, on the

other *Transparent Chennai's* rectification process involves manually verifying the addresses of

the public toilets.

Transparent Chennai employs participatory community mapping exercises and

interviews to solicit inputs from communities for its practices of augmentation, but it does not

provide communities with a more direct mechanism of feedback. In considering to implement an

SMS-based feedback system for public toilets, *Transparent Chennai* wrestles with concerns over

such a system unfairly benefiting certain communities over others. Such ethical considerations

suggest that data intermediary practices are not free from potential bias.

Transparent Chennai's practices of dissemination and connection create opportunities for

data-driven dialogue between city government officials and communities through public

meetings. Its association with the *Corporation of Chennai,* as evidenced by its work to determine

the location of proposed public toilets in the city and the collaboration on the *Chennai Data*

Portal, suggests that data intermediary practices, which demonstrate the potential of civic data and highlight the lack of good quality data on civic issues, can stimulate an assessment of existing data collection and dissemination practices of the government, and perhaps even help inform and shape these practices.

Through this discussion I have sought to highlight some of the salient aspects of a data intermediary in the Indian context. I have also examined in a preliminary way some of the similarities and differences in the data practices of *596 Acres* and *Transparent Chennai,* who operate in very different open data contexts, and I continue this comparison in greater detail in the following chapter.

The Comparative Contexts of Data Intermediaries and Some Considerations for Further Work

In an effort to understand the role of data intermediaries as a crucial basis for open government data usage, we may define the data intermediary function, as one constituted by a range of data practices that data intermediaries employ in making open government data intelligible and usable for communities. Ordinarily, we assume that the motivation behind data intermediaries is civic, or, to support greater comprehension, transparency and utility of data by the public. Having explored data practices of two intermediaries working to make open government data useful for communities to effect change in the built environment, I would like now to compare and contrast the data practices of the two intermediaries using the framework proposed in Chapter Two, and then to discuss some of the challenges faced by data intermediaries, and especially three opportunities for future research.

5.1 The Data Intermediary Function: A Comparative Perspective

In examining the similarities in data practices of the two intermediaries in New York City and Chennai, I will highlight data practices that seem to be effective in diverse settings. I will also underscore some differences in the two sets of practices and explore how certain data practices depend on open data context and consider the limitations of open government data specific to an intermediary.

5.1.1 Practices of Aggregation

While both *596 Acres* and *Transparent Chennai* collect and combine multiple

government datasets, the data practices they employ are informed by the open data context they

operate in. In order to create a list of public vacant lots in New York City, for example, the *596*

Acres team used two datasets available for download from the NYC open data portal in a

machine-readable and reusable format. In contrast, the government dataset on public toilets in

Chennai was not available in a readily accessible and machine-readable format, and the

Transparent Chennai team had to approach ten zonal offices in Chennai individually for the

release of public sanitation data. Even when the Chennai data was released, it was often given in

a format that was not machine-readable. For instance, one particular zonal office released the

data as a hand-written list, which the *Transparent Chennai* team had to manually type up as an

excel sheet.[236] After receiving this data from zonal offices, the team had to repeat the process of

collecting public toilet data by filing for an *RTI* request that also involved following up with each

of the zonal offices and in some cases took more than thirty days for the release of data.

Comparing the aggregation practices of *596 Acres* and *Transparent Chennai* reveals that

aggregation contexts are critical to data intermediary functions. While *596 Acres* could combine

multiple datasets readily available from the NYC data portal in an open format to create an initial

list of public vacant lots in New York City, *Transparent Chennai* had to undergo a time and

resource intensive process to get access to public sanitation data from multiple city offices, and

to consolidate it to create an initial list of public toilets in the city. In the case of *596 Acres*, the

practices of aggregation primarily serve a technical function, involving downloading and

---

[236] Many government offices in India, still use paper-based registers for recording administrative information.
Nithya Raman, "Collecting data in Chennai City and the limits of openness," *The Journal of Community*
Informatics 8, no. 2 (2012), accessed January 18, 2015, http://ci-journal.net/index.php/ciej/article/view/877/908.

combining multiple datasets, but in the case of *Transparent Chennai* the practices of aggregation serve a predominantly civic function, involving the acquisition and opening up of critical public sanitation data through navigating complex bureaucratic processes that pose a significant barrier of access to government data.

5.1.2 Practices of Rectification

In rectifying government data, both *596 Acres* and *Transparent Chennai* use methodological triangulation that involves comparing and combining data collected from multiple data sources.[237] *596 Acres* verified the initial public vacant lot data by examining the lots through virtual mapping resources such as Oasis NYC, Google Streetview, and satellite imagery, and by soliciting information from neighbors both through online and offline methods. In contrast, *Transparent Chennai* verified the initial data collected from the government by conducting its own survey of the public toilets in zone 4 and by interviewing the caretakers of the public toilets on the usage and conditions of public toilets. Both *596 Acres* and *Transparent Chennai* were able to update the initial data to reflect more closely the existing conditions on the ground. It is important to note, however, that *Transparent Chennai* was unable to use virtual resources, because such resources in India are often less robust than in the United States. Google Street view is only available in selected Indian cities and a spatial information system comparable in granularity and detail to the OasisNYC does not exist for the city of Chennai.[238] As the case of *596 Acres* suggests, examining spatial data through virtual information systems

---

[237] Methodological triangulation is the use of two or more methods during a research study to check the validity of the data collected and the findings gleaned.
[238] OasisNYC is a spatial information system that is a resource for community maps in the New York City.

can prove to be a useful precursor to the actual verification of the spatial data on the ground.[239]

However, the *Transparent Chennai* team cannot employ a similar strategy, because the geographic and spatial information resources in the case of Chennai are not well developed yet.

Note that by helping the neighbors of a lot report the current status of the lot through its online interactive tool, *596 Acres* has managed to involve the neighbors of a lot in the rectification process. *596 Acres* has been able to create a mechanism that allows for updating and correction of the public lot dataset that it maintains. In rectifying government data on public toilets, *Transparent Chennai* also solicits data from the neighbors of a public toilet, but it does so by interviewing caretakers and other local individuals during the zonal survey. One limitation of this approach is that it does not provide communities with a more direct mechanism of feedback, and lack of such mechanisms prevents data updating from becoming a self-sustained activity.[240] By facilitating comments and inputs, the organizing pages of *Living Lots NYC* support civic behavior and serve as online spaces for organizing and watching the continually changing conditions of a particular lot. So, while the practices of rectification of *596 Acres* serve a critical and civic function by supporting neighbors of a vacant lot to report the current condition of a lot, those of *Transparent Chennai* serve primarily a critical function.

---

[239] Using the virtual spatial information system, the *596 Acres* team was able to remove many public lots that it deemed unfit for community use, and therefore was able to save the resources that might have gone into the verification of these resources on the ground.

[240] *Transparent Chennai* was planning to implement a SMS (short messaging service) based feedback mechanism to allow for feedback from users of a toilet. For a more detailed discussion on the ethical considerations of doing so, consult the previous chapter.

Table 1: Summary of 596 Acres' Data Intermediary Practices

Aggregation	Rectification	Augmentation	Interpretation
Technical	Critical, Civic	Critical, Civic	Critical, Technical
1) Combining data from two datasets on the NYC open data portal to create an initial list of public vacant lots. Function: Technical	1) Removing vacant lots that are not fit to be used as community gardens by using virtual mapping resources. Function: Critical, Technical	1) Soliciting notes and images from neighbors on the web page of the lot to make the data more qualitative. Function: Civic	1) Removing lots that were being used as community gardens to make the dataset more community and use focused. Function: Critical, Technical
	2) Correcting the database to reflect the current use by soliciting inputs from neighbors of a lot. Function: Critical, Civic	2) Adding new vacant lots to the database based on comments on the web page. Function: Critical, Civic	

Representation	Dissemination	Connection
Representational, Civic	Civic, Technical	Civic
1) Disagreggating the dataset and representing it on an online map to make the data contextual. Function: Representational	1) Creating posters based on the dataset to highlight the availability of vacant lots. Function:Representational, Civic	1) Providing contact name and contact details of relevant city council representatives and elected officials, and guidelines for how to approach them. Function: Civic
2) Representing the dataset as individual datapoints that open up as commentable web pages, facilitating formation of communities of interest around a particular lot. Function:Representational, Civic	2) Helping communities in using the online tool and data by organizing outreach workshops. Function: Civic	2) Facilitating in-person meetings and collaboration between organizers of a particular lot through the online tool. Function: Civic
	3) Releasing data and source code to facililitate creation of online land access advocacy tools. Function: Civic, Technical	

Table 2: Summary of Transparent Chennai's Data Intermediary Practices

Aggregation
Civic

1) Aggregating a list of public toilets by personal visits to zonal offices in the city and filing an *RTI* request.
Civic

Rectification
Critical

1) Comparing the list obtained from office visits and *RTI* request to highlight its discrepancies.
Critical

2) Validating and correcting the addresses of public toilets by physically mapping the location of toilets.
Function: Critical

Augmentation
Critical, Technical

1) Adding attributes such as GPS data, cleanliness, ligting, etc. to the data making it more useful to assess and evaluate the condition of toilets.
Function: Civic, Technical

2) Using community mapping exercises to document public toilets and to help communities in voicing their concerns.
Function: Civic

Interpretation
Critical

1) Highlighting the shortcomings of government data practices in collecting, storing, and disseminating data.
Function: Critical

2) Articulating the demand for opening up of government data by demonstrating the potential of civic data for planning and service delivery.
Function: Civic

Representation
Representational

1) Representing each public toilet as an individual data points on a map, which open up to show profiles of the toilet.
Function: Representational

2) Representing ward-wise aggregated data on a map to compare and contrast the public sanitation conditions in different wards.
Function: Representational

Dissemination
Civic

1) Distributing paper-based maps and reports in local languages to support local communities in the use of data.
Function: Civic, Representational

2) Sharing data and findings from its research with city officials and policymakers.
Function: Civic

Connection
Civic

1) Organizing public meetings to support data-based interactions between communities and elected representatives.
Function: Civic

5.1.3 Practices of Representation

596 Acres and *Transparent Chennai* both disaggregate datasets into individual data points and present them in a spatial context through maps, and use online interactive maps in conjunction with paper-based maps to represent government data. Moreover, both Paula Z. Segal and Nithya V. Raman argue that maps are useful in giving context to government data and in helping people to relate to the data being represented.

In the case of 596 Acres, *Living Lots NYC,* the online interactive tool, represents the public vacant lots in the different boroughs of New York City as individual data points. Each data point not only represents the public vacant lots but also facilitates formation of communities of interest around a particular lot. Thus, the representational strategy employed by *Living Lots NYC* supports organizing around particular pieces of land, facilitating action of individuals who want to organize around particular lots.

In the case of *Transparent Chennai,* each public toilet on the map is represented as an individual data point, and on clicking, opens up to show the underlying data in a granular form. But, this online map does not support further data based interactions. The online maps were created to serve primarily as a tool of dissemination, and while the citizen-generated layer of the tool was created to solicit data inputs from individuals, individuals who are most affected by the lack of public toilets, will be not be able to use these tools without facilitation.[241] Hence, we can argue that while both *596 Acres'* and *Transparent Chennai's* practices of representation serve a civic function by making the data contextual. *596 Acres,* practices of representation also supports civic behavior by facilitating communities to form around data.

---

[241] Nithya V. Raman (Director, *Transparent Chennai*) in discussion with the author, March 2015.

5.1.4 Practices of Augmentation

In augmenting government data, both *596 Acres* and *Transparent Chennai* use participatory methods. In the case of *596 Acres,* this happens when members of the *596 Acres* team and organizers of a particular lot comment on its organizing page with additional data, including notes and pictures about the present condition of the lot and comments on a lot's history.[242] In the case of *Transparent Chennai*, the data obtained from government (the list of public toilets and their addresses) is updated and augmented to include additional attributes such as administrative and sectoral boundaries, geographical coordinates, images, and other qualitative attributes such as description of usage and infrastructure. This is done through a survey conducted by volunteers of *Transparent Chennai* who interview caretakers and local individuals around public toilets. In addition, *Transparent Chennai* also made use of paper-based maps for participatory data collection from communities affected by public sanitation issues. So, practices of augmentation of both *596 Acres* and *Transparent Chennai* contribute in making government data more textured and granular.[243]

It is important to note, however, that while *596 Acres'* augmentation practices serve a civic function by supporting communities in adding qualitative data about the lot, *Transparent Chennai's* augmentation practices serve a civic function by adding attributes to government data that can help in assessing, evaluating, and monitoring the conditions of toilets. Practices of augmentation employed by both *596 Acres* and *Transparent Chennai* highlight the importance of capturing the knowledge that exists within communities about the built environment, which is often not captured in government datasets. Indeed Raman, observes that in the Indian context

---

[242] See Chapter Three for a more detailed discussion.
[243] For a more detailed discussion on the critical need of granular data see the following section.

there is a great opportunity to capture qualitative data about an issue through community

mapping exercises:

> If you show a map of a neighborhood [to] people, [they] can identify; here are
> the problems, here is where the flooding happens, here is where there are
> houses with less sanitation. [244]

While there is an opportunity to create data through participatory methods, it is important to

note that data generation often is a very resource-intensive process. Both *596 Acres* and

Transparent Chennai create data to address the gaps in government data, but the decision of

whether or not to generate this data involves weighing the effort required to create the data

against its anticipated impact.

5.1.5 Practices of Dissemination

Founders of both *596 Acres* and *Transparent Chennai* underscore the critical need to use

a combination of online and offline methods to disseminate data, suggesting that through their

dissemination efforts data intermediaries attempt to reach individuals with a diverse range of

data literacies. Segal notes, "Without support communities will not be using tools. Academics

and people who like [to use] tools, [only] they would be using the tools."[245] Needless to say,

without facilitation, the tools and data created by data intermediaries would fail to reach

individuals and communities who lack the technical skills to access and use these tools.

Similarly, Raman highlights the critical importance of community outreach, workshops, and

---

[244] Nithya V. Raman (Director, *Transparent Chennai*) in discussion with the author, March 2015.
[245] Paula Z. Segal (Executive and Legal Director of the NYC land access program, *596 Acres*) in discussion with the author, January 2015.

publications in local languages that help disseminate data to affected communities in an easy to understand form.[246]

Dissemination happens not only through the use of data and tools created by intermediaries but also through the diffusion of data practices developed by intermediaries. *596 Acres* makes the code used for developing *Living Lots NYC* freely available for download and has partnered with organizations in other cities to create land access tools. Similarly, *Transparent Chennai* has shared its data intermediary practices with the *Corporation of Chennai* and helped with the creation of the *Chennai Data Portal,* the first city-level data portal in India. In sharing the techniques and methods they have created and honed, intermediaries' dissemination practices serve a technical function. In the process, these practices also support the creation of similar tools and data serving a civic function.

5.1.6 Practices of Connection

Both *596 Acres* and *Transparent Chennai* support communities in developing interactions with city officials. *596 Acres* connects communities to city officials by listing the name and contact number of the decision makers from the relevant city department responsible for a particular lot of land. It also offers a step-by-step guide for calling NYC city agencies and using the data from the *Living Lots NYC* to support their conversation with city officials, and securing an interim license for starting a community garden on a public vacant lot.[247] Through a detailed description of the materials required for approaching the city agency (a letter for the

---

[246] Nithya V. Raman (Director, *Transparent Chennai*) in discussion with the author, March 2015.
[247] "I'm about to Call a NYC Agency -- What Do I Say?" Living Lots NYC, accessed May 3, 2015, http://livinglotsnyc.org/resources/im-about-call-nyc-agency-what-do-i-say/.

community board, a name for the garden, a proposal for land use, and so on), *596 Acres*

stimulates participation of individuals in the civic processes of a neighborhood.

Similarly, *Transparent Chennai* organizes ward meetings in which members of the

Transparent Chennai team present findings from the surveys to ward councilors and residents of

a ward, supporting residents in raising their concerns backed by data from *Transparent*

Chennai.[248] By connecting communities with city agencies and officials, *596 Acres* and

Transparent Chennai support communities in exerting agency in their neighborhoods. It is

important to note here that the two data intermediaries not only produce data that is useful for

neighborhood level advocacy but they also support a range of complementary practices that

facilitate communities in having a data-based dialogue with city officials and agencies. It would

require a longitudinal study to evaluate whether or not such practices enable data intermediaries

to stimulate long-term civic behavior in communities or to make such dialogue a self-sustained

activity.

5.1.7 Practices of Interpretation

While practices of interpretation are inherent in some of the practices discussed

previously, in this section I would discuss two sets of practices through which *596 Acres* and

Transparent Chennai interpret government data to highlight the needs of the community.

596 Acres, starts with a list of public vacant lots and through a combination of practices

removes lots that might not be appropriate for community use. To begin with, it removes vacant

lots that are already being used as community gardens by comparing the initial dataset of public

---

[248] Ward is an administrative unit in the city of Chennai, and each ward is represented by a ward councilor. Vinaya Padmanabhan, "Findings from Ward 57 Public Meeting" (public meeting between residents and ward councilor of Ward 57, Chennai, India, May 2013).

vacant lots with data on community gardens in New York City. In doing so, it also emphasizes

the perceived value of community gardens in government datasets.[249] Subsequently, it uses

virtual mapping resources to eliminate public vacant lots that are small in size and therefore not

fit for community-use. Through these practices *596 Acres* interprets the data to make it more

community-focused.

By mapping public toilet data collected from zone 4 against data on slums in Chennai,

Transparent Chennai reveals that public toilets in zone 4 are typically not available in areas

where they might be needed the most. In this process, it brings to attention the lack of planning

and thought in the design and provision of civic amenities, and thus makes a case for the need of

better civic data for the city of Chennai.[250] Through these practices of interpretation both *596*

Acres and *Transparent Chennai* emphasize the limitations of government data, and highlight the

need to evaluate and assess the data collection, storage, and dissemination practices of

government.

5.2 Challenges and Opportunities

5.2.1 Challenge: Lack of Granularity and Specificity of Open Government Data

One of the critical aspects of the quality of government data is its granularity. Granular

data is data that is available at the "level of the individual observation."[251] More detailed data

offers a better representation of the object it represents and supports researchers and practitioners

---

[249] Government datasets list lots that are being used community gardens as vacant land. For more on this discussion consult Chapter Three.
[250] For more discussion on this see Chapter Four.
[251] "Exploring Open Data's Microdata Frontier, "Sunlight Foundation, accessed May 1, 2015, https://sunlightfoundation.com/blog/2014/10/24/exploring-open-datas-microdata-frontier/.

in making and testing specific arguments.[252] Yet, even though some definitions highlight granularity as one of the key characteristics of open government data and suggest that open government data be made available in disaggregate form and with the highest level of detail whenever possible, this is often not the case. Government datasets frequently lack granularity and specificity.[253] For instance, a report by *Capgemini,* a global consulting firm, analyzed public data in 23 countries in 2013 and found only 22% of the open data released to be granular.[254]

In the context of the built environment, granularity of government data means that it has spatial and geographical attributes such as latitudinal and longitudinal attributes or an exact address, and if government data lacks these attributes it is difficult to represent it in the context of the built environment, considerably limiting its usability. For instance, data on the total area of public vacant lots in a city might highlight the city-level opportunity for releasing this land for community use; however, this aggregate data does not necessarily support communities in demanding the release of particular pieces of land. The same data when disaggregated presents each public vacant lot as a data point, and if endowed with appropriate attributes such as the correct address of the lot, and its current use, can enable communities to organize around that particular lot. In order to be effective and to make an impact on the built environment, open government data should be available in a format that is granular and specific. If this data is not available from the government, evaluating whether or not to create this data, requires considering the resources required in producing this data vis-à-vis its potential impact.

---

[252] Ibid.
[253] "The 8 Principles of Open Government Data," OpenGovData.org, accessed April 30, 2015, http://opengovdata.org/.
[254] Dinand Tinholt, "The Open Data Economy: Unlocking Economic Value by Opening Government and Public Data," Capgemini (2013): 4, accessed May 01, 2015. http://ebooks.capgemini-consulting.com/The-Open-Data-Economy/.

5.2.2 Opportunity: Designing for Mechanisms of Feedback and Augmentation

While government data might often not be available in granular form, I have shown that data intermediaries may contribute to the granularity of government data through practices of augmentation. *596 Acres* does so by allowing individuals to contribute qualitative data such as images and notes, and *Transparent Chennai* does so by conducting community-based mapping exercises and surveys. By releasing this augmented data, data intermediaries considerably improve the quality of data that exists in the open government data ecosystem.[255] Data practices of *596 Acres* and *Transparent Chennai* demonstrate that there is an opportunity to design initiatives that feed back into the open government data ecosystem through participatory data updating and collection practices, creating loops of data rather than a one-directional flow of data from data providers to users. In designing for participatory data augmentation however, it is important to consider whether or not individuals with a diverse range of technical skills can participate and whether or not this can be a self-sustained activity, a challenge that I will consider in the following section.

While integrating mechanisms of feedback and augmentation to their data practices, data intermediaries should consider how these mechanisms could interact with existing data collection and dissemination practices of the government. For example, *SeeClickFix*, an online tool that allows citizens to report issues in their neighborhood and subsequently directs them to the relevant city agency, was recently integrated with the *open 311* system that the city of

---

[255] For more discussion on the eco-systemic view of open government data consult Chapter Two.

Chicago uses for tracking civic issues. [256] After this integration, an issue reported on *SeeClickFix*

would directly be stored in the city's 311 database.[257] Furthermore, some open government data

platforms already allow users to discuss and comment on government datasets. Socrata, which is

an open data platform used by multiple city level data portals in the United States, allows

discussions to occur around particular datasets. As city governments explore such feedback

mechanisms, I believe that there is an opportunity for data intermediaries such as *596 Acres* and

Transparent Chennai, who employ participatory data augmentation methods, to inform and

shape these data practices.

5.2.3 Lack of Support for Sustained Use of Data

Founders of both *596 Acres* and *Transparent Chennai* indicate that the bulk of their work

is in supporting communities to make sustained use of data. While data collection, creation and

dissemination are all resource intensive processes, it is particularly difficult to find financial

resources to facilitate the use of data or an online tool after it is created. Community facilitation

efforts are resource intensive, and yet are often not appreciated by funders. At the same time,

founders of both *Transparent Chennai* and *596 Acres* have indicated that their work involves

working closely with communities on the ground, through a range of activities including, but not

limited to, organizing public meetings and workshops to stimulate conversations around data,

distributing paper-based maps, and so on. Segal observes that it is particularly difficult to find

financial resources for the work of facilitators and translators, who support communities in

---

[256] Previously these requests had to be manually entered. Betsy Isaacson, "SeeClickFix Is Now Your 311 (If You Live In Chicago)," *The Huffington Post*, accessed May 4, 2015, http://www.huffingtonpost.com/2012/10/25/seeclickfix-is-now-your-311_n_2017518.html.
[257] Ibid.

continued use of online tools and data created by intermediaries.[258] As I have discussed in

Chapter Two, one of the prominent critiques of open government data is that the uneven use and

access of open government data can create data divides by limiting the use of open data to those

with the necessary skills and resources. While data intermediaries support communities by

creating tools that facilitate interaction with government data, lack of support for continued use

of these tools is definitely a pressing challenge that limits their usefulness.

5.2.4 Opportunity: Designing for Diversity of Uses, Range of Skills, and Replicability of Practices

I believe that the challenge of sustained use of data can be tackled by considering that a

broad range of users including community-groups, land access activists, city officials, and urban

planners, can use open government data to effect changes in the built environment. To support

this, data intermediaries should enable a variety of uses for a range of skills to broaden the

diversity of individuals who have the incentive and ability to participate in data related activities.

For instance, *596 Acres* and *Transparent Chennai* cater to advanced users through their online

interactive tools and to beginners through their community outreach tools and paper-based maps.

In addition, I believe data intermediaries should explore ways to integrate learning resources in

their tools for developing data literacies in communities, and thus facilitate long-term data use in

a more systematized manner.

Another strategy through which the work of data intermediaries can be useful in multiple

contexts is by designing for replicability of data practices and tools. As we have seen in the

previous section, even though the methods used to aggregate and collect data might vary based

---

[258] Paula Z. Segal (Executive and Legal Director of the NYC land access program, *596 Acres*) in discussion with the author, January 2015.

on the open data context that the data intermediaries operate in, data intermediaries can still develop data practices and processes that are applicable to diverse settings. For instance, after creating its New York City land access tool, *596 Acres* partnered with other organizations in Philadelphia, Los Angeles, and New Orleans to create similar community land access tools for those cities. Moreover, it has made the source code used to develop *Living Lots NYC* freely available for download, and offers consultations to organizations that seek to develop similar tools for their cities.[259] Best practices, toolkits, and source codes when openly shared can prove to be very useful for communities who might not have the technical and monetary skills to develop and implement similar tools from scratch.[260]

5.2.5 Challenge: Ethics of Participatory Data Collection

In acting as an intermediary between communities and government data, data intermediaries often mediate the interactions between city officials and citizens, and in doing so become a part of an existing power structure and can either replicate or ameliorate it. In collaborations between data intermediaries and communities, it is important therefore to consider how power is distributed, who exerts it, and what measures are taken to address power imbalances.[261] These questions are especially critical for intermediaries working with vulnerable and affected communities, because the power imbalances in such scenarios are greater.[262] As data practices around open government data evolve there is a critical need to consider both

---

[259] "Maps For Other Cities!" 596 Acres, accessed May 4, 2015, http://596acres.org/en/about/other-cities/.
[260] Ibid.
[261] Sarah Banks et al. "Everyday Ethics in Community-based Participatory Research," *Contemporary Social Science* 8, no. 3 (2013): 267.
[262] Nithya Raman mentions the ethics of managing expectations when working in collaboration with vulnerable communities, and notes that in such scenarios the approval for dissemination of the data collected collaboratively should come from within the community. Nithya V. Raman (Founder, *Transparent Chennai*) in discussion with the author, March 2015.

possibilities: how practices of intermediaries working closely with communities on the one hand can help amplify voices of the community and on the other circumvent community needs altogether.

5.2.6 Opportunity: Designing for Community-Focused Goals, and Formation of Communities around Data

A way in which data intermediaries can address concerns over power imbalances in collaborative relationships with communities is by supporting goals that are rooted in the local context. Founders of both *596 Acres* and *Transparent Chennai* underscore the importance of using data to serve community goals and concerns. Indeed *596 Acres* and *Transparent Chennai* see data as a means to amplify and validate concerns already existing in the community, and a means to serve community-focused goals. These goals are facilitated even further by supporting the formation of communities of interest around open government data. As in the case of *596 Acres*, if communities form around open government data, they are able to form consensus, share skills, and articulate their demands more clearly, and therefore have a greater potential to exert an agency in their neighborhood. In such cases, the community becomes the primary driver for change in the neighborhood, with the intermediary acting as a facilitator and not vice versa.

5.3 Conclusions

Governments have long collected data. However, emerging developments that automate data creation (such as networked ubiquitous sensing devices) have led to an unprecedented growth in data production, and as a result increased the complexity of data practices needed to analyze, update, and use data. At the same time, discussions on big data and open data, and their transformative potential for diverse fields, including governance, have become increasingly

prominent. A critical discussion on the constructed and deliberate nature of data has emerged

problematizing this growing influence and impact of data. In "Six Provocations for Big Data"

Kate Crawford and Danah Boyd critique the notion of "data as fact."[263] They argue that practices

of data production, dissemination, and processing are almost always prone to interpretations, and

therefore emphasize the need to recognize the interpretive and representative nature of data in

our interactions with it. In *Raw Data is an Oxymoron*, Lisa Gitelman and her contributors

demonstrate that a nuanced understanding of the practices bound up in the production, storage,

and dissemination, can help reveal its implicit biases and assumptions.[264]

These discussions raise important questions about contemporary data usage and practices,

namely: what are the data practices that help reveal the assumptions and biases in data? How

does data become intelligible and useful for communities? Can communities make use of data to

effect change in their civic circumstances? It is critical that we develop ways to thoughtfully

consider these questions, because our current understanding of data usage and practices will have

an impact on its future use. In this thesis, I have argued that data intermediaries and their

practices are critical to answering these questions, and highlighted the need to create a

framework that helps analyze practices of data intermediaries in diverse contexts.

To develop an understanding of current data practices, I have framed the discussion

within the historical context to reveal a tradition of critically engaging with government data that

resonates with contemporary concerns of data usage. In addition, drawing on the work of

researchers who have studied the patterns of use of open government data, and my broader

---

[263] Danah Boyd and Kate Crawford, "Six Provocations for Big Data" (paper presented at Oxford Internet Institute's conference, "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society." Oxford, United Kingdom, September 21, 2011): 1, accessed December 13, 2014, http://ssrn.com/abstract=1926431.

[264] Lisa Gitelman, ed. *Raw data is an oxymoron* (Cambridge: MIT Press, 2013), 70.

research on the data practices of intermediaries, I have proposed an analytical framework that defines the data intermediary function, as constituted by a range of technical, representational, critical, and civic data practices. This framework is distinct from earlier definitions of data intermediaries in its emphasis on their civic and critical functions.

Underscoring the critical function of data intermediaries helps identify practices that reveal biases and assumptions in data. In the discussion of examples in New York City and Chennai, I have shown that discrepancies in data have a material impact on the built environment. Intermediaries are capable of teasing out and addressing these biases through a range of data practices (some involving inputs from communities) that rectify and augment data. In this process, data comes to reflect more closely the realities on the ground. However, data intermediary practices are not free from potential bias, as evidenced by the discussion of the ethical considerations of implementing an SMS-based feedback system in the Chennai example. Hence, it is crucial to evaluate whether data intermediaries help reveal biases in data or introduce new assumptions in data through their data practices.

Highlighting the civic function of data intermediaries helps identify practices that support data usage by communities. At present the wider public has the capacity to only minimally engage with data, and intermediaries are critical in making data intelligible and accessible to communities. In the discussion of the examples in New York City and Chennai, I have shown that intermediaries operate in the civic context, and through their data practices facilitate civic behavior and engagement in communities. When data intermediaries facilitate connections between data and communities, it is crucial to evaluate whether intermediaries emerge out of community needs or are more closely linked to the needs of government and other institutions.

I believe that the analytical framework provides a basis to evaluate these considerations, and helps us develop a detailed understanding of the capacities and limitations of data intermediary practices and their functions. By applying this framework to the examples in New York City and Chennai, I have shown that the framework is useful in identifying data intermediary practices that are effective in different contexts. In addition, I have shown that using the framework to compare data intermediary brings to attention challenges particular to open data contexts as well as challenges such as quality and sustained use of data that are common to different open data contexts. For researchers and practitioners working with open government data, I have proposed three opportunity areas to tackle challenges of quality and sustained use of data that face open data intermediaries today:[265]

1. Designing for Mechanisms of Feedback & Augmentation

2. Designing for Diversity of Uses, Range of Skills, and Replicability of Practices

3. Designing for Community-Focused Goals, and Formation of Communities around Data

At a moment when the practices and research around open government data are still evolving, I believe that these opportunity areas provide a fertile ground for further exploration.

The two cases discussed in this thesis have provided insight into data practices that translate open government data into information that supports communities in having an impact on their built environment. I have tried to demonstrate that while most open data initiatives at the national and city level emphasize top-down approaches for data dissemination, researchers and practitioners can evaluate how open government data can be made more community-focused and user-centric by focusing attention on the practices of intermediaries who work in close

---

[265] For a more detailed discussion on these three opportunity areas consult the previous section.

collaboration with communities. Indeed through the cases I have described, I have tried to illustrate that there is a great opportunity to support the wider public in engaging and interacting with open government data. To tackle challenges related to sustained use of open government data, I believe that in the future it will be important to evaluate whether these interactions enable individuals and communities to make data supported arguments, and whether or not they are able to do so without the facilitation of data intermediaries. There is a great opportunity to account for the opportunities, particularities and challenges of bottom-up data practices assisted by data intermediaries to help shape the future of open data initiatives.

Bibliography

Addams, Jane. *Hull House Maps and Papers*. Boston: Thomas Crowell & Co, 1895.

Banisar, David. "Freedom of Information Around the World: A Global Survey of Access to Government Information Laws." Privacy International, 2006. Accessed April 21, 2015. http://freedominfo.org/documents/global_survey2006.pdf.

Banks, Sarah et al. "Everyday ethics in community-based participatory research." *Contemporary Social Science* 8, no. 3 (2013): 263-277.

Benjamin, Solomon, R. Bhuvaneswari, and P. Rajan. "Bhoomi:'E–governance', or, an anti–politics machine necessary to globalize Bangalore?" Working Paper, Collaborative for the Advancement of Studies in Urbanism through Mixed Media, January 2007. Accessed December 12, 2014. https://casumm.files.wordpress.com/2008/09/bhoomi-e-governance.pdf.

Booth, Charles. *Life and Labour of the People*. London: Williams and Norgate, 1889.

Boyd, Danah and Kate Crawford. "Six Provocations For Big Data." Paper presented at Oxford Internet Institute's conference "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society." Oxford, United Kingdom, September 21, 2011. Accessed December 13, 2014, http://ssrn.com/abstract=1926431.

Canares, Michael et al., "Ground-up Open Data Intermediaries – Who? Where? How?" Session at the Open Knowledge Festival, Berlin, Germany, July 17, 2014. Accessed April 22, 2015. https://pad.okfn.org/p/Ground-up_open_data_intermediaries.

Chattapadhyay, Sumandro. "Opening Government Data through Mediation: Exploring the Roles, Practices and Strategies of Data Intermediary Organisations in India." Open Data Research, 2014. Accessed March 18, 2015. http://ssrn.com/abstract=2549734.

Chui, Michael, et al. "Open Data: Unlocking Innovation and Performance with Liquid Information," McKinsey & Co., 2013. Accessed February 12, 2015. http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information.

Currie, Morgan. "Participation in the City: A Typology of Open Government Data Use." Poster presented at the iConference at University of North Texas, Fort Worth, Texas, February 12-15, 2013. Accessed March 12, 2015. http://hdl.handle.net/2142/42062.

Davies, Tim. "Open Data Barometer: 2013 Global Report." World Wide Web Foundation and Open Data Institute, 2013. Accessed March 13, 2015. http://www. opendataresearch. org/dl/odb2013/Open-Data-Barometer-2013-Global-Report.pdf.

Davies, Tim. "Open Data Barometer Global Report – Second Edition." World Wide Web Foundation and Open Data in Developing Countries, 2015. Accessed April 19, 2015. http://barometer.opendataresearch.org/report/summary/the_barometer.html.

Davies, Tim. "Open Data, Democracy and Public Sector Reform: A Look at Open Government Data use from data. gov. uk." M.Sc diss., University of Oxford, 2010.

Davies, Tim. "Open Data: Infrastructures and ecosystems." Open Data Research Network, 2011. Accessed March 18, 2015. http://ssrn.com/abstract=2549734.

Davies, Tim and Zainab Ashraf Bawa. "The promises and perils of open Government data (OGD)." *The Journal of Community Informatics* 8, no. 2 (2012). Accessed February 12, 2015. http://ci-journal.net/index.php/ciej/article/view/929/926.

Deahl, Erica Sachiyo. "Better the Data You Know: Developing Youth Data Literacy in Schools and Informal Learning Environments." S.M. thesis, Massachusetts Institute of Technology, 2014.

Donovan, Kevin. "Seeing Like a Slum: Towards Open, Deliberative Development." *Georgetown Journal of International Affairs* 13, no. 1 (2012): 97-104.

Feldman, Jonathan. "Asheville's Open Data Journey: Pragmatics, Policy, and Participation." In *Beyond Transparency: Open Data and the Future of Civic Innovation*, edited by Brett Goldstein, 51-62. San Francisco: Code for America Press, 2013.

Gitelman, Lisa, ed. *Raw data is an oxymoron*. Cambridge: MIT Press, 2013.

Goldsmith, Stephen, and Susan Crawford. *The Responsive City: Engaging Communities Through Data-smart Governance*. San Francisco: John Wiley & Sons, 2014.

Goldstein, Brett. "Open Data in Chicago: Game On." In *Beyond Transparency: Open Data and the Future of Civic Innovation*, ed. by Brett Goldstein, 13-26. San Franciso: Code for America Press, 2013.

Gurstein, Michael B. "Open data: Empowering the empowered or effective data use for everyone?" *First Monday* 16, no. 2 (2011). Accessed December 18, 2014. http://firstmonday.org/ojs/index.php/fm/article/view/3316.

Huijboom, Noor, and Tijs Van den Broek. "Open data: An International Comparison of Strategies." *European journal of ePractice* 12, no. 1 (2011): 1-13.

Janssen, Marijn, Yannis Charalabidis, and Anneke Zuiderwijk. "Benefits, adoption barriers and myths of open data and open government." *Information Systems Management* 29, no. 4 (2012): 258-268.

Johnson, Jeffrey Alan. "From open data to information justice." *Ethics and Information Technology* 16, no. 4 (2014): 263-274.

Kitchin, Rob. *The Data Revolution: Big data, Open Data, Data Infrastructures and their Consequences*. London: Sage Publications, 2014.

Lee, Gwanhoo, and Young Hoon Kwak. "An Open Government Maturity Model for Social Media-based Public Engagement." *Government Information Quarterly* 29, no. 4 (2012): 492-503.

Living Lots NYC. "About." Accessed March 23, 2015. http://livinglotsnyc.org/about/about/.

Magalhaes, Gustavo, Catarina Roseira, and Sharon Strover. "Open government data intermediaries: A Terminology Framework." In *Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance*, 330-333. Association for Computing Machinery, 2013.

O'Connor, Shannon. "Methodological Triangulation and the Social Studies of Charles Booth, Jane Addams, and WEB Du Bois." *Sociation Today* 7, no. 1 (2009). Accessed December 13, 2014. http://www.ncsociology.org/sociationtoday/dubois/three.htm.

Pollock, Rufus. "The Economics of Public Sector Information." Working Paper in Economics, Faculty of Economics, University of Cambridge, United Kingdom, 2008. Accessed January 12, 2015. http:rufuspollock.org/papers/economics_of_psi.pdf.

Raman, Nithya. "Collecting data in Chennai City and the limits of openness." *The Journal of Community Informatics* 8, no. 2 (2012). Accessed January 18, 2015. http://ci-journal.net/index.php/ciej/article/view/877/908.

Ribes, David, and Steven J. Jackson. "Data bite man: The work of sustaining a long–term study." In *Raw Data is An Oxymoron*, edited by Lisa Gitelman, 147-166. Cambridge: MIT Press, 2013.

Sawicki, David S., and William J. Craig. "The Democratization of Data: Bridging the Gap for Community Groups." *Journal of the American Planning Association* 62, no. 4 (1996): 512-523.

Shekhar, Satyarupa and Vinaya Padmanabhan, "Still Open Ended." Research poster, Transparent Chennai, Chennai, 2014. Accessed March 18, 2015. http://www.opendataresearch.org/project/2013/tc.

Stanley, Matthew. "Where Is That Moon, Anyway? The Problem of Interpreting Historical Solar Eclipse Observations." In *Raw Data is an Oxymoron*, edited by Lisa Gitelman, 77-88. Cambridge: MIT Press, 2013.

Tauberer, Joshua, *Open Government Data: The Book*. Joshua Tauberer, 2014. Accessed April 24, 2015. https://opengovdata.io/2014/open-government-big-data-mediators/.

Tauberer, Joshua. "Visualizing Metro Ridership." In *Open Government Data: The Book*. Accessed April 26, 2015. https://opengovdata.io/2014/visualizing-metro-ridership/.

Tinholt, Dinand. "The Open Data Economy: Unlocking Economic Value by Opening Government and Public Data." Capgemini, 2013. Accessed May 01, 2015. http://ebooks.capgemini-consulting.com/The-Open-Data-Economy/.

Townsend, Anthony M. *Smart cities: Big data, Civic hackers, and the Quest for a New Utopia.* New York: WW Norton & Company, 2013.

Transparent Chennai. "About." Accessed March 16, 2015. http://www.transparentchennai.com/about/.

Transparent Chennai, "Public Toilets in Chennai," Research Report, Chennai, India, 2011. Accessed March 18, 2015.

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ve
d=0CB4QFjAA&url=http%3A%2F%2Fwww.transparentchennai.com%2Fwp-
content%2Fuploads%2F2011%2F05%2FPublic-Toilets-Transparent-Chennai-Issue-
Brief.pdf&ei=XI5MVeGyC9e5oQTpjoCoDQ&usg=AFQjCNHSwxsjPmvASc2kmg3CFDGIDC
DWww&bvm=bv.92765956,d.aWw.

Ubaldi, Barbara. "Open Government Data: Towards Empirical Analysis of Open Government
Data Initiatives." Working Paper on Public Governance, No. 22, Organisation for Economic Co-
operation and Development, 2013.