Ready for Data Analytics? Data Collection and Creation in Local Governments

Minyoung Ku
John Jay College of Criminal Justice, City University of New York
524 West 59th Street
New York, NY 10019
mku@jjay.cuny.edu

J. Ramon Gil-Garcia
University at Albany, State University of New York
Universidad de las Americas Puebla
187 Wolf Road, Suite 301
Albany, NY 12205
jgil-garcia@ctg.albany.edu

ABSTRACT

A good understanding of local government data, particularly the contexts in which the data are collected and created, is essential in order for data guardians and users to effectively manage that data and to draw accurate and rich information from them. However, our knowledge of both local government data and data contexts is very limited, which poses a wide range of challenges to data initiatives inside and outside local governments, from open data to data analytics. This paper explores the mechanisms of data generation and the determinants of data contexts in local governments. To do so, we conducted an in-depth case study of data collection and creation in a city government in New York State in the U.S., focusing on administrative data. We found that local government data are collected or created in three ways: original raw data collection, original raw data creation, and second or higher-order data creation. Both the internal and external environments of local governments add complexity to data management by influencing who, what, where, when and how to collect and create data for administrative purposes through these processes.

CCS CONCEPTS

• **CCS → Social and professional topics → Professional topics → Management of computing and information systems → System management** → Quality assurance

KEYWORDS

Data management, data lifecycle, data generation, data context, local government, case study

dg.o '18, May 30-June 1, 2018, Delft, Netherlands

ACM ISBN 978-1-4503-6526-0/18/05…$15.00
https://doi.org/10.1145/3209281.3209381

ACM Reference format:
Ku, M. and Gil-Garcia, J. R. 2018. Ready for Data Analytics? Data Collection and Creation in Local Governments, In *Proceedings of the 19th Annual International Conference on Digital Government Research*, May 30 – June 1, Delft, the Netherlands

1 INTRODUCTION

Advanced computing technologies and high-bandwidth networks have dramatically increased the data that governments at all levels generate. In particular, the advancement of e-government has accelerated the creation and accumulation of digital data in local governments for the past four decades. It is now common to see local government employees use computers for diverse and important aspects of their work, including online processing of forms and recording public meetings, hearings, workshops, and conferences to provide public access to video clips and/or audio files. More recently, the penetration of mobile phones and tablets and the evolution of machine-to-machine technology, including sensors and Radio Frequency Identification (RFID), have brought about enormous change in the size, diversity, and speed of data gathered on a daily basis at local levels.

As local government digital data grow exponentially, their potential and actual benefits increase. Opening government data can facilitate transparency and strengthen citizen engagement in legislative processes and governance [1-3]. When shared with the private sector, government data can fuel economic growth by helping industries open more doors to innovation, transforming the way companies do business [2]. In addition, the use of local government data within the public sector, with the aid of intelligent technologies such as artificial intelligence and machine learning, is expected to alter governance structures and operations by personalizing public services and tailoring government interventions in markets and civil societies based on rich and timely information [4].

However, capitalizing on government data inside and outside local governments is challenging. Prior studies have identified some of the obstacles to releasing and leveraging government data, which include privacy and legal concerns, public employees' fear of data misinterpretation [2], lack of information about the data such as their location and

ownership [2], and data quality issues, such as inaccurate, incomplete, or outdated data [5-6]. Many of these challenges arise in the very early stage of government data management—data collection and creation—and result in cascading problems along the data lifecycle, including problems with data release and sharing, data processing, data analysis, and data archives. These problems can consequently harm the usability and usefulness of government data and ultimately prevent the achievement of benefits from government data release and mining [5]. Nonetheless, little is known about how and in which contexts government data are collected and created.

Although most government data that are directly related to communities and citizens are produced at local levels, data collection and creation in local governments have received little attention by both scholars and practitioners. In particular, information about data contexts, which are mostly determined when data are collected or created, is essential for data users to decide how to analyze and interpret data, as well as whether the data can be used to achieve a specific goal. Over the past decade, the absence or insufficiency of such information has been consistently acknowledged as one of the biggest barriers to effective management and use of government data in local governments. However, our knowledge of how data contexts are determined is very limited. Therefore, this paper explores the collection and creation of local government data, and data contexts. To do so, we seek answers to the following questions: *How do local governments generate data? What determines data contexts by influencing these data generation processes?*

2 LITERATURE REVIEW

2.1 Defining Data

2.1.1 Definition of Data. Data has been used interchangeably with information in the literature from many disciplines and fields, including information science, information systems, communication, organization science, and e-government. However, the two are not identical, but rather distinct concepts. Data are symbols with no value, while information is processed data that delivers contextually relevant content that can answer questions [7]. According to Ackoff, both data and information represent the properties of objects and events, but data do not have value until they are processed so that people can infer meaning from them. Based on these definitions, Ackoff suggests that there is a hierarchal relationship between data and information in that information is derived from data. Similarly, Zeleny distinguishes information from data by the interpretive meaning conveyed via information [8] and Nonaka defines information as "a flow of messages" [9] (p. 15). Although there have been controversies over the hierarchical relationship between data and information, previous studies have consistently revealed distinctions between data and information [10] and programmability and algorithmicity [11].

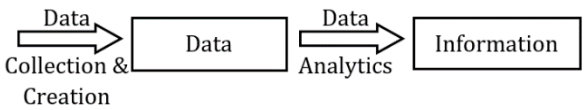

Figure 1: Transforming data into information

2.1.2 Data Format. Data are multifaceted and previous studies have suggested different classifications of data, depending on which dimension is highlighted. Data in organizations can be digital or non-digital. In human history, data have existed in the form of letters, symbols, characters, and numbers that are not digitized but displayed and stored on paper or conveyed through verbal or behavioral communications between individuals. It has been less than half a century since the creation of digital data and the digitization of legacy data began, attracting the attention of scholars and practitioners [12].

2.1.3 Data Type. As digital data explodes, the type of data collected and created by organizations has become enormously diverse. Recent literature in the fields of information systems and computer and information sciences has proposed three types of data by the degree to which the data are organized: structured, semi-structured, and unstructured data [13-16]. Structured data have clearly defined fields so that they can be entered into an array of rows and columns, such as data entries in spreadsheets and relational databases [12-15]. On the other hand, unstructured data are all formats of data whose fields are not defined, such as text documents, pictures, videos, and audio recordings [13-14]. Semi-structured data fall between the two extreme types, representing loosely formatted data with fields and hierarchies that are not associated with data tables, such as XML documents [14] [16].

The materials now classified as data, such as text documents, images, and voices, have traditionally been considered the display of information rather than its source. Our understanding of data is still deeply rooted in this approach. For example, specifying methods for representing data and information, Zins argues that data contain symbols, letters, characters, and numbers, whereas information takes the form of texts, figures, and other forms of communication that are meaningful and help people better understand a subject [17]. However, recent advances in technology allow researchers and practitioners to tap into a greater volume of digital data with greater speed [18].

2.1.4 Data Source. As data sources have become more diversified, from human agents to computer operating systems and satellites, scholars and practitioners have recently attempted to classify data by its source. Monash classified digital data into three categories: human-generated data, machine-generated data, and hybrid data. Human generated data (HGD) are data created as "the direct result of human choices" [19]. Credit card transaction records, doctors' prescriptions, tweets, emails, government reports, and policy memoranda are all examples of HGD. Meanwhile, data can be

produced automatically by mechanic and electronic devices with minimal or no human intervention, known as machine generated data (MGD) [19]. MGD includes log files of computer operating systems, sensor readings from factories, RFID chip readings, and global positioning system (GPS) and global navigation satellite system (GNSS) outputs [19]. Both HGD and MGD can be records about human beings. However, while the former records human choices, the latter records observations of humans [19]. Hybrid data have characteristics of both HGD and MDG. Monash illustrated the hybrid type using web logs; a web log often consists of commentaries and graphics that an individual(s) creates, as well as data components from operating web servers and networks [19].

2.2 Data Lifecycle

Data, like other resources, are utilized throughout different stages of development in organizations [20-23]. Jagadish and his colleagues argued that understanding the data pipelines from generation to deletion helps data users have a better understanding of the data assets available to them. By doing so, it enables them to better handle uncertainty and error in data and to grasp the opportunities that data from diverse sources bring [18].

The stages of the data lifecycle identified in prior research vary; even for the same stage, different terms are used. For example, in the case of data generation, some studies use the term data acquisition [18] [20], while others call it data creation, data collection, data gathering, or data capture [21-22]. However, the models for data lifecycle management suggested in the literature commonly include the following five steps in sequential order.

- Data generation: data collection and creation
- Data cleaning and processing
- Data analysis and interpretation
- Data preservation
- Data re-use or deletion

Data starts its lifecycle by being collected through observation and measurement or by being created through creativity and intelligence [18-20]. More specifically, data collection is to gather or measure attributes of objects or events to understand a phenomenon of interest. Examples of data collection are survey questionnaires or interviews; recording and observing human behavior, traffic conditions, and weather through closed-circuit televisions (CCTV); and measuring sea levels through sensors. Data can also be created through creative and intelligent activities as the records of data creators' expression of judgement, knowledge, logic, and emotions. Taking pictures, writing doctors' notes in word documents, recording songs in audio files, and writing emails can be classified into this category of data generation. Jagadish and his colleagues pointed out that data collection and creation do not always involve human beings' senses, but can frequently be conducted by machines whose primary functions may or may not be recording or creating data [18]. As a result, data generated by machines are common today [18] [22-23]. Monash argues that the growth of machine-generated data has surpassed data recorded by human beings and that this trend will intensify as computing, networking, and intelligence technologies advance [18-20].

To extract meaningful messages (i.e., information) from the data collected or created, data need to be analyzed and interpreted to produce value [24]. Since data are not always error-free or machine readable, data cleaning and processing are often performed before analyzing data [18]. Existing studies point out that since data errors frequently occur and most, if not all, data properties are determined during data collection and creation, the speed and accuracy of data processing is closely linked to the conditions of data collection and creation processes in an information system [18] [22]. Thus, properly managing data from a very early stage of the data lifecycle and building a successful meta-data catalog for data users, which can provide enough information about the contexts of data collection and creation, are critical to reducing the cost of data cleaning and processing, as well as to extracting accurate and meaningful information from the data [18] [22].

2.3 Local Governments as Data Contexts

2.3.1 Six Dimensions of Data Contexts. Data do not exist in a vacuum. Rather, data are embedded in contexts. Data users' knowledge of both the data itself and its contexts plays an important role in obtaining relevant, accurate, and rich information through data analytics [25-26]. From a data management perspective, data contexts refer to data about data (or *meta-data*) that describe *who, how, what, when, where, and why* [27-28]. The identity of a data producer—*who* collected or created the data—can become an important indicator of data credibility for data users [29]. In addition, such information also helps data guardians who manage data collected or created by others and allows data users to make inquiries about the data and direct data requests to the appropriate parties. The *how* refers to the tools and methods used during data collection or creation. Data collection and creation methodologies determine the properties of the data. Fully understanding these properties is necessary for selecting appropriate tools and methods to process, store, and analyze the data. More importantly, data collection methods and instruments are directly linked to the reliability (when a method or a tool returns the same result with repeated tests) and validity (when a method or a tool records what it is intended to record) of the data and in turn the accuracy, precision, and trustworthiness of findings from data analysis.

Understanding *what* the data are about is the first step in managing and using data. In particular, data users need this information to figure out what questions they can or cannot answer by using the data. *Where* the data are collected and created is as important as what to collect or create, since it establishes the purpose and scope of data analytics. For example, data collected or created in one county may not be applicable to another county. The location of data collection or creation may limit the generalizability of the findings from data analysis [29]. Further, having information about the timeline for data collection or creation (the *when*) is also necessary for determining whether the data fit a particular purpose. Things

change over time and outdated information about an object or an event may not be useful in answering time-sensitive questions [29].

All the dimensions of data contexts discussed above—who, how, what, where, when—are designed to fill a purpose. Knowing the reason *why* the data were collected or created can guide data users along the data lifecycle to determine whether the data are appropriate or sufficient for answering a question, to choose the right tools and methods to process and analyze the data, and ultimately to draw accurate and useful information from the data by embracing the strengths and limitations of the data.

Since the 1980s, scholars and practitioners in the fields of computer science, engineering, and information science have pursued efforts to more effectively manage data and provide users with more complete information about those data. Most of these scholars and practitioners have viewed meta-data issues as technological problems and their research has been dedicated to building successful meta-data schemas and computer repositories. However, the questions of meta-data—who, how, what, when, where, and why to collect or create data—are socially constructed. While answers to those questions can and should be displayed technologically in an organized way to data guardians and users, they are fundamentally human questions. To develop and improve a meta-data system, it is necessary to understand the dynamics in the formation of data contexts: where and how the contents of meta-data are formed.

2.3.2 Data in Local Governments. Data generation is one of the important functions of governments. As the first-tier administrative division of the states in the U.S., local governments are a major source of data that are directly related to citizens' daily lives in communities [30]. The majority of data produced by or residing in local government agencies are administrative in nature. Administrative data are the data produced for organizational and managerial purposes [30]. In particular, administrative data in local governments are collected or created for the purposes of preserving records and supporting the functions and operations of government agencies, including the provision of public services, policy making and implementation, and organization management at the local level.

Unlike research data, the production of administrative data tends to take place as a byproduct of governmental entities' diverse activities [31]. Research data are collected or created through scientifically and systematically designed and purposefully selected methods and tools in settings that are controlled by data collectors or creators to some extent, in order to answer research questions. However, much of administrative data in local governments, such as registration data, payroll data, and sensor data, are not intended to answer preset questions. In addition, they tend to be collected or created in much less controlled settings than data produced in laboratories. Thus, analyzing and interpreting such data for purposes that were not considered in the early stages of data collection and creation can be challenging. Information about data structure, content, and context is critical to extracting relevant and accurate information from administrative data in local governments.

3 RESEARCH DESIGN AND METHOD

To explore the mechanisms of data generation and the determinants of data contexts in local governments, we conducted a case study. Case studies allow researchers to explore contemporary real-life phenomenon with detailed contextual information [32] (p. 2). The method is especially useful for conducting empirical inquiries about phenomena that are complex and not well understood, providing rich information about them [33] (p. 23). The generation of government data at the local level is deeply embedded in the functions of local governments, which are highly complex. Data generation in local governments has rarely been studied to date. Therefore, an in-depth case study was chosen as the most appropriate approach for the study.

This study relies on qualitative data collected from 32 public employees at multiple levels of local government, including top- and middle-level managers and street-level staff in an anonymous city in New York State's capital region in the U.S. We conducted focus groups from July 2015 to September 2015 and interviews between August 2015 and November 2015. The departments in which the participants were affiliated were accounts and disbursements, assessment, buildings and code, engineering, facilities, finance, fire, human resources, information technology, law, parks and recreation, planning and development, police, purchasing, public works, utilities, and water and sewer departments. The positions of the participants vary, including building and housing inspectors, code enforcement officers, police officers, firefighters, paramedics, assessors, city attorneys, city clerks, treasurers, city engineers, and information technology staff. However, it should be noted that this case study is part of a project that looks at code enforcement data management and data sharing among municipalities in Upstate New York. Thus, our analysis may rely more on data collected in the departments whose duties are closely related to property code enforcement—the departments of buildings and code, public works, planning and development, fire, and police—than in other departments.

4 CASE: CODE ENFORCEMENT DATA IN A U.S. LOCAL GOVERNMENT

We explore data generation and data context in Turian (a pseudonym), a city government in Upstate New York in the U.S., which participated in a government-university partnership to build a code-related data sharing system with neighboring cities between 2015 and 2017. The municipality is a medium-size city, whose population was approximately 65,000 as of 2013. As of 2015, the number of city employees was just under

600 and the city government consisted of 19 departments, including the mayor's office.

Turian has experienced economic declines over the past decade since the onset of the global financial crisis in 2007. One of the effects of the economic downturn is an increase in vacant and abandoned properties in the city. In 2015, the number of non-seasonal vacant properties in the city reached more than 1,000 in 20,379 parcels. In the city government, the concern was raised that vacant and abandoned properties may increase crimes, such as litter, assault, arsenal, prostitution, vandalism, and drug dealing, and decrease property values not only in the distressed area, but also in the neighboring areas. Thus, a team of 18 inspectors, administrators, and support staff members affiliated with three different departments (Buildings and Code, Assessment, and Law) started closely working together to meet the growing demand for code enforcement services, ranging from building and housing inspection to demolition of foreclosed properties in Turian. Later, police, fire, and planning and development also joined the efforts to tackle urban blight in the city by sharing and analyzing data collaboratively from a broader perspective. However, the team encountered obstacles and challenges that slowed down the initiative and found that the most crucial problems they faced lay in data.

Code enforcement tasks are by nature evidence-based. Code enforcement officials must examine the physical and legal conditions of properties to make code-related decisions. The property code-related data, such as property owner records, property addresses and conditions, and the history and status of permits and violations, were necessary not only for the team's day-to-day administrative decision-making. Such data was also critical to policy decision-making for public safety, sustainable urban planning, land-use control, and taxation. However, the data were scattered in multiple offices and infrequently shared between departments even in the city government. Since many of the datasets were also not updated on a regular basis, code enforcement officials frequently had to make decisions on problematic properties based on inaccurate information extracted from the outdated data with many missing data entries. Consequently, the poor quality of data hindered efficient code enforcement and timely responses to time-sensitive complaints in the city. It also considerably delayed foreclosures on vacant and abandoned properties, and sales and demolitions of those foreclosed properties. On top of that, the inaccurate information about properties in the local area, such as floorplans and physical conditions, put firefighters and police officers in danger when emergencies occurred inside the homes and buildings. Therefore, the city mayor launched the collaborative initiative to improve the city's information systems by enabling effective data management inside the organization and effective data sharing across cities in the same region, partnering with a university and neighboring cities.

5 ANALYSIS OF THE CASE

5.1 Data in Local Governments: How are the Data Generated?

The Turian city government generated both digital and non-digital data through administrative activities. The digital data they generated exist in various forms, ranging from structured, such as tables in spreadsheets, to unstructured data, such as text documents and emails that the city employees produce, pictures of the exterior and interior conditions of abandoned properties taken by code enforcement officials, and video recordings captured by in-car and body-worn cameras that police officials use. The digital and non-digital data were generated through three major mechanisms: original raw data collection, original raw data creation, and second or higher-order data creation. First, the city government collected data from internal and external sources by measuring, observing, and recording properties of an object or an event. This data collection was conducted through the activities of the labors hired by the city government and/or by city-owned or -managed machines; for example, traffic sensors and closed-circuit televisions (CCTV) automatically collected and transferred data from humans, artificial structures, and natures.

Second, original raw data creation takes place through creative and intelligent activities of individuals hired by the city government, whether as individuals or in a group. Fire and police reports, policy memoranda, budget plans and survey data in spreadsheets, and webpages on the city portals are examples of original raw data created in the city government.

Third, the city government also created new data by processing, integrating, analyzing, and interpreting existing data collected or created within and outside the organization through creative and intelligent activities of individuals and groups or using computer programs that were purchased or leased by the city government. Data sharing across departmental or organizational boundaries occurred much less frequently than it did inside the departments. However, data that were shared not only from internal, but also from external sources, were used as raw materials to create second or higher-order data. Data from external sources were an important source of data creation when the data users did not have the legal authority to get access to, or enough resources to collect data directly from, the object or event of interest. For example, the police department is legally authorized to collect, but not to store, fingerprints of all visitors at the police department building entrance; the scanned fingerprints are stored on servers managed at the federal level. The fingerprint data (i.e., first-order data) were shared with the local government upon request, and police officials in the city government were able to use the data to write investigation reports and crime statements (i.e., second-order data).

5.2 What Determine Data Contexts in Local Governments?

5.2.1 The Division of Labor: Departmentalization and Job Description. The division of labor, which is the way specialized jobs are grouped, determined who, what, where, and when to collect and create the local government data in Turian. The division of labor by specialty in the city government dictated the data content that members in each department should generate, manage, and use. Thus, depending on the role of a department, the data that the departmental members gather or create about the same object or event vary, and the locations and times of their data collection or creation may be different. For example, the public safety officials in the fire and police departments who we interviewed reported that they collected empirical data, like numbers and pictures about incidents at properties in the city and created reports on the events, but that the data they collected from the sites and put in the reports were not always the same. A paramedic in the fire department who we interviewed stated below:

"When a serious incident involving injuries happens, we paramedics are dispatched to the site. Of course, policemen are there too. But, our data jobs are different. On the spot, policemen usually collect the information about the incident itself, such as detailed descriptions of the event, victims and suspects involved in the event, and actions taken. We paramedics record patients' health status, such as how they breathe, if they have any disability, if so what disability they have, and how much they are traumatized. We also record the patients' medical conditions, and medical treatments we performed to them. The records that we make and keep are more about health conditions of humans involved in that event, rather than the event. We collect such information [data] from patients mostly on the site. But it could be done after we left the site, like while moving patients to a hospital."

In a department, depending on the description of each job, the types of data collected and created by the departmental members vary. Job descriptions, which are the roles and responsibilities of each position as officially described, can also determine the sequence of tasks performed among different positions within and across departments. This, in turn, decided where the first-order data should go and who would create second-order data by using the original data. For example, building and housing inspectors in the buildings and code department were the collectors of data about the interior and exterior conditions of properties in the city. By using the data, they created new data about code violations of the property owners. However, it was code department secretaries who decided who would collect and create the data and when by deciding, and creating data about, the inspectors' schedules. Once the property-related data was gathered by the inspectors, the staff in the planning and development department created tables, graphs, and charts and wrote documents to establish a policy to design and construct the city's facilities to be green and sustainable and to evaluate the plan.

5.2.2 Forms and Work Processes. The Turian city government collects and creates data in various ways, including conducting surveys, writing documents, taking pictures, and recording videos. The data collection method most frequently mentioned by the participants in both the focus groups and the interviews is forms. The forms that the city government agencies use are formatted with the information about who, what, where, and when to collect data. In particular, the items in a form determine the fields of data. The design of the form can considerably determine the content and properties of the data collected by using it. For example, code enforcement officials collect property-related data, including physical condition, ownership, and code violation of properties, mostly by using forms. Building registration forms, field checklists, and code violation notice forms are three important examples. Due to differences in the design of the forms, data entries that contain the same information are collected in different formats; the condition of a property is identified as good, fair, or bad in a form, while another form requires inspectors to check fail or pass and then write a narrative description in a box. Similarly, the prehospital care report form decides when, where, and what data paramedics should collect from patients and about an incident. This form requires them to collect data such as incident year, month, date; dispatched paramedic's code and arrival time; patient's name, address, date of birth, medication, allergies; the status of patient's breathing, skin color, skin temperature, cap refill, and pupils; and responsible party's name and phone number. The report form requires paramedics to collect the data by selecting one or more of the prelisted options, putting numbers, and writing words or narratives.

Work processes play a critical role in both data collection and creation in the city government, deciding who, what, where, when, and how to create and collect. The processes for code-related services, including permits and licenses, are not electronic but paper-based. The building and code department accepts paper-based construction permit applications and related documents; not all of the applications then become digital data. The construction permit review committee in the city government meets every other week and decides on approvals. Once the review process is complete, only data from the approved applications are manually entered and saved in the city's electronic database by an administrative staff member in the department. The applications that are declined and their related paper documents are sent to a cabinet in the office for future use.

The inspection appointment process in the department also indicates who, what, when, and where to collect data about building complaints and site visit appointments. The building and housing inspector who we interviewed stated that "both administrative staff in our department and I receive incoming calls about building and housing complaints and fill appointment sheets. However, it's me, not the department secretary, who collects information [data] about how the

complaints are resolved. Once we receive a complaint about a building or a house, the secretary schedules my site visit. Then I bring the appointment sheet to the site, since it must be signed by a landlord or an agent who is responsible for managing the property right after an inspection, regardless of the inspection results. I bring it back to the office, and then the secretary enters the data in the appointment sheet into the city government's information system."

5.2.3 Employees' Perceptions and Experience. Forms and Work Processes. Data whose collection or creation entails human decision-making tend to be affected by employees' perceptions and experiences. Unless data collection or creation is completely automated by sensors or other machines, in particular, members' ideas, knowledge, wisdom, and personal experiences can be reflected in the processes and products of the data collection or creation in the city government. For example, building inspectors collect data about properties in the city by filling in inspection forms. When using the forms, they make decisions about the conditions of properties and either indicate pass or fail or rate them as good, fair, or bad. Although they have inspection standards, not all of the forms clearly list the pass/fail or good/fair/bad standards in detail. The inspectors do not use any tools that give them quantified information about the conditions of ceilings, walls, and floors of a property. Rather, they make their decisions by examining a property with their senses and by using knowledge and wisdom gained from their experiences as professional inspectors.

The building and housing inspectors also write inspection reports after their site visits. There is a building inspection report form which is itemized by category of site work and certification necessary for maintaining the building properly. In the report, inspectors diagnose the conditions of the building by section, including structure, electricity, heating, air conditioning, ventilation, pluming, roofing, interior, exterior, insulation, and life safety fire protection; they then make a final judgement on the building's grade and provide comments about their assessments. In this process, building inspectors' understanding and experience considerably affect their decision-making about what to write about the properties in the form, whether or not a check-up on a property is necessary, and what the timeline of the report should be.

5.2.4 Leadership. Data creation and collection are also the product of leadership in the city government at both the department and organization levels. Organizational-level leadership tends to make the decisions that can influence what data is created or collected by what office and/or department, while departmental-level leadership tends to directly influence the decisions about who, what, when, where, and how to create and collect data in the office and/or department. For example, to reduce urban blight and prevent it from spreading by reducing unseasonal vacant and abandoned houses and buildings, the mayor of Turian initiated a property analysis program, called "ReNew" (a pseudonym), which is a coordinated effort of the buildings and code and police departments. A group of code and police officers patrol the streets together, checking the external conditions of properties. If they find a problematic property, they issue and post a warning notification on the front door of the property, which informs the owner(s) that they should fix the problem within 30 days. If the problem is not fixed within that time period, a violation notice is issued. This program has created a new dataset, which contains information about high-risk, potential code violators as well as the physical and legal conditions of their properties, which were not collected or created before. While the mayor officially initiated the program, it was the chief inspector and the chief police officer who laid the plan out and administered the program in their departments.

5.2.5 Information Technologies. Information technologies inside and outside the city government heavily affect data collection and creation, particularly what, where, and how to collect and create, and where and how to save the data collected or created. The Turian city government has their own information system, called "RUNIS" (a pseudonym), which is internally managed and maintained. RUNIS has brought changes in data management practices to the city government. Specifically, it has become the central system where data are created, collected, and saved in the city government.

Due to its very limited functions, however, RUNIS is considered as a database, rather than a data management system in which the members can look up information and run programs to visualize and analyze data. Since RUNIS was initially developed for school districts primarily to store data, the system has very limited functions and cannot satisfy even the basic needs for data management in all departments of the city government. Most of the departments have adopted different information management systems within their offices, which are developed by vendors in the private sector and are customized to their needs. Therefore, the departments use the outsourced information systems to get their work done on a daily basis by entering, saving, retrieving, and analyzing data. They access RUNIS only when they are required to upload or update master data. The respondents we interviewed commented that they considered the government information system merely as a record keeper. As a result, aside from the information systems required by the federal or state government, such as the Real Property System, the members use six information systems for different purposes, and some of the systems direct and store all the data entries from the city government to the vendors' databases.

The development of technology outside the organization has also changed the way the city collects and creates data. Employees in the Turian city government, including code, police, and fire officers, use tablets and mobile phones to perform their daily tasks. Housing, building, and plumbing inspectors enter all the data generated during property inspections by using tablets and the electronic inspection forms loaded onto the devices. They also communicate with

administrative staff members in their office and in other departments by using smartphones that can connect to the internal email server. However, because the tablets can only connect to the internet through Wi-Fi, in most cases the data are collected at sites and saved temporarily in the devices, then moved to the city servers back at the office.

5.2.6 Lawes and Policies. Laws and policies can affect the generation of original raw data from diverse sources by the city government in two ways: (1) providing a legal mandate and the authority to perform data collection or creation and (2) defining the specific conditions of data collection or creation that a local government should satisfy, such as specifying an object or event of interest and defining data fields and structures. The Turian city government, as any of the 1,607 local governments in New York State, is subject to both federal and state laws [46]. The New York State government recognizes the city as a legal entity "with its own governing and taxing authority" [46]. Thus, not only does the city government have the duty to provide public services to citizens in Turian, they are also granted the authority to collect data from and create data about the citizens and communities necessary for the provision and improvement of public services.

More specifically, the city government data activities, including data collection and creation, are legally bounded. For example, Title 19 New York Codes, Rules, and Regulations Part 1203 requires local governments that are charged with administration and enforcement of the Uniform Code to collect particular data fields in order to make decisions on building permits, including a description of the proposed work; tax map number and street address; and the occupancy classification of any affected building or structure. Such legal regulations determine what forms the city government should create to enforce codes, what items should be included in the forms, when the forms are used to collect the data items, and by whom the forms should be handled in local governments in the state.

The participants in the focus groups and the interviews said that the Freedom of Information Act and New York State's Open NY initiative are the respective legal and policy contexts that influence the city's data collection and creation strategies and practices. The Freedom of Information Act is a law that gives citizens in New York State the right to access information from local agencies. Open NY is the policy that New York State Governor Andrew Cuomo initiated to widen the accessibility and usability of the data generated by all government agencies in the state. One of the respondents mentioned that "it is true that we care more about the quality of information [data] we collect and create in our office because of the open data policy. For the open government policy also requires us to display our data in specific formats, for example, in Excel spreadsheet, Jason, and HTLM, we organize our data in those formats too. … Even though there is a three-year or five-year rule for recordkeeping in New York State, but we do not throw out almost any of our data since we don't know what records the citizens request to provide."

5.2.7 Events Happening. Data collection and creation by the city government are both proactive and responsive. The documents on the city planning strategies, pictures of abandoned and vacant properties in a testing zone, and statistical datasets created during the preparation and enactment of the urban blight prevention program, ReNew, are examples of proactive data collection and creation. On the other hand, data from external sources tend to be gathered responsively and such data collection is highly affected by events happening outside the organization. A respondent from the buildings and code department said that "the city's code enforcement system is very responsive. Our data collection begins mostly when citizens need it. The inspectors in our department receive inspection assignments for the day from the chief code inspector every day. The list of inspection assignments is totally up to the calls, voicemails, and emails our office receives from property owners and agents who have complaints. We inspect the overall conditions of a property at a site, but what to examine depends on what complaints we received from them."

This means that although the data fields that the inspectors collect and enter into the information system are predetermined by the forms, when and where the data collection is conducted and what data are recorded, particularly in boxes for narrative descriptions, are influenced by events happening at the properties in the city. The chief code inspector commented in an interview that the kinds of data the department collects from citizens are affected by the economic situation of the city. Data related to foreclosures and code violations tends to be collected more during economic downturns. In contrast, during an economic boom, new data entries through the foreclosure forms and violation reports decrease, while data entries through building permit and business licenses applications tend to increase.

Table 1: Determinants of Five Dimensions of Data Contexts in Local Governments

		DJ	FW	EPE	LD	IT	LP	EH
Dimension of data	Who	V	V		V		V	
	What	V	V	V	V	V	V	V
	When	V	V	V	V		V	V
	Where	V	V	V	V	V	V	V
	How		V	V	V	V		

Note: DJ: Departmentalization and job description; FW: Forms and work processes; EPE: Employees' perceptions and experiences; LD: Leadership; IT: Information technologies LP: Laws and policies; EH: Events happening

6 DISCUSSIONS AND CONCLUSIONS

The case study demonstrates that local government data are generated in diverse forms, ranging from tables in spreadsheets to sensor outputs to text documents, images,

videos; they are collected from internal or external sources or created through the creative and intelligent activities of organizational members. The findings of the study highlight that much data in local governments are *socially constructed* rather than mechanically produced. Data contexts in local governments, which describe who, what, when, where, and how data are created and collected, are affected not only by information technology, but also by the division of labor, including departmentalization and job descriptions, forms and work processes, employees' perceptions and experiences, leadership, laws and policies, and events happening outside the organization.

Such findings theoretically suggest that the socio-technological perspective, which was originally developed to explain the adoption of information technology in organizations, might also be useful in understanding data and their management lifecycle in local governments. More specifically, the 11 determinants of local government data contexts identified in the study can be classified into four categories: *organizational elements*: the division of labor, forms and work processes, employees' perceptions and experiences, and organization information systems; *technological elements*: technological hardware, software, and applications available in the market; *legal and political elements*: laws and policies; and *externally contingent elements*: external events. This means that decisions about what to collect or create and when, where, and how to collect or create the data in local government agencies are not determined or influenced solely by the technologies available, but also by social elements in organizational, legal, and political contexts and even contingent events. Moreover, the technological and social elements that influence the collection and creation of local government data reside both in and outside the local government. Thus, figuring out the contexts of local government and managing and leveraging data assets in local government agencies may be more complex than we think.

From a practical perspective, our findings suggest that the operation of organizational, legal and political, and externally contingent elements, coupled with the complex deployment of information technology resources, may contribute to the failures of data initiatives in the public sector. Much of the government data created in Turian are collected and created manually by the employees. In addition, the collection and storage of the data from electronic devices, such as sensors, office computers, tablets, mobile phones, CCTVs, and in-car and body-worn camera are restricted by laws and policies and influenced by organizational dynamics. Multiple forms ask the same questions in different ways or are formatted differently to collect the same information. Much of the data automated by machines is stored and managed in information systems that reside separately in each department as silos and do not communicate with one another. These problems exacerbate the issues of duplicate and inconsistent data in the city government.

Furthermore, despite the wide adoption of cutting-edge technologies, a large portion of the data collected and created in the local government are non-digital. The paper-based data collection system still widely used in the city government hinders real-time data collection and update and, in turn, the improvement of government information systems from record keeping systems to intelligent decision-support systems that enable seamless data sharing, integration, and analytics. This may explain why many attempts to reinvent government information systems by adopting cutting-edge technologies have not been successful to date.

Despite these contributions, however, this case study has some limitations. This study is a pilot based on a case from a mid-size city government in New York State in the U.S. Moreover, as noted earlier, most of the data used in the study were collected from the city employees who originally participated in a project initiated to tackle urban blight by reinventing code-related information management. Therefore, to generalize the results of the study, we encourage future research to examine data collection and creation in local governments of different sizes, in different regions and countries, and with data from more diverse departments.

REFERENCES

[1] Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government information quarterly*, *27*(3), 264-271.

[2] Conradie, P., & Choenni, S. (2014). On the barriers for local government releasing open data. *Government Information Quarterly*, *31*, S10-S17.

[3] Kassen, M. (2013). A promising phenomenon of open data: A case study of the Chicago open data project. *Government Information Quarterly*, *30*(4), 508-513.

[4] White House (2014). *Big data: Seizing opportunities, preserving values*. Executive Office of the President.

[5] Dawes, S. S. (2010). Stewardship and usefulness: Policy principles for information-based transparency. *Government Information Quarterly*, *27*(4), 377-383.

[6] Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, *29*(4), 258-268.

[7] Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*(1), 3-9.

[8] Zeleny, M. (1987). Management support systems: Towards integrated knowledge management. *Human Systems Management*, *7*(1), 59-70.

[9] Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science*, *5*(1), 14-37.

[10] Bernstein, J. H. (2011). The data-information-knowledge-wisdom hierarchy and its antithesis. *Nasko*, *2*(1), 68-75.

[11] Awad, E. M., & Ghaziri, H. M. (2004). *Knowledge management*, 2004. ed: Prentice-Hall, Upper Saddle River, New Jersey.

[12] Smith, A. (1999). Why digitize?. *Microform & Imaging Review*, *28*(4), 110-119.

[13] Baars, H., & Kemper, H. G. (2008). Management support with structured and unstructured data—an integrated business intelligence framework. *Information Systems Management*, *25*(2), 132-148.

[14] Li, G., Ooi, B. C., Feng, J., Wang, J., & Zhou, L. (2008, June). EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 903-914). ACM.

[15] Power, D. J., & Sharda, R. (2009). Decision support systems. In: Nof S. (eds) *Springer handbook of automation* (pp. 1539-1548). Springer: Berlin Heidelberg.

[16] McCallum, A. (2005). Information extraction: Distilling structured data from unstructured text. *Queue*, *3*(9), 48-57.

[17] Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the Association for Information Science and Technology*, *58*(4), 479-493.
[18] Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. Communications of the ACM, 57, 86-94.
[19] Monash, C. (2010). *Three broad categories of data*. Retrieved from http://www.dbms2.com/2010/01/17/ three-broad-categories-of-data/
[20] Ku, M. (2018). [IT & Future Strategy] Analysis of Four Public Policies Key to the Growth of Data-Driven Business. National Information Society Agency: Seoul, South Korea
[21] Ofner, H. M., Straub, K., Otto, B., & Oesterle, H. (2013). Management of the master data lifecycle: a framework for analysis. *Journal of Enterprise Information Management*, *26*(4), 472-491.
[22] Otto, B., Hüner, K. M., & Österle, H. (2012). Toward a functional reference model for master data quality management. *Information Systems and e-Business Management*, *10*(3), 395-425.
[23] Reid, R., Fraser-King, G., & Schwaderer, W. D. (2007). *Data lifecycles: managing data for strategic advantage*. John Wiley & Sons.
[24] Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information systems management*, *29*(4), 258-268.
[25] Benz, U. C., Hofmann, P., Willhauck, G., Lingenfelder, I., & Heynen, M. (2004). Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. *ISPRS Journal of photogrammetry and remote sensing*, *58*(3-4), 239-258.
[26] Perrin, P., & Petry, F. E. (2003). Extraction and representation of contextual information for knowledge discovery in texts. *Information sciences*, *151*, 125-152.
[27] Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata principles and practicalities. *D-lib Magazine*, *8*(4), 1082-9873.
[28] Fegraus, E. H., Andelman, S., Jones, M. B., & Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: An introduction to Ecological Metadata Language (EML) and principles for metadata creation.
[29] Yau, N. (2013). *Understanding data – Context*. Retrieved from http://bigthink.com/experts-corner/understanding-data-context.
[30] National Academies of Sciences, Engineering, and Medicine. (2017). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. National Academies Press.
[31] Smith, G., Noble, M., Anttilla, C., Gill, L., Zaidi, A., Wright, G., Dibben, C., & Barnes, H. (2004). The Value of Linked Administrative Records for Longitudinal Analysis, Report to the ESRC National Longitudinal Strategy Committee. Swindon: ESRC.
[32] Zainal, Z. (2007). Case study as a research method. *Journal Kemanusiaan*, *5*(1).1-6.
[33] Yin, R. K. (2013). *Case study research: Design and methods*. Sage Publications.