# Table of Contents

Introduction ....................................................................................................................... 2  
1.1 Background ............................................................................................................... 2  
1.2 The Need for Guidance on Transit Data Sharing ....................................................... 2  
1.3 Research Scope and Key Questions ......................................................................... 5  

2 Study Status Report ..................................................................................................... 7  
2.1 Completed Tasks ...................................................................................................... 7  
2.2 Upcoming Tasks ....................................................................................................... 8  

3 Data Collection and Management in Public Transit Agencies .................................... 13  
3.1 Transit Data Types and Attributes for Data Sharing ............................................... 13  
3.2 Preparing Data for Sharing ..................................................................................... 20  

4 Sharing of Public Transit Data .................................................................................... 28  
4.1 Models for Public Transit Data Sharing ................................................................... 28  
4.2 Factors Impacting Transit Agency Decisions About Data Sharing ......................... 35  
4.3 Deeper Dive: Sharing Route, Schedule, and Vehicle Location Data and the Customer Experience ........................................................................................................... 47  

5 Accessing External Data Sources ............................................................................... 51  
5.1 External Data Types and Transit Agency Uses ......................................................... 51  
5.2 Models to Access External Data .............................................................................. 52  

6 Key Findings and Next Steps ...................................................................................... 56  

7 Bibliography ................................................................................................................ 58  

8 Appendix ...................................................................................................................... 63  
I. Transit Agency Interview Guide .................................................................................. 63  
II. Private Sector Interview Guide ................................................................................ 65  
III. Public Sector Interview Guide ................................................................................ 68  
IV. Utility Industry Interview Guide ............................................................................. 70
1 INTRODUCTION

This is the Interim Report for TCRP J-11 (31), which is aimed at understanding the value of data sharing for transit agencies. This project will culminate in a final report including a guidance section that can be consumed as a stand-alone document or as part of the report. This Interim Report summarizes the progress so far and also lays out the next steps that will be pursued to complete this project. This report summarizes information collected through a review of literature, web, and news sources and through a series of interviews. Findings are summarized in chapters on transit data management, sharing of public transit data, and access to external data by transit agencies. This report also includes a chapter on the tasks completed and upcoming tasks in this project. This report concludes with a summary of findings and next steps.

1.1 Background

Data has become a critical component of many fields, and public transit is no exception. Public transit systems generate a substantial amount of data, including vehicle location records, records of passenger boardings and fare transactions, information on routes, schedules, and real-time alerts. Public transit customers in many cities are accustomed to receiving information on transit services in real-time on their smartphones, using apps developed primarily by private companies and fed by public data.

The reliance of transit agencies and customers on data is only expected to increase, as predicted by the World Bank. The literature reveals that researchers and transit agencies use transit data for a variety of planning and forecasting applications, including:

- Making network and service adjustments
- Forecasting demand
- Informing fare policy

“The transportation industry is a leader in creating the Internet of Everything, generating vast volumes of data each day through sensors in passenger counting and vehicle locator systems and ticketing and fare collection systems, to name just a few.”

(Rosado, 2014)

1.2 The Need for Guidance on Transit Data Sharing

Just as there is value in having data, there can be value in sharing data. The mobility landscape is changing, with partnerships in service provision across modes and between public and private mobility providers (for example, partnerships between transit agencies, bikeshare providers, and transportation network companies like Uber and Lyft). These models can facilitate and in fact often necessitate data sharing.
In 2017, Citymapper, a transit app company, launched two pilot bus routes in partnership with Transport for London. Citymapper identified this opportunity and selected the routes based on data collected from their app users. This example illustrates a growing phenomenon: that both within and outside the field of transportation planning, mobility data obtained from users is becoming increasingly valuable. Today, marketers use location information and mobility patterns to tailor ad campaigns – pushing mobile ads for nearby restaurants or targeting mobile users on their way home from work. Data from cell phone apps, GPS probes in vehicles, and Bluetooth and WiFi signals in mobile devices are collected, processed, and sold by companies including HERE, INRIX, and Cuebiq.

Engaging with other players in data sharing has the potential to create benefits for transit agencies and their users. At the same time, there are challenges and risks that discourage agencies from making these agreements. Additionally, the context for making these agreements is evolving. As new forms of data become available, new possibilities for sharing, trading, and selling data emerge. As shown in Figure 1, these factors drive the need for guidance.

“When we think about what transport will look like in the future, one of the key things we know is that it will be filled and underpinned by data.”

(Peralta Quiros, 2018)

![Figure 1 The Need for Guidance on Transit Data Sharing](image)
There are many other reasons transit agencies may wish to share data. Sharing data can:

- Promote transparency and engagement with customers
- Spur innovation and research
- Save costs
- Generate revenue
- Positively impact customers

A decision framework for data sharing must weigh these benefits against the cost of providing data and the risks of sharing it. Preparing data for sharing can be resource-intensive and can be both technically and organizationally challenging for some transit agencies. Technically, data sharing requires knowledge of data cleaning, processing, and storing, including the application of appropriate cybersecurity and privacy protection measures. Data sharing also often includes terms of use or a data sharing agreement, and transit agency staff must know which terms to attach in each case. Organizationally, transit agencies are working to find the most efficient staff structures to manage their growing volumes of data and the data sharing requests they receive.

The development of organizational structures and technical knowledge for data management can help transit agencies develop protocols for assessing risks when they make decisions about data sharing. The risks of sharing data include privacy risks, security risks, risks of data misuse, and strategic risks.

These risks are defined as follows:

- **Privacy risks** are present whenever data has the potential to identify individuals. Sometimes, the potential for a data set to be combined with other data sets increases this risk. Transit agencies can take steps to protect privacy that include encryption of identifiers, aggregation, and adding noise (random variation) to data to obfuscate individual patterns.

- **Security risks** can be present if data provides special insight into infrastructure and the locations of the people who use transit that could be used in a physical attack.

- **Risks of data misuse** can be present whenever data is shared. While transit agencies seek to mitigate this risk through data documentation, some users may intentionally or unintentionally misinterpret data, drawing conclusions that are incorrect.

- **Strategic risks** are defined as the risk that sharing data could compromise the transit agency’s ability to serve its customers. This includes risks to the transit agency’s reputation and the risk that the information is used against the transit agency (for example by competitors).

Finally, data sharing decisions require understanding of the laws that govern information sharing and data privacy. In many instances, as public agencies, transit agencies are required to share data, if requested. However, these laws generally include provisions stating that data must not be shared if the privacy of individuals is at risk. Including the correct contractual provisions in a data selling, sharing, or trading agreement also requires technical understanding of the data as well as legal expertise.
1.3 Research Scope and Key Questions

There are several models for data sharing. Broadly, this report considers the following types of data sharing:

- **Public data sharing**, often called open data, releases data to the public, sometimes including terms of use. This model has successfully spurred innovation and research in many industries, including public transit. However, it limits the ability of the transit agency to direct the way that data is used. It is only appropriate for data that has low privacy and security risks.

- **Private data sharing** includes a variety of data sharing models in which data is shared directly with an individual or partner, sometimes with non-disclosure provisions that prohibit further sharing.

Either of these types of data sharing can generate value for transit agencies. As noted, value can be generated through external innovation and research, through increased awareness of transit services, through cost savings, and in some cases through revenue generation. This report does not explicitly consider data selling as a separate category of data sharing because there appear to be no documented instances of data selling by transit agencies. However, in some cases transit agencies generate revenue by selling unique analysis of transit data or by using transit data to increase advertising revenue.

Another form of private data sharing is a data exchange or data trade, in which a transit agency provides data to a partner and receives data from the partner in return. This specific model appears to be rare for transit agencies. Instead, there are many examples of transit agencies providing data and receiving other benefits in return (such as data analysis, publicity, customer information platforms). Alternatively, there are also examples of transit agencies leveraging service agreements and their relationships with customers to access external data sources. These examples are not data-for-data trades and are therefore defined in the broader context of data sharing agreements.

As such, this report tackles the broad concept of public transit data sharing from two perspectives. The first is the sharing of public transit agency data and the generation of value from this data. **Chapter 3** discusses different types of data collected by transit agencies and their potential for data sharing. It also covers the process of preparing data to be shared. It focuses on questions including:

- What data should be collected?
- How should the data be cleaned, processed, and stored?

**Chapter 4** introduces models for data sharing and discusses the factors that should be considered in decisions about data sharing, tackling questions including:

- What models exist for public transit data sharing?
- What is the legal context for public transit data sharing?
- What are the potential risks of public transit data sharing?
- What benefits can public transit agencies receive from sharing their data?
- What are the costs of data sharing?
The second part of the report (**Chapter 5**) looks at public transit agency access to external data types and considers the mechanisms transit agencies can use to access this data. Prior to the reporting on data sharing, this Interim Report includes a status report on the project, in **Chapter 2**. Finally, **Chapter 6** summarizes the key findings and discusses next steps.

Because of the large number of acronyms in this research, we have created a glossary of acronym definitions, provided in Table 1.

**Table 1 List of Acronyms**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFC</td>
<td>Automated Fare Collection</td>
</tr>
<tr>
<td>APC</td>
<td>Automated Passenger Counter</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>AVL</td>
<td>Automated Vehicle Location</td>
</tr>
<tr>
<td>FTA</td>
<td>Federal Transit Administration</td>
</tr>
<tr>
<td>GDPR</td>
<td>General Data Protection Regulation</td>
</tr>
<tr>
<td>GTFS</td>
<td>General Transit Feed Specification</td>
</tr>
<tr>
<td>GTFS-RT</td>
<td>General Transit Feed Specification Real-time</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>ITS</td>
<td>Intelligent Transportation Systems</td>
</tr>
<tr>
<td>MOD</td>
<td>Mobility on Demand</td>
</tr>
<tr>
<td>UITP</td>
<td>International Association of Public Transport</td>
</tr>
<tr>
<td>TNC</td>
<td>Transportation Network Company</td>
</tr>
</tbody>
</table>
2 STUDY STATUS REPORT

The study is well underway with Tasks 1, 2, and 3 completed and work on Task 4 in progress.

<table>
<thead>
<tr>
<th>Task</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Kickoff &amp; AWP</td>
<td>Complete</td>
</tr>
<tr>
<td>2 Information Collection: Transit Agency Experience</td>
<td>Complete</td>
</tr>
<tr>
<td>3 Interim Report</td>
<td>Complete</td>
</tr>
<tr>
<td>4 Additional Information Collection: Other Institutions and Industries</td>
<td>Underway</td>
</tr>
<tr>
<td>5 Development of Guidance</td>
<td>To Be Done</td>
</tr>
<tr>
<td>6 Finalize Report</td>
<td>To Be Done</td>
</tr>
</tbody>
</table>

2.1 Completed Tasks

We began the project with a review of literature, news, and guidance related to the value of data, transit data types and uses, and transit data sharing, which are listed in the bibliography and are referenced throughout this document.

As part of Task 2 we distributed an online survey to transit agencies designed as a screening tool to identify transit agencies for interviews. We received 32 responses and went on to interview representatives from 11 transit agencies. The interview guide is included as an appendix to this report. The transit agencies interviewed were:

1. MBTA (Boston)
2. Houston Metro
3. PSTA (Florida)
4. Metro Transit (Minneapolis-St. Paul)
5. Los Angeles Metro
6. San Francisco MTC
7. TANK (Northern Kentucky)
8. RTD (Denver)
9. New York City Transit
10. TransLink (Vancouver)
11. Transport for London
Several of the transit agencies requested to not be named in the report. Hence, this report does not refer to specific transit agencies except when citing from information that is available on transit agency websites or in published documents.

The findings of the literature and information review and from the interviews with transit agencies are integrated and reported on in the remaining chapters of the report.

We interviewed the following subject matter experts:

- George Kocur, expert in smart card systems and data
- Gabriel Sanchez Martinez and Jay Gordon, MIT research scientists and founders of Korbato
- Professor Konstantinos Gkiotsalitis, University of Twente
- Professor Rongfang (Rachel) Liu, New Jersey Institute of Technology
- Professor Kari Watkins, Georgia Tech

We also reviewed literature on data management and sharing by other government and publicly-regulated agencies and among private data providers. That literature was used to inform the interview plan for Task 4, which is described in the following section on Upcoming Tasks. The final report will integrate the literature review and findings from the additional interviews to draw insights that are most relevant to transit agencies.

### 2.2 Upcoming Tasks

Task 4 is underway and consists of additional information gathering with a focus on the legal context and on key perspectives outside of transit agencies. It includes:

- A review of the legal context for public transit data sharing
- Interviews of private mobility providers and private data owners
- Interviews of public agencies
- Interviews focused on data sharing in the energy sector

The legal analysis will be conducted by Steptoe and Johnson, LLP, under the guidance of the core research team. Through this analysis, we seek to answer the following questions:

- What is the current legal framework for data sharing by transit agencies?
- What, if any, potential changes in the legal framework should transit agencies be prepared for?
- What guidelines should transit agencies follow to protect the privacy of individuals when sharing data that includes individual records?
- What terms should transit agencies include when they share data, either openly or with an individual or partner organization?
Insights from the legal analysis will be integrated into the final report and will inform the Guidance Document developed in Task 5.

The interviews for Task 4 are underway. Discussion guides for each set of interviews are included as an appendix to this report.

The interviews of private mobility providers and data owners include a variety of company types and data collection models. The planned interviews include the following:

1. Mastercard (Miguel Gamiño, Jr., Executive Vice President for Global Cities). Mastercard’s collaboration with Carto (Mastercard Retail Location Insights) provides a case study for private-sector data exchange.

2. INRIX (Rick Schuman, Vice-President, Public Sector). INRIX is an industry leader in real-time GPS and mobile data analytics. They have a range of private-sector and public-sector partnerships.

3. CARTO (Stuart Lynn, Head of Research and Wenfei Xu, Data Science). CARTO provides location intelligence in a SaaS platform and has both public sector and private sector clients and partners.

4. Cuebiq (Brennan Lake, Senior Director of Research Partnerships). Cuebiq is a location intelligence and consumer insights industry leader, with a platform that claims to collect data from over 61 million monthly active U.S. smartphone users on over 180 mobile apps.

5. Zipcar (Lauren Alexander, Manager, Data Science). Zipcar is an industry leader in the shared-mobility landscape and provides an example of a long-established mobility company.

6. Uber (Andrew Salzberg, Head of Transportation Policy and Research). Uber is a now-ubiquitous part of the transportation landscape and has committed to sharing anonymized data through the Shared Streets data platform.

7. Sidewalk Labs (Corinna Li, Urban mobility). Sidewalk Lab’s Replica program offers location and movement insights to cities by gathering and anonymizing cell phone location data.

8. Skyhook. Skyhook was an early innovator in Wi-Fi-based location data and provides location insights to various private-sector partners.

9. Lyft (Michael Frumin, Research Science Manager, NYC). As a leader in the emerging mobility on demand market, Lyft has access to vast amounts of transportation data.

10. Populous AI (Regina Clewlow, CEO). Populous AI’s aim is to facilitate data sharing between private mobility companies and cities.

11. Transit App (Stephen Miller, Communications Lead). Transit App integrates data from public transportation and ride-hailing operators into a streamlined app. Their business model is made possible through collaboration with public-sector transit agencies.

The public sector interviews are focused on organizations that are at the forefront of data management and sharing and include cities and agencies of various sizes across the United States. We have included some transit agencies to gather additional perspectives not included in the first round of transit agency interviews.
The planned interviews are:

1. **Dallas Area Rapid Transit System First and Last Mile Program, Todd Plesko.** As part of the Mobility on Demand Sandbox program, DART has implemented an integrated transit app (GoPass++) with expanded mobility offerings to increase efficiency and reduce the first and last mile problems faced by users. The app collects trip and rider information that is accessed by DART and used to evaluate the transit system. We are particularly interested in the development of the app and the ownership structure of the collected data.

2. **City of Denver, Office of Transportation and Mobility.** In their Smart City Challenge Proposal, the City of Denver identified three major challenge areas: Mobility on Demand, Transportation Electrification, and Intelligent Vehicles. Although Denver's proposal was not selected for funding, their application is informing future transportation planning decisions. The city hopes to forestall the negative impacts of rapid population growth by leveraging the data they collect into a responsive transit monitoring system. Private partnerships with mobility companies are key to their proposal, including the development of an integrated transit application. Through our interview, we will investigate the proposal to make all collected data (except that with personally identifiable information) open for reuse.

3. **Kansas City, Eric Roche.** Kansas City’s Smart City Challenge Proposal offered three “pillars” of smart transportation development: constructing a rapid transit line, automated/connected vehicles, and connecting/empowering communities. Although Kansas City was not selected as the final recipient of funding, we felt it was important to include a midsized city in our research. We hope to compare the public-private data sharing strategies between midsized and larger cities.

4. **Bay Area Rapid Transit.** Joy Bonaguro, Blake Valenta, Ash Fellow. BART has partnered with the carpool company Scoop to limit the number of vehicles parking at BART stations. Carpools through Scoop will be given guaranteed or priority parking.

5. **Vermont State DOT, Vermont Agency of Transportation.** The Vermont Agency of Transportation faces a unique challenge: how to leverage mobility apps that are built for dense, urban areas. Through their Mobility on Demand Sandbox proposal, VTrans will build out an application that includes flexible ride services, including those without GTFS compatibility. The VTrans application will be built on open-source code, making it available for replication.

6. **City of Cincinnati.** The Cincinnati Mobility Lab is an innovate three-year partnership between the city and Uber. Our team learned about the data sharing model in our interview with TANK and will gain more insight by speaking to the city.

While the literature review considered data sharing models across a variety of industries, the energy sector was selected for a deeper analysis. The sector has made progress in establishing data standards and promoting opt-in models that enable individuals to share their data. The sector has many interesting partnerships between publicly-regulated utilities and private companies that are similar to relationships between public transit agencies and private mobility and data providers.
The planned interviews are:

1. London Hydro, Zoran Stojanovic. London Hydro’s “Connect My Data” is the utility industry’s first example of a cloud-based platform set up to host the utility’s customer data, any third-party device, and software apps.
2. Pacific Gas and Electric, Amy Kight. PG&E is transitioning from a decentralized data sharing architecture to a centralized data sharing platform (like London Hydro’s “Connect My Data”). We will ask about their decision to revise their data sharing platform, process, and lessons learned to date.
3. Commonwealth Edison. ComEd is selling anonymized customer data to third parties.
4. Avista. A utility in western Washington state that is collaborating with Duke Energy on OpenFMB.
5. Open Energy Solutions. Was announced as the company that the OpenFMB companies are investing in as a platform for interoperability and node computing and may be more available/interested in talking than the utilities are.
6. DataCapable. Small startup that integrates social (Twitter, Facebook, etc.) and GIS data for utilities to manage power outages

Following the information collection phase, we will develop the guidance document that will be part of the Final Report. The Final Report will follow a structure similar to the Interim Report but will include additional detail on the legal context and lessons learned from the research and interviews of the private sector and other public sector agencies. The following is an outline of the Final Report.

1. Introduction
2. Research Methods
3. Data Collection and Management in Public Transit Agencies
   a. Transit Data Types and Attributes for Data Sharing
   b. Preparing Data for Sharing
4. Sharing of Public Transit Data
   a. Models for Public Transit Data Sharing
   b. Factors Impacting Transit Agency Decisions about Data Sharing
   c. Deeper Dive: Sharing Route, Schedule, and Vehicle Location data and the Customer Experience
5. Accessing External Data Sources
   a. External Data Types and Transit Agency Uses
   b. Models to Access External Data
6. Guidance
7. Conclusions
The guidance will take the following form:

1. Establishing Transit Agency Goals
   a. Research and analytical goals
   b. Customer information goals
   c. Strategic goals including transparency and competitiveness
   d. Information goals – external data

2. Choosing Models to Meet Transit Agency Goals
   a. Process for choosing models for sharing transit agency data
      i. Evaluating legal context
      ii. Evaluating risks
      iii. Evaluating benefits and costs
      iv. Selecting a model and finding partners
   b. Process for choosing models for accessing external data
      i. Identifying points of leverage
      ii. Finding the right partner

3. Making it Happen: Data Management
   a. Identifying and fulfilling the organizational and technical needs based on data sharing goals and selected data sharing and data access models
Public transit data comes in many forms with a variety of potential external uses, if shared. Decisions about data sharing, as described in Section 4, must factor benefits, costs, and risks of data sharing into account. These factors vary by data type. Hence it is useful to review public transit data through the lens of data sharing, considering the benefits, risks, and costs of sharing. In addition, the effort required to share data depends on the practices that transit agencies maintain to manage data internally. If a transit agency cleans, processes, documents, and catalogs data for internal use, the additional effort to prepare this data for sharing is likely to be minimal. Therefore, the foundation of any discussion on transit data should include a review of transit data collection and management practices.

Section 3.1 inventories transit data types and comments on the characteristics of each as related to data sharing. Next, Section 3.2 describes the process of preparing data to be shared, beginning with data collection, and including data cleaning, processing, documenting, and cataloguing. This chapter reports on documented examples from the literature and web materials and from the transit agency interviews completed for this project.

### 3.1 Transit Data Types and Attributes for Data Sharing

Figure 2 summarizes the sharing characteristics of the major public transit data types, based on the review of literature and the transit agency interviews. Public transit data can be classified into two broad categories – data pertaining to passengers and data pertaining to the public transit system itself. One exception is incident data, as incidents may involve passengers as well as infrastructure, vehicles, and transit agency staff.

Data pertaining to passengers includes passenger count data and survey data, as well as newer data types including fare or bank card transactions, video, Wi-Fi, Bluetooth, and app and webpage usage data. Most, but not all, types of passenger data contain records of individual passengers or records pertaining to a specific card or device that has the potential to identify an individual. This is a critical distinction for data sharing, as the sharing of individual records poses a privacy risk.

Data pertaining to the system includes route and schedule data, vehicle location data, maintenance, staff and operations data, and financial data. These data types typically do not contain privacy risks, however, there may be security risks associated with the sharing of some of these data types. While security was not raised as an issue in the majority of the transit agencies interviewed for this study, a few transit agencies indicated that they refrained from sharing some detailed system and system...
Figure 2 Transit Data Types and Sharing Characteristics
usage (passenger) data due to concerns that it could be used to stage an attack on public transit infrastructure and the people who use it.

There are many reasons that transit agencies share data, including legal obligation. The factors that govern transit agency decisions to share data are discussed in more detail in Section 4.2. This section on data types highlights some of the key impacts of transit data sharing based on the literature and transit agency interviews. These include:

- Data sharing promotes **transparency** and can increase awareness of the transit agency and engagement with transit customers.
- Data sharing can support **research** that can help transit agencies plan better service and operate more efficiently.
- Data sharing can spur innovation in the private sector that can benefit customers, including through better **customer information**.
- Data sharing can generate revenue, for example through **advertising**.
- Data sharing can support other community functions, informing **municipalities, real estate developers**, and even **law enforcement agencies**.

The sharing of any data type promotes **transparency**. The majority of transit agencies interviewed reported that they sharing data improves relations with their customers, promoting trust and awareness. However, transit agencies must weigh this impact against the potential risks of sharing data and the costs of preparing the data to be shared. For example, despite potential transparency benefits, data that could be used to identify individuals, and hence presents a privacy risk, is generally not shared publicly or in a disaggregated form.

All data types also have the potential to spark **research**, though different data types enable different research questions to be addressed. Example research questions for each data type are noted below. In addition, Section 4.1 describes models transit agencies use to ensure that external research answers questions are relevant to the agency’s needs.

Route and schedule data and vehicle location data are the primary sources of **customer information**, often through third party apps used for trip planning and to check the timing of transit vehicle arrivals at stops and stations. In the future, data on passengers may feed customer information platforms as well. For example, vehicle crowding information could be presented to customers in real-time.

Any form of data on passenger use of the system has the potential to be leveraged to increase the transit agency’s **advertising** revenue, as advertisers are interested in knowing the volume of passengers their in-station or on-vehicle ads will be exposed to, as well as characteristics of those passengers. Because of privacy concerns, transit agencies typically do not provide this information directly to advertisers. Instead, they may provide depersonalized and aggregated data (Transport for London, 2018). At least five of the transit agencies interviewed had already used data to generate advertising revenue or were considering doing so.
Transit agencies also reported that data on passengers is requested by real estate developers, municipal planners, and law enforcement officers. Transit agencies attempt to support these community needs while also protecting private information on their customers.

**Fare and Bank Card Data**

When passengers use smart cards or bank cards for boarding, as is becoming increasingly prevalent, it produces records of boardings or station entries and also introduces the possibility of tracking the use of a given card over time. While this characteristic of the data raises privacy risks, it also enables transit agencies and researchers to understand more about how passengers use the network.

Some researchers and agencies, including Boston’s MBTA, Washington Metropolitan Area Transit Agency (WMATA), and New York City Transit (NYCT) have applied methods to infer destinations of passengers within their systems based on automated fare collection (AFC) data combined with automated vehicle location (AVL) data (Barry et al., 2002; Gordon et al., 2013). This origin-destination level data is valuable to transit agencies and researchers as it typically provides a large, year-round sample of passenger demand patterns. Smart card data can also be used to understand how the travel behavior of users changes over time, including in response to fare or service changes, or weather, even at micro levels such as a bus stop or zone (Morency, 2007).

Smart card data also facilitates the grouping of transit users by their distinct trip sequence structure. Using Transport for London data for a four-week period, researchers clustered transit users based on their activity patterns. They found that 40 percent of frequent transit users did not follow a conventional trip activity sequence involving one trip to work in the morning and another trip home in the evening (Goulet-Langlois et al., 2016). Information on the passenger segments that use different stations or routes may be provided to advertisers to enable them to customize ads to different passenger types (commuters, visitors, etc.). As such, this data is not only valuable for research purposes, but can also be used to generate revenue through advertising.

**Wi-Fi and Bluetooth Data**

While smart card and bank card data can provide stop or station level origin-destination information, Wi-Fi and Bluetooth connection records enable even more detailed tracking of passenger movements within stations or within a gated transit system. For example, Transport for London engaged in a pilot study collecting Wi-Fi signals from passengers’ phones, which they used to understand passengers’ route choices within the subway network (Transport for London, 2017). This aspect of Wi-Fi and Bluetooth data makes it especially valuable for research, and also makes it an important source for informing advertising. Transit agencies can estimate not only the number of passengers who pass through a station, but also the number of passengers who pass a specific location within a station. This has the potential to enable the transit agency to generate more advertising revenue by providing these detailed statistics to advertisers (Cheshire, 2017). Like fare card transaction data, the fact that this data tracks individual devices means that there are some privacy risks associated with sharing the data in disaggregate form.
Video Data

Many transit vehicles and stations are equipped with video cameras. From a research and planning point of view, this data can provide insight on crowding and passengers left behind at stations. This data can also be valuable for police investigations, and there are examples of transit agencies sharing this data with law enforcement agencies. In a recent example in Boston, MBTA video surveillance data and fare card data were used to locate a kidnapped woman (Flanigan, 2019). Video system vary; some, but not all systems enable “pan-tilt” or “zoom-in” features which allow for facial recognition. Particularly in systems that permit facial recognition, there are significant privacy risks associated with storing and sharing this data (Thomas, 2018).

Transportation App and Webpage Usage Data

Some transit agencies have developed or commissioned the development of a customer-facing transportation app for trip planning, ticketing, or both. When transit agencies develop or own their apps, they can harvest data from these apps. In trip planning apps, app users’ destination requests are saved. In bus or train arrival apps, the specific bus or route information requested is saved. In addition, many apps save data on users’ locations while using the app, which can be used to infer origins and destinations for trips made (Lu et al., 2015).

In a research context, this data is interesting because it can be used to draw insights about mode choice and alternatives. Users may search for transit directions in an app, but ultimately choose to use a different mode. This type of information can be inferred from app data, making it a sought-after source for transit competitiveness and mode choice studies.

These apps and the data they generate also have the potential to be used in location-based advertising and other geotargeted information. According to a recent study, most transit agency ticketing apps are location-aware, but this feature is only used to locate nearby stops/routes. While this may be an untapped revenue source for transit agencies, initial research suggests there may be push-back to this type of advertisement (Brakewood, 2017).

Any discussion of transportation app data merits noting that the majority of transit agencies rely on third-party developers to provide customer-facing transportation apps to their customers. Some transit agencies have access to the data from these apps, while others do not. Different models for app development and data access are described in Section 4.3.

Even transit agencies that do not have a proprietary app have a public website that often includes a route planning tool. While usage of web planning tools may be more limited and does not provide location information, transit agencies may nonetheless be able to generate value from their web traffic analytics. Transit agencies can draw insight from the type of information their web visitors access, and this information may also be useful for local planners and developers.

“The review of location-based advertisement (LBA) suggests that there may be mixed levels of receptiveness to LBA by mobile phone users.”

(Brakewood, 2017)
Survey Data

Transit agencies conduct surveys of their users regularly. While some of the information collected in surveys, such as origin-destination patterns, can now be inferred from other sources, surveys continue to be a valuable source of information on things like trip purpose, trip alternatives considered, and demographic characteristics of transit customers. Surveys are also used to assess customer satisfaction and collect information about transit passengers’ preferences and priorities. In short, surveys often provide information that cannot be gathered from other transit agency sources, nor can it be inferred from external data sources, such as cellphone and GPS data. As a result, this data may be valuable to researchers and others, particularly if it can be combined with other data sets. Raw survey data can pose privacy risks because responses can contain identifying information such as home addresses and demographic characteristics of the respondent. Instead of sharing survey data openly, transit agencies typically share aggregated reports on the surveys they conduct. Of the 11 transit agencies interviewed, one publishes survey responses to its customer satisfaction survey, aggregated by month. Six other transit agencies interviewed publish reports that summarize survey findings.

Passenger Count Data

Some systems for monitoring passenger movements provide aggregate information on passenger counts without tracking individuals. Automated passenger counting (APC) systems use sensors to estimate the number of passengers that board a vehicle, and “load weigh” data (data on the weight of a train or vehicle and its occupants at different points along a route) allows analysts to estimate passenger loads. Some fare collection systems also produce anonymous count data, for example, estimates of the number of people who passed through a turnstile or interacted with a farebox. Fareboxes often record all boardings regardless of fare payment type, and even unpaid boardings recorded by the driver. As a result, this data can support studies of fare evasion. More broadly, it is used to analyze crowding and productivity of routes and lines. Except in the case of very small samples (for example if just one person boards a bus at a stop during the period reported on), this data does not enable the identification of individuals and therefore does not elicit privacy concerns. Four of the transit agencies interviewed reporting sharing APC data with researchers (in three cases) or municipalities (in one case).

Incident Data

Transit agencies collect data on incidents, including details on the cause of incidents and the operational response. They also collect data on passenger injuries and claims. Sharing this data can support research that helps transit agencies improve incident response protocols or prevent incidents. In rare cases, incident data may pose privacy risks if individuals involved in the incidents are described in an identifiable way. One of the transit agencies interviewed noted that they do not release incident data publicly because of the staff effort that would be required to read descriptive data fields to confirm that they could not be used to identify individual passengers. In general, incident data is not released publicly by the transit agencies interviewed (though some publish real-time alerts about incidents). However, on transit agency interviewed shares this data with a research partner who has analyzed incident response and passenger disruption impacts.
Route and Schedule Data

Route and schedule data is commonly shared publicly by transit agencies. The General Transit Feed Specification (GTFS) is a standardized format for this data, though some agencies use propriety formats from scheduling software companies. Nearly all transit agencies surveyed in a 2015 survey provided this type of data free-of-cost (Schweiger, 2015). This data is used in trip planning and real-time transit information applications (Antrim, 2013; Schweiger, 2015).

Transit agencies also have more detailed system data, such as station diagrams. This data can be useful for research on how passengers move through the system, but some transit agencies opt not to share it widely due to security concerns.

Automated Vehicle Location Data

AVL data tracks the location of vehicles over time. As described in the section on fare collection data, AVL data is often a critical input to analysis that infers passenger destinations (Gordon, 2013). In addition, AVL data can be used to track and display system performance, evaluating headway variability and schedule adherence.

Transit agencies use data from AVL systems to provide information to customers on the next train or bus arrival. Many transit agencies share AVL data streams publicly, and app developers use this data to fuel transit arrival apps (Schweiger, 2015). In many cases, this is accompanied by real-time alert information. GTFS-Realtime is a standardized feed specification for this type of data, although not all transit agencies use this format for the published data (Barbeau, 2018A).

System and Vehicle Maintenance Data

This category of transit data may include records of failures and maintenance activities, as well as maintenance facilities and maintenance costs. The National Transit Database (NTD) collects records of annual failures and maintenance facility types and sizes from transit agencies. The use of transit vehicle maintenance data is not a prevalent topic in recent literature. It is primarily discussed in in the context of life cycle cost assessments (Chester and Horvath, 2010). Some researchers have considered the possibility of using sensor data to predict maintenance needs (Corazza et al., 2018).

Staffing and Operations Data

Staffing and operations data includes crew and vehicle assignments, absenteeism data, and operational procedures. This data can support research on operational efficiency and scheduling, which may ultimately allow the transit agency to operate more efficiently. However, there was little discussion of these data types in the transit agency interviews on data sharing. In the area of crew scheduling, there was considerable research in the past, but there are now off-the-shelf solutions that transit agencies use. One of the transit agencies interviewed provided operations data to a research partner who helped them pilot a new bus operations method.
Financial Data

Financial data includes transit agency spending and subsidies. The NTD collects information on transit agency spending. Sharing this data helps transit agencies maintain transparency and accountability. One of the transit agencies interviewed posts budgeted and actual expense and revenue data on a monthly basis.

3.2 Preparing Data for Sharing

One of the key findings from the public transit interviews is that sharing data is no easy task. Once collected, data must be cleaned. Next, many data sources require additional processing. Sometimes different sources must be reconciled and combined. Data is sometimes aggregated either for storage reasons or to preserve privacy. Sometimes, data must be converted to a standard format. Data processing tasks vary in part by the audience that data will be shared with. Less technical audience tend to prefer data that has already been aggregated, while researchers prefer to be able to manipulate disaggregate data. Prior to sharing, data should also be documented. Some transit agencies opt to create data catalogues, making data more accessible internally and easier to share externally. Figure 3 summarizes the process of preparing data for sharing.

![Figure 3 The Process of Preparing Data for Sharing](image-url)
Fulfilling these tasks can present technical challenges for some agencies that lack technical staff (Brauneis and Goodman, 2017; Lawson, 2016). These challenges only increase with large-scale data that can require machine learning techniques for processing and requires scalable data storage and mining (Zaslavsky et al, 2013). In fact, this is often the reason that public agencies partner with private companies or universities who can help complete some of these data processing tasks both for the agency’s internal use and for broader sharing. However, some argue that these partnerships take power away from the public agencies, particularly when external partners fail to transparently describe the methodologies they use to process data (Brauneis and Goodman, 2017). The transit agencies interviewed spoke positively about the technical assistance they receive through data sharing.

In general, the transit agencies interviewed were less concerned with having technical skills in-house and more concerned with having the time to dedicate to data preparation tasks. Many expressed that the structure of their agency contributed to a lack of effort devoted to data management tasks. Most transit agencies interviewed do not have staff or divisions dedicated to data management, which means staff have other priorities. Those transit agencies that most actively analyze data internally tend to be most well-equipped, both technically and organizationally, to prepare data for sharing. Because of their internal capabilities, these transit agencies may be least in need of external research and innovation using their data. The needs of small agencies to develop data sharing infrastructure require special attention. Organizational and technical challenges in public transit data sharing are discussed in more detail in Section 4.2.

The following sections describe each step of data preparation for sharing, highlighting challenges associated with each.

**Data Collection**

Decisions about data collection determine the types of data and the data quality and coverage available to be shared. However, data collection is often more of a byproduct of system design than a dedicated analysis effort. The data generated is often dependent on functional aspects of the system, such as operations and fare collection rather than potential data analysis or the value that can be generated through sharing data. Some systems pre-date modern conceptions of open data and in some cases, considerable effort is required just to extract data from the system. One exception is survey data, which is typically collected expressly for the purpose of analysis and evaluation. Several of the transit agencies interviewed identified data collection and harvesting data from existing systems as major challenges.

For example, two transit agencies interviewed described issues with their APC systems that limited the value of their APC data. In one case, there were issues with both quality and coverage, as not all vehicles had APC systems and there were inaccuracies in the data. As a result, they were hesitant to share this data. At the time of the interview, they were upgrading to an APC system that would produce higher quality data and would outfit all vehicles with passenger counting systems. The other agency recounted a similar story in which a previous APC system had provided data with false precision. They opted to commission a new system that now provides accurate and precise data.

Another data type that appears to present analysis challenges is maintenance data. There is little research on public transit system and vehicle maintenance in the literature and the transit agencies
interviewed revealed little activity around maintenance data. Likely, this data is not collected in a format that is easy-to-use and as a result it is not a data type that is widely distributed.

For smart card data, there are several data collection factors that impact the potential for analysis and therefore can impact the value of this data for both internal and external use. First, traditional smart card systems were “card centric” with data stored on the cards themselves. This data was notoriously inaccurate and unreliable. Now, many transit systems are moving to account-based systems, with data stored on a server. Several of the transit agencies interviewed are in the process of developing account-based smart card fare payment systems, but many transit agencies, including large systems, do not have account-based smart card data.

Another challenge with smart card data collection is that it is an example of a data type in which the ultimate value of the data depends on user participation. If only a small share of customers pay with smart cards, little insight can be drawn from the data. The use of smart cards has been incentivized in many systems through fare reductions or volume rebates. In addition, smart card users are sometimes eligible for discounts or loyalty programs at commercial partners (Pelletier et al., 2011). Still, two of the transit agencies interviewed noted that smart card usage was very limited in their systems. In one case, smart cards had been rolled out only to specific audiences through partnerships with universities and businesses. In another case, a large share of users have passes that do not require any tap on or off, so smart card data is available for only about 50 percent of bus riders. Even in systems with high shares of smart card usage, there may be biases in which people are left out (see side bar on data biases).

Two other factors impact the value of smart card data. First, many systems require customers to tap on to vehicles or in to fare gates, but do not require tapping when passengers exit vehicles or leave stations. This means that destination stops or stations are not recorded in the data. The decision against “tap off” systems is typically made for a variety of reasons, including the fare collection structure and the ease of use of the

---

**Key Issues for Data Collection:**

- **Data Quality.** Is the data sufficiently accurate and precise for its potential applications?
- **Data Coverage.** Does the data have sufficient coverage to draw insights? Are there biases in which people or vehicles are included?
- **Data Ownership.** Who owns the data? If data pertains to individuals, is there a mechanism to get their permission to use the data?

---

**Data Biases**

The value of data can be limited by biases in terms of which data exists and which data is missing. In the mobility data field, this issue is discussed when app or GPS data from smartphones is used. This data excludes information on people who do not own or use smartphones, which may disproportionately include specific population groups such as low income and older people (Windmiller et al., 2014). Similar data biases can occur in transit agency smart card data. In most systems, not all passengers use smart cards and it is important to identify which passenger demographics are more likely to use smart cards as well as the trip types more likely to be paid for with smart cards (Erhardt, 2016).
system by passengers. Transit agencies should weigh these tradeoffs against data analysis benefits when they design or update their systems.

The other factor that impacts the value of smart card data is the level of privacy protections implemented. In almost all cases, smart card systems track individual card IDs to ensure that passes and discounts are applied appropriately. However, not all systems store this information for analytical use. In some systems, the data stored for internal use by the transit agency has a new ID for each trip or for each day, preventing the tracking of smart cards across trips (in the former case) or across multiple days (in the latter). While these measures protect individual privacy, they limit the potential for analysis. In many systems, persistent encrypted IDs are stored for internal use, with precautions to preserve privacy taken when the data is shared externally. This shifts the decision about privacy protection to the data censoring phase.

**Who owns the data?**

Even when good data is collected, a final hurdle for transit agencies can be data ownership. This issue arises when transit agencies partner with private companies to provide services, in which case the private partner may not be required to turn over data to the transit agency. This is discussed in Chapter 5 on accessing external data. In other cases, vendors that install and maintain systems such as AFC, APC, or AVL may retain ownership of the data generated. Transit agencies must be careful to consider the potential value of data sources and ensure that they have ownership of valuable data. While there has been a shift in ownership of AVL and AFC systems to transit agencies, this issue may still persist for other data types. For example, if maintenance is outsourced, some maintenance data may be owned by the maintenance company rather than the transit agency.

The issue of data ownership is further complicated in the context of data on individuals. In the European Union, General Data Protection Regulation (GDPR) rules guarantee individuals’ ownership of their own data (See Section 4.2). While these rules do not apply in the United States, transit agencies should consider the possibility that laws around individual data will change and specifically consider mechanisms in which individuals can give the transit agency permission to use their data as part of the data collection process.

**Data Cleaning**

Once data is collected, it must be cleaned (to check for anomalous or inconsistent data and correct or delete erroneous information). In the cleaning process, transit agencies identify data quality issues like the APC accuracy problems described above. There may be missing data, miscoded data, issues with clock alignment, and a variety of other problems to address.

Data cleaning requires staff time, and transit agencies must judge how much effort to devote to data cleaning. Accurate data is more important for some data types and applications than others. Clean, accurate data is critical for good customer information. If the route, schedule, and vehicle location information that transit agencies share is inaccurate, it may dissuade customers from using the system, and could even have implications for customers’ safety. In interviews, transit agencies emphasized the importance of vehicle arrival prediction quality and described issues such as “ghost buses”—in which bus arrivals are predicted but do not occur—that they are actively working to combat.
This same level of data quality is not necessary for all data types. Of the agencies interviewed, the two that release the most data publicly both noted that data does not have to be perfect to be released. They see benefits from releasing data even if it has minor flaws. As long as the issues and caveats are described in the data documentation, releasing data promotes transparency and can spur research and innovation. In some cases, external data users can actually help the transit agency identify and fix problems with the data.

**Data Processing**

Once generated, data does not have inherent value. It must be processed and often must be connected to other information to produce value (Hemerly, 2013). Transit agencies make decisions about which data to process. Common examples include merging APC and farebox data, connecting data to GTFS identifiers, and assigning data to trips or vehicles. These processes make data easier to use and increase its value both internally and externally.

The transit agencies interviewed noted that these processes require significant effort and sometimes are not prioritized. As an example, one transit agency noted that they do not assign boarding data to trips and hypothesized that this effort had not been prioritized because it does not fall clearly under the responsibility of a single staff person or group. In contrast, other transit agencies noted that they undergo this effort for their own strategic initiatives and analysis, and therefore did not have to undergo additional processing effort to share the data.

Yet even the agencies that perform significant analytics internally draw the line somewhere. There are countless data sets that transit data could be linked to, including public data sets such as census data and weather data. Transit agencies could opt to produce linked data sets, which would likely provide additional value to researchers and other data users, such as real estate developers and journalists. However, the transit agencies interviewed did not tend to go to these lengths, at least not specifically for the purposes of publishing data. Transit agencies opt instead to publish data and allow external users to link it to other data sets.

Other data processing takes place in response to specific data requests. This may include simple tasks such as aggregating data to the geographies or time periods requested, or more involved tasks that involve joining data sources together. As an example, one transit agency received requests for bus emissions data. As this data had not been used internally, processing it into a useful format required significant effort. One transit agency noted the importance of making data sets easy to use before sharing, for example by adding GTFS identifiers to other data types. This helps protect against misuse of the data.

**Data Aggregation**

Data aggregation refers to any process in which individual records are combined to produce summary data, for example combining individual boarding or origin-destination data to provide estimates of average weekly ridership on a route. The transit agencies interviewed reported that

---

**Key Question for Data Cleaning:**

How clean does the data need to be?

Answer: It depends on the application.
they aggregate data for a variety of reasons, including making data easier to use and understand (particularly for non-technical audiences), minimizing data storage needs, and protecting individuals’ privacy.

Aggregation is an important tool given the variety of audiences for transit data. One of the transit agencies noted that different audiences are interested in different levels of aggregation. While researchers typically prefer disaggregate data, journalists, advertisers, and real estate developers typically seek some level of aggregation so they can draw conclusions and make decisions about actions to be taken, without having to perform a significant amount of analysis themselves. Providing aggregated statistics on things like ridership, on-time performance, and vehicle crowding can also prevent some types of data misuse, in which external users misunderstand aspects of the data and perform analysis that leads to incorrect conclusions. However, more detailed disaggregate data, when analyzed correctly, can spur research that generates new insights that can benefit the transit agency. Some transit agencies, including one of the transit agencies interviewed, provides both disaggregate data for download and an interactive dashboard that allows the user to view aggregated information, with data grouped by time period and route.

In interviews, transit agencies also discussed the challenges of data storage. For example, one of the transit agencies interviewed explained that they are grappling with the best procedure for storing origin-destination data long term. They are weighing trade-offs, including storage needs, when deciding how long to keep disaggregated data and what level of aggregation to use for long term storage. In addition to storage costs, the transit agency is also weighing the privacy risks of maintaining individual records, particularly the risk of a cyber-attack on stored data. These costs and risks are being evaluated against the impact of aggregation on analysis potential. For example, aggregation may prevent analysis of how individuals’ behavior changes over time. As a compromise, some transit agencies maintain individual records for a small sample of IDs. Transit agencies are trying to understand the best way to navigate these challenges in the fast-changing world of data storage (scalability of cloud services) and data usage.

The transit agency interviews also revealed many examples of data aggregation for privacy protection. With a few exceptions, transit agencies generally aggregate individual records by geographic zone, time period, or both prior to sharing.

### Data Censoring

Even with aggregation, transit agencies recognize the potential for small sample bins to compromise privacy. Several transit agencies interviewed have developed rules for minimum sample size (varying from 5 to 10 individuals), under which level values are suppressed prior to data sharing. Data

---

**Why Aggregate?**

*For internal use -*

- Aggregation reduces data storage needs and protects against cyber-attacks of individual records

*For external use -*

- For some audiences, aggregation helps them understand the data and prevents misuse
- Aggregation of individual records prior to sharing can protect individuals’ privacy
Censoring may also be required for data that consist of written descriptions. For example, one transit agency noted that while most of the information in their incident reports was likely not sensitive and could be released for transparency purposes, some reports may occasionally contain descriptions of individuals that present a privacy risk. As a result, sharing this data would require significant effort to review the data and scrub any sensitive information.

**Data Standardization**

In many cases, data reusability and value are promoted by data standardization. Data standardization across transit agencies can enable external partners to repeat analyses for multiple transit agencies with limited additional effort. This can encourage private companies and researchers to develop standard tools that can benefit transit agencies. Standardization is highlighted as a salient need in the transit data industry (Sanchez-Martinez and Munizaga, 2016), but also a major challenge, with data formats varying significantly across organizations.

In public transit data, GTFS is the noted outlier, a standard format for route and schedule information that is widely used across transit agencies. This has allowed app developers to use data across agencies to provide customer information (Schweiger, 2015). A newer standard, GTFS Real-time (GTFS-RT), attempts to do the same for real-time vehicle location data.

The National Transit Database is another example of transit data that is both standardized and consolidated. To date, smart card data has not been standardized in the same way. Two organizations – the Integrated Transport Smartcard Association and the Secure Technology Alliance (formerly the Smart Card Alliance) in the United States – have developed standards for interoperability of smart cards, but these standards are focused on secure fare collection, not on data generation and formatting.

Transit Intelligent Transportation Systems (ITS) Data Exchange Specification (TIDES) and GTFS-ride are two projects around developing standards for transit ridership data from passenger counters and fare collection. Still in their early stages, these standards look to support tools and applications for transit analysis.

A major challenge with data standards is adoption. Transit agencies and/or the vendors they employ need to cooperate, and this requires effort from one or both parties to convert existing systems to meet new standards. Generally, standards are adopted either when there is a clear benefit (the proverbial carrot), or when their adoption is mandated (the stick). As an example, transit agencies quickly adopted GTFS because it allowed their information to be displayed in apps their customers were using. In contrast, transit agencies submit standardized data to the NTD because it is a requirement to receive funding.

**How do Data Standards Get Adopted?**

1. Good standards require champions and resources to support a standards-making activity, including respected experts.
2. Well-designed standards have few optional fields and can evolve over time.
3. For adoption, there needs to be either a clear benefit or a mandate.
A standard needs champions as well as resources. Resources are key to supporting a standards-making activity. For traction, the activity should include experts who are respected within the industry. Well-developed standards minimize the number of optional fields, which limit the usefulness of the standards. The standards-making activity should include testing, and a certification system may need to be developed to evaluate compliance with the standard. Good standards can evolve over time. As an example, GTFS has limited ability to describe fares, but has the potential to be extended to handle more complex fare policies (Wang, 2014).

In summary, the process of data standardization depends not just on an individual transit agency’s technical and organizational ability to apply standards, but also on a strong coalition that has built effective, flexible, and respected standards, and on motivational “carrots” or “sticks” to promote the standard’s adoption. The majority of transit agencies interviewed recognized the need for more data standards but felt that external organizations or regulators would be required to implement them.

**Data Documentation**

Data documentation is important for internal use, ensuring that data is used appropriately across groups and staff changes. Data users need to know field definitions, as well as any assumptions and caveats. This process is especially important when data is shared externally. All the transit agencies interviewed provided some documentation with the data they share. Some expressed that this process can require significant effort and that it is sometimes a challenge to determine what level of detail of documentation is sufficient. Good documentation can help prevent misinterpretation and misuse of data but takes time to develop. An additional benefit of data standards is that transit agencies can rely on centralized documentation of data following the standard format.

**Data Cataloguing**

Not all transit agencies have data catalogues. In fact, most of the transit agencies interviewed noted that their agency does not have a centralized data repository, and that data was stored in a variety of locations across the agency. The advantages and need for a centralized data catalogue were explained in several transit agency interviews. A centralized data catalogue can serve both internal data analysis and data sharing. Parts of the data catalog may be made open to the public, with access to other parts granted to certain partners or limited to transit agency staff. The catalogue can help internal staff find and use data collected across divisions and can also ease the data sharing process, saving the transit agency time responding to data requests.

One of the transit agencies interviewed developed a public-facing dashboard where users can view and download many types of data. The interviewee noted that the dashboard saves time responding to internal data requests as well, because people from other divisions can “help themselves” to data. Having such a catalog requires staff effort to maintain. Many of the transit agencies interviewed noted that the lack of a staff member or group dedicated to such an effort was the reason they did not have a catalogue.
4 SHARING OF PUBLIC TRANSIT DATA

With the wide variety of data types that transit agencies collect, and the often resource and time-intensive processes required to prepare data for sharing, transit agencies are faced with decisions about which data to share with which audiences. The first part of this chapter describes different models for transit data sharing, identified both from the review of literature and web sources and from the transit agency interviews. Then, the second part of the chapter describes the factors that transit agencies consider when choosing whether to share data and which data sharing model to employ. These are factors that were either reported in the transit agency interviews or noted in the literature on data sharing.

4.1 Models for Public Transit Data Sharing

Broadly, models for sharing data can be classified as public or private. Public models, often classed as open data, make data available to everyone, typically by publishing it online. In private models, data is shared with an individual, institution, or group of individuals or institutions. Much of the time, this data is shared under a data agreement, often including clauses about non-disclosure, preventing the data from being shared more widely. Within each of these classes of models there is considerable variation, as described in the sections that follow.

Public Data Sharing

Public data sharing is prevalent among transit agencies. In a 2015 study that surveyed 67 transit agencies, 83 percent provided open data (Schweiger, 2015). All the transit agencies interviewed for this study publish data on their websites. Sharing data openly promotes transparency and can spur innovative use of their data. Additionally, public data sharing was touted by transit agencies in interviews for its efficiency. Transit agencies receive large volumes of public records and other data requests, and publishing the data online allows transit agency staff to quickly point requesters to the online portal rather than having to fulfill requests individually. Data users may also download data directly from transit agency websites without interacting with the transit agency at all. While there is an upfront cost to putting data online and providing necessary documentation, transit agencies believe putting frequently-requested, non-confidential data online saves them staff time in the long term.

On the other hand, posting data publicly generally means that public transit agencies cede control over how the data is used. In general, transit agencies do not publicly share any data containing records that relate to individuals in disaggregate form because of the potential for privacy violations.
Transit Data Types That Are Shared Online

The data that transit agencies publish online includes route and schedule information, system alerts, and the real-time location of transit vehicles. In most cases, route and schedule information is published in the standard General Transit Feed Specification (GTFS) format, and vehicle location data follows the GTFS Real-Time (GTFS-RT) standard. According to the 2015 study, the most common examples of open data among transit agencies are route and schedule information and vehicle location feeds (Schweiger, 2015).

In addition to route, schedule, and vehicle location data, many of the transit agencies interviewed for this study publish information on performance indicators, including route or line level ridership, passenger counts at bus stops and train stations, on-time performance, and reliability indicators. Transit agencies also provide summaries of survey data, including travel surveys and customer satisfaction surveys. Finally, at least one of the transit agencies interviewed publishes financial data.

Figure 4 summarizes the types of data that transit agencies share according to our interviews of 11 transit agencies of different sizes. It refers specifically to the data transit agencies provide on their websites and not to data shared through formal reporting processes such as the National Transit Database.

Figure 4 Online Data Sharing Reported in Transit Agency Interviews

Online Sharing Formats: Reports, Repositories, Dashboards, and APIs

Transit agencies share data publicly in a variety of forms. All the transit agencies interviewed have data and reports that can be downloaded from their websites. In addition, two have interactive dashboards that allow users to interact with the data in a controlled way. Most of the transit agencies share route, schedule, and vehicle location data using an Application Programming Interface (API), which is essentially a set of methods for retrieving data that makes it easy for developers to use the data.
These formats have advantages and disadvantages, as shown in Figure 5. While static reports are easy for all audiences to understand, they do not allow researchers and innovators to manipulate the data, which can limit new insights that could be drawn from the underlying data. On the other hand, static reports protect against data misuse as the analysis is performed by transit agency staff.

Interactive dashboards typically also limit the chances of misinterpretation of data because they allow for only controlled data manipulation. For example, one of the transit agencies interviewed has a dashboard that allows users to look at service reliability and ridership for a specific route, date, and period of the day. Backend calculations are programmed by the transit agency, preventing incorrect analysis of the underlying data. Dashboards can be a very convenient way for people with a wide range of technical abilities to interact with transit agency data. Of course, if the underlying data is not provided, dashboards do not promote new analysis. For example, while the dashboard described allows the user to retrieve reliability information, defined as the on-time percentage for a route (for low-frequency service), the underlying data would enable a researcher to answer more detailed questions about the extent and patterns of schedule deviation. In addition, developing an interactive dashboard requires significant effort on the part of the transit agency. In some cases, third parties have produced public dashboards based on open data. For example, the Bus Turnaround Coalition developed a dashboard that reports on the performance of New York City’s bus routes.

![Figure 5 Methods for Sharing Data Online](image)

Most of the transit agencies interviewed have a developer website designed for use by software developers. These sites house the transit agency’s API. Developers use these interfaces to access data that is then provided to customers in travel planning and real-time information apps. Some of the transit agencies interviewed require the users of this data to register to access an API key. At least two of the transit agencies interviewed noted that this model enables them to cut off users who overburden the system with too many data requests. Travel planning and real-time information
apps have become a key source of transit customer information. As such, transit agencies are reevaluating the best way to leverage their route, schedule, and vehicle location data to provide customers with information. Section 4.3 discusses this debate in detail.

Many transit agencies use a combination of the mechanisms in Figure 5 to share data and information publicly. For example, the MBTA’s interactive performance dashboard allows users to select specific lines, dates, and periods when viewing reliability and ridership data. In addition, the underlying data is available for download. The agency also has a developer API and publishes reports that share transit agency insights and analysis of their data.

**Cross-Agency Data Sharing Platforms**

In some cases, multiple public transit agencies share data with an organization that shares the data using a centralized platform. Such platforms typically require a standardized data format and may therefore require additional effort from transit agencies. Conversely, such models can allow transit agencies to share costs of processing, storing, and documenting data, as well as addressing any legal implications of data sharing, potentially reducing transit agency effort.

The Federal Transit Administration’s (FTA’s) National Transit Database (NTD) is an example of a cross-agency data repository. Recipients of FTA grants are required to submit system, ridership, and financial data to the database. The standardized format makes it highly accessible to researchers, who produce studies that can benefit transit agencies.

The National Association of City Transportation Officials (NACTO), which includes both cities and transit agencies, launched the SharedStreets initiative in 2018 (NACTO, 2018). SharedStreets is a data platform in which public entities and private companies can provide data on physical infrastructure and vehicle activity. SharedStreets highlights four core functions of its collaborative platform: (1) to standardize data on physical infrastructure and vehicle activity, (2) to build open source tools to use data, (3) to anonymize sensitive data on individuals, and (4) to establish a foundation of collaboration and trust (SharedStreets, 2018). These functions illustrate the potential benefit of cross-agency platforms to transit agencies. Efforts to standardize and anonymize data can be shared across agencies, and open source tools that operate on standardized data can provide value back to these transit agencies.

These initiatives also represent a way for transit agencies to access private data. Uber, Lyft, and Ford Motor Company are all involved in the SharedStreets initiative. Chapter 5 discusses transit agency access to external data sources in more detail.

**Terms of Use for Public Data**

The inclusion of terms of use or legal notices with open data provided online varies among transit agencies, including the transit agencies interviewed. For example, the performance data that powers the MBTA performance dashboard can be downloaded without agreeing to any terms of use. Houston Metro, however, includes a legal notice with its data downloads indicating that METRO retains ownership of the data and that the data is provided “As Is.” TransLink has similar Terms of Use for their Open API data (https://developer.translink.ca/Home/TermsOfUse).
Public Data Sharing Challenges

The transit agencies interviewed identified two key challenges that they face in providing data online. The first is that different audiences want different level of detail. Researchers are typically interested in raw, disaggregate data, while journalists or real estate developers may desire more aggregated data. For some data types, disaggregate data poses privacy risks, while for other types, this is not the case. Transit agencies must choose an appropriate level of data aggregation for each data set that is shared publicly. Some transit agencies provide both raw and aggregated data. Adequately documenting this data to ensure different audiences understand how to use it requires significant effort from transit agency staff.

Another challenge with open data is that transit agencies do not have control over how the data they publish is used. Several transit agencies expressed concerns that data would be misunderstood and used incorrectly. Even if used correctly, public data sharing models do not give transit agencies direct control over whether data is used to produce analysis and tools that support the transit agency’s needs and goals. One innovative solution to this challenge was a data competition hosted by one of the transit agency’s interviewed. Not only did this bring attention to the transit agency’s open data, but it encouraged participants to compete to answer questions posed by the transit agency. As already noted, several transit agencies are also considering different methods to ensure the route, schedule, and vehicle location data they publish is used to best serve their customers (See Section 4.3).

Private Data Sharing

Some data is more sensitive but still sharable under the right conditions. Or, other data is not commonly requested and therefore has not been published to the transit agency’s website but can be shared when requested. To fulfill these needs, all the transit agencies we interviewed share data directly with partners and data requesters. This category of data sharing includes data sharing with partners such as research institutions, municipalities, and private sector contractors or real estate developers and fulfillment of public records requests. When sharing sensitive data with partners, transit agencies may require non-disclosure agreements and training prior to releasing the data.

Research Partnerships

Most transit agencies interviewed described the importance of sharing data with researchers. Five of the transit agencies interviewed have strong partnerships with a specific university or research

Opt-In Models

Researchers often want to access individual records, which some transit agencies are hesitant to share due to privacy concerns. There may be potential to address this challenge with opt-in models, in which individuals agree to share their data (UITP, 2018). One of the transit agencies interviewed discussed this option for accessing users’ wi-fi and app usage data. If transit agencies can show customers that they will use data to the customers’ benefit and establish trust with their customer bases, opt-in models can allow transit agencies to maintain sensitive data internally and use it for planning purposes. In some cases, transit customers may even opt in to sharing their data with external trusted partners such as researchers and municipalities if they are made aware of potential benefits.
institute. In at least two cases, the transit agency pays the university to complete research that aligns with the transit agency’s needs. In other cases, transit agencies do not pay the university, but nonetheless have a long-standing collaborative relationship that allows them to shape their partners’ research agendas. In some cases, transit agencies are willing to share sensitive data, including individual customer records, with researchers in these universities, after the researchers are trained and have signed non-disclosure agreements. Researchers are often able to spend time on questions that transit agencies are unable to, providing significant benefits to the transit agency. Long-standing partnerships allow the transit agency to have a standing data agreement and established trust in the researchers.

Public Records Requests

All transit agencies interviewed complete public records requests. While different states have different open records acts, they generally impact transit agencies in similar ways. As public agencies, transit agencies are required to share data that is requested as long as it does not violate their customers’ privacy. Some transit agencies also specified that they would not release data when it may pose security concerns. In some cases, transit agencies said that their lawyers review requests if privacy is a concern. Transit agencies also discussed the use of rule-of-thumb rules to protect privacy. For example, one agency does not share data that contains records that pinpoint the location of fewer than 10 individuals in a given time period. Privacy concerns are discussed in more detail in Section 4.3, which covers the risks of data sharing. Most transit agencies interviewed fulfill public records requests free of charge. However, at least one transit agency noted that they have the option to charge requesters for the staff time required to complete the request if the data exceeds a certain threshold. Several transit agencies also described that if a specific type of data was requested frequently, they opted to put it on their website to save time in the long run.

Other Data Sharing Relationships

At least two of the transit agencies interviewed shared data with local municipalities. In addition, several of the transit agencies provide data to real estate developers when requested. One transit agency specifically mentioned advertisers as a recipient of data.

Another private data sharing model is followed by the American Bus Benchmarking Group. Established in April 2011, the group consists of mid-sized bus organizations in the United States. Members share performance data which can be accessed confidentially by members of the group. The objective is to establish benchmarks that can help members understand their agency’s performance and identify best practices to improve performance (American Bus Benchmarking Group, 2019).
Exchange of Funds

Almost all the transit agencies interviewed stated that they had not sold data and were not considering selling their data. One transit agency noted that they could charge requestors for the effort of fulfilling public records requests. This agency also noted that they had received some payment for data provided to a public entity. Another agency stated that they occasional sell analysis of their data. In general, transit agencies do not see much potential in receiving funds directly for their data. Most transit agencies noted that because they are required to provide data in response to public records requests, they could not expect to charge for it.

Given the expectation of free and open data, public agencies have seen push back when they attempt to sell data. Dutch agencies attempted to provide release data to some partners for free while selling to others, which led to conflicts (Conradie & Choenni, 2014). The tone of news stories about Transport for London’s potential financial gain from Wi-Fi data similarly suggested that public agencies profiting off of their data can provoke a negative response (Cheshire, 2017).

Terms of Use for Private Data Sharing

As noted, sensitive data shared with a partner is typically accompanied by a non-disclosure agreement, according to the transit agency staff interviewed. In addition, at least two transit agencies require any publications from their academic partners to be reviewed by the agency prior to publication.

On the other hand, for non-sensitive data shared through public records requests, it is common to provide it without any provisions for use. At least two of the transit agencies noted that they do not attach any provisions when they fulfill non-sensitive data requests, which may include aggregate statistics on ridership or boardings and system-level information.

Private Data Sharing Challenges

In some cases, private data sharing agreements are initiated by the transit agency, which seeks out a third party to perform certain research or analysis tasks. In that case, transit agencies may struggle to find a good partner. In other cases, individuals or organizations approach the transit agency with data requests. Many transit agencies described the fact that data requests may be sent to a variety of divisions within the agency and are handled by different staff depending on the data type. Only one transit agency noted that they had created an information management and governance group to handle all outside data requests. Due in part to these organizational challenges, transit agencies interviewed noted that responding to individual data requests can be resource intensive.
### 4.2 Factors Impacting Transit Agency Decisions About Data Sharing

When transit agencies decide whether to share their data, who to share it with, and which model to use to share it, they consider several factors. For many transit agencies, the first factor to consider is any legally binding rules. Some data sharing is required. Other data cannot be shared to protect the privacy of the individuals it pertains to. Even when specific laws may not apply, transit agencies seek to protect the privacy of their customers. The risk of privacy violation is one of the risk factors transit agencies evaluate when deciding about data sharing. The concept of a risk assessment is introduced in the International Association of Public Transport (UITP’s) guide on public transit data sharing. In addition to privacy violation, risk factors include reputational risks from misuse of data and security concerns.

After considering legal requirements and risks, transit agencies can think of decisions about data sharing in terms of benefits and costs. According to the UITP report, one question agencies should ask is, “Which types of data could improve customer experience or create efficiency gains?” This question asks the evaluator to consider the potential benefits of sharing data. The report also recommends consideration of the costs to produce, store, and maintain data.
In a paper on public policy considerations around data, Hemerly (2013) cautioned that technology has been developing more quickly than legislation to keep up with it, which can lead to concerns and conflicts over what kind of data is public and what data is private. In the area of transit data, there are two recent sources of legal advice for data: the TCRP Legal Research Digest published a summary of legal issues related to transit agency use of electronic customer data (Thomas, 2017) and a discussion of the legal implications of video surveillance on transit systems (Thomas, 2018).

**Figure 6 Framework for data sharing decisions**

**Rules and Legal Issues**

In a paper on public policy considerations around data, Hemerly (2013) cautioned that technology has been developing more quickly than legislation to keep up with it, which can lead to concerns and conflicts over what kind of data is public and what data is private. In the area of transit data, there are two recent sources of legal advice for data: the TCRP Legal Research Digest published a summary of legal issues related to transit agency use of electronic customer data (Thomas, 2017) and a discussion of the legal implications of video surveillance on transit systems (Thomas, 2018).
Both of these reports focus on internal maintenance and use of these data types, but there are implications for data sharing. Specifically, both electronic fare transaction records and video surveillance have the potential to identify individuals, presenting privacy violation concerns. The reports cite various rulings that may apply in cases of privacy violation involving these data sources and note that a decision that impacts the way transit agencies manage these data sets may eventually be handed down by the Supreme Court. At present, some states have guidelines about the maintenance of electronic records by government agencies, and a few states have laws specifically about sharing video data that require documentation of the “custodians” with whom data is shared.

One transit agency interviewed pointed out that laws can appear arbitrary or out of date. A law in their state exempts smart card data from freedom of information act requests on the basis that it contains individual records. However, data from smartphone apps, which the agency is planning to collect, will not be protected from release under the same law.

In the European Union, the General Data Protection Regulation (GDPR), approved in 2016 and enforced beginning in 2018 defines a comprehensive set of regulations around privacy. It requires that consent and terms are clear, transparent, and written in easily-understood language. It specifies that individuals own their data and must be informed if their data is being transferred to another party or if there had been a data breach.

In addition to laws and rulings that may prevent data sharing, there are laws that mandate data sharing by government agencies. All transit agencies interviewed fulfill public records requests, as mandated by state laws. However, the transit agencies also noted that they refuse to fulfill some public records requests if they determine that the requests violate individuals’ privacy or pose a security risk. At least two of the transit agencies specified that they have attorneys on staff who review requests when these concerns arise.

Risks

The primary risks that may impact public transit data sharing decisions are: privacy, security, data misuse, and strategic risks. Each of these four classes of risk is discussed below, with an understanding that risk assessment needs to consider both the likelihood or risks occurring and the extent of the impact of a risk occurs.

Privacy

There are several sources of privacy concerns with public transit data, including:

- Personal data collected, such as registration information associated with fare cards (names, addresses, etc.).
- Anonymized individual data that risks re-identification when combined with other data sets.
- Anonymized individual data that risks re-identification even without additional data sets. As public transits increasing integrate their electronic fare systems with other modes and payment systems (such as credit cards) re-identification becomes increasingly possible.
- Facial recognition of video data.
Examples of re-identification of anonymized data occur across fields. In 2008, Netflix released data on movie ratings by individuals that they believed had been anonymized, but researchers at the University of Texas at Austin proved that they could identify individuals (NAS, 2018). Similarly, when the New York City Taxi and Limousine commission released data on taxi rides in 2014, a data scientist was able to identify individual trip origin and destinations and amount paid by combining the data set with medallion numbers visible in celebrity photographs (Lubarsky, 2017). In interviews, transit agencies expressed the need for guidance and protocols to follow to assess and reduce privacy risks.

**How important is privacy?**

In assessing privacy risks, transit agencies may consider how important privacy is to their customers and their customers willingness to provide personal information to public agencies in return for benefits. Studies have shown that people are willing to trade privacy for benefits. According to a 2012 Pew study almost three quarters of smartphone owners get location-based info on their phones. But people also appear to be selective in which sources they provide information to. The study found that more than half of app users surveyed had uninstalled or chosen not to install an app because of privacy concerns (Brakewood and Paaswell, 2017).

In a focus group on transit agency apps, most users said they did not read app privacy policies, though 72% said they understood that their smartphone’s locations could be identified. In a survey on the same subject, most respondents said that transportation apps should know their location (71%) and 60% said they were not concerned about this (Brakewood and Paaswell, 2017).

However, people are concerned with their data being shared, especially if it is not for transportation planning purposes. 50% were “strongly concerned” about having data shared for marketing purposes, compared to only 13% “strongly concerned” about having it shared for transportation purposes. This study also found that while 35% of survey respondents were “strongly concerned” with data from transportation apps being shared with a private agency, only 18% were “strongly concerned” with this data being shared with a public agency.

Understanding these tradeoffs is important, because there is a cost to maintaining data privacy. Erhardt (2016) argues that strong privacy restrictions, such as data obfuscation requirements, can limit the usefulness of smart card data. Lerner (2012) discusses data privacy regulations in the context of online advertising and suggests that they may inhibit innovation, by posing obstacles to start ups and thus favoring established large companies. Transit agencies may be similarly burdened by privacy regulations relative to private mobility providers and private mobility data collectors.

**How much aggregation is necessary to protect individual data?**

The majority of transit agencies interviewed indicated that they never shared individual records. Those that share individual records do so only with trusted partners who sign a non-disclosure agreement and undergo training in the handling of such data. Instead, transit agencies typically opt to aggregate individual records prior to sharing.

Transit agency interviews revealed that the level of aggregation varies. One transit agency indicated that they never release aggregated data with containing fewer than 10 records within a given
sample bin. For example, if a data requester asked for hourly boardings at a stop and the stop had fewer than 10 boardings in an hour period, they would not release data for that hour. Another transit agency described a similar rule but set their minimum at 5 records. A third transit agency interviewed indicated that they only release data aggregated to a census tract or TAZ level. Yet another agency specified that they only supply average daily boarding information, typically aggregated to an entire year. They noted that this choice was not only for privacy reasons but due to lack of consistent data and data quality concerns.

Multiple transit agencies commented that their privacy policies felt arbitrary and that guidance on privacy protection would be appreciated.

**Strategies and Lessons Learned Around Data Privacy**

Transit agencies have developed a variety of techniques to address privacy concerns when they collect and share data. One agency discussed the importance of transparency. They noted that they used several methods to inform customers of Wi-Fi data collection in their stations. Notices described that data would be used to benefit customers through improved service planning. Being upfront about data collection helps mitigate the risk of privacy concerns being raised after the fact and establishing the benefit to customers may create buy-in.

Another transit agency specified that they have a privacy officer who reviews data requests with privacy concerns and who also conducted a privacy impact assessment. This organizational structure and proactive approach may also mitigate privacy risks.

Outside of transit, there are other frameworks for privacy that can guide transit agencies. One framework for assessing privacy risk, put forth in the National Academy of Sciences (NAS) 2018 report categorizes data in three tiers: open data, restricted data, and highly restricted data that are collected under a pledge of confidentiality. Open data is data for which privacy concerns do not exist. Restricted data may have privacy concerns associated with it and should only be shared with appropriate provisions. Highly restricted data generally should not be shared, and individuals should be informed of its collection and uses.

The National Center for Health Statistics follows a “Five Safes” framework to guide decisions about data access (NAS, 2018). The “Five Safes” are:

- Safe projects, in which they consider the specific use of data and determine whether it is “appropriate, lawful, ethical, and sensible”;
- Safe people, in which they evaluate the researchers who will be analyzing data;
- Safe data, in which they look at the information contained in the data and evaluate any potential confidentiality breach;
- Safe settings, in which they consider the security of the facilities where data is stored and accessed;
- Safe outputs, which considers what types of findings will be released based on the data analysis and evaluates risks, particularly of re-identification.
There are also technical approaches to privacy protection. One of the transit agencies interviewed described their process of encrypting data using a salt, an unknown character string, that is added to a unique identified prior to encryption. This serves as protection against decryption. This agency, which has a pilot to collect mobile phone data in collaboration with a private company, also noted that they had a process for automatically randomizing data relating to a sample of fewer than 10 individual devices.

Chen et al. (2012) describe the potential of the differential privacy framework, a statistical process for protecting user privacy in data sets consisting of individual user data by adding noise to the datasets. Their case study for the Montreal Transportation System demonstrated that they could successfully apply the differential privacy framework to smart card data, producing a privacy-protected data set from which the agency could perform standard analysis tasks.

**Security**

In the context of public transit data sharing, security risks are defined as the risk of someone using transit data to inform an attack on transit infrastructure. Compared to privacy issues, security concerns were not emphasized in the interviews of transit agencies. However, one transit agency noted that they were often prevented from releasing data, for example on stop-level boardings, that was deemed security sensitive. Another transit agency noted that they only release data if the requester can demonstrate a research or business need for the data. If not, they infer that the request produces a security concern. Security was not mentioned in the other transit agencies interviews with the exception of one interviewee who specifically noted that their agency does not perform a security risk assessment for data requests.

The lack of emphasis on security risks likely stems from the fact that the likelihood of an infrastructure attack that relies on transit data is low. There are no past examples of attacks that depended on transit data.

**Misuse**

While security was discussed only occasionally in interviews of transit agencies, the risk of data misuse was raised in almost every interview. Misuse may be deliberate or accidental, with most transit agencies more concerned with accidental misuse, which they perceive as much more likely. One interviewee noted that data users often do not have the full picture. Because they see only part of the data, they may make incorrect conclusions. Another interviewee noted that they were concerned that users would select the wrong data source or use old, stale data to drive their analysis. One transit agency described an example in which a third-party app misrepresented the data the transit agency had published, leading to complaints to the transit agency from their customers.

Information about transit agencies that is relayed to customers through apps or published on websites and in newspapers can significantly impact the way customers view transit agencies. While transit agencies cannot prevent misuse of published data, they can take steps to prevent it. The transit agencies interviewed noted the importance of checking data before it is published and of fully documenting data that is published online or provided to partners. In terms of route, schedule
and vehicle arrival data shared with customers through third-party transit apps, several transit agencies are taking steps to actively manage what information is shared (See Section 4.3).

**Strategic Risks**

The final set of risks are strategic. These risks comprise any consequences of data sharing that impact the transit agency’s ability to serve its function. For example, if data sharing can impact the way the transit agency is perceived by its customers or its ability to provide good service to its customers, there is a strategic risk.

In the transit agency interviews, several interviewees described concerns about public perception. Particularly when asked about the possibility of selling data, they noted that this could cause them to lose their customers’ trust. There are also varying approaches for viewing the strategic impacts of open data. One interviewee commented that some transit agencies are concerned with releasing data that shows things like poor on-time performance or over-crowding on their system. In contrast, the interviewee believed that releasing data promotes transparency and provides their customers with the best information available to navigate the system. In short, there may be strategic risks associated both with releasing and with not releasing data.

The UITP identifies a different set of strategic risks in their guidance document on data sharing. They discuss that there may be a strategic risk of sharing data of high commercial value for free (UITP, 2018). Their report hypothesizes that when certain data sets are shared, it may actually cause power to shift away from the transit agency. As an example, when transit agencies share GTFS and GTFS-RT data and it is used by third-party apps, the third-party apps collect information on customers that the transit agencies may not have access to. This information asymmetry may disadvantage transit agencies and hamper their ability to best-serve their customers. This particular instance of information asymmetry was discussed in several of the transit agency interviews and is discussed in more detail in Section 4.3.

**Benefits**

The benefits associated with sharing data are wide ranging and can be difficult to quantify. In the transit agency interviews, interviewees frequently commented on the need for methods to assess the value of data and particularly the value of sharing data.

**Innovation Outside the Agency**

In a report on the value of data, Abella et al. (2017) identify the ability of data to spur innovation as a key benefit. The innovation impact of open data was pinpointed in the context of public transit route, schedule, and vehicle location data in TCRP Synthesis 115 (Schweiger, 2015) and also in a report on the value of Transport for London’s open data (Deloitte, 2017). Across the United States and abroad, private developers have responded to open GTFS and GTFS-RT data by developing travel apps the provide trip planning and vehicle arrival information to customers. In the transit agency interviews, interviewees commented on the ability of external partners to innovate in quick-changing contexts, such as app development.
A study of Transport for London’s open data estimated that the gross value added to the economy from companies that develop apps using Transport for London’s open data is between £12 and £15 million and directly supports approximately 500 jobs (Deloitte, 2017).

Open, standardized GTFS data has also led to the development of additional open-source resources, including products such as OpenTransit Indicators, which calculates performance indicators from GTFS data, and TransitWand, a tool for collecting GTFS in the field (Lawson, 2016).

In addition to innovative products, data sharing supports public transit research. Nearly all the transit agencies interviewed for this study also discussed the benefit of external research conducted using their data. Five transit agencies discussed long-term partnerships with academic institutions in which they could work with researchers to ensure that academic work would benefit the transit agency, while also meeting research goals of the institutions. Interviewees noted that within the transit agency there often is not time to focus on research-oriented questions. Transit agencies could point to specific examples of research conducted by external partners that benefits their transit agency. These include:

- a bus turnaround dashboard
- an origin-destination inference algorithm
- a passenger segmentation model
- an electrification study
- optimization of dispatcher assignment of work

Sharing data can also spur innovation through the combination of public transit data with external data sets. One of the transit agencies interviewed discussed this potential, noting that transit data might be combined with other data sets such as health care or census data to create new insights. For transit agencies that seek to pursue innovative, multi-modal collaborations, some level of data sharing often becomes a necessity.

**Cost Savings**

Even if the work that outside partners contribute is within the capabilities of transit agency staff and could potentially fall into its purview, outsourcing this work through data sharing can save transit agencies money. For example, although some transit agencies have developed their own transit planning and real-time information apps, several of the transit agencies interviewed noted that developing their own apps in-house would be time-consuming and inefficient, compared to allowing external partners to develop them. One transit agency also noted that by opening up data, external users of the data help the transit agency more quickly identify problems with the data sets. Increasing the data user pool saves the transit agency time spent looking for missing data and data anomalies. Clearly, this benefit must be weighed against the risk of releasing data that has not been fully-vetted. However, for some data sets and partners, this may be a useful model.

As discussed in Section 4.1 on data models, cost savings can also be accrued by releasing data publicly in batches, rather than repeatedly releasing data on a case-by-case basis through individual public records requests. Publishing frequently-requested, non-sensitive data online saves transit agency staff time in the long term.
Transparency

According to the transit agency interviews, the general public expects that public agencies should publish data in free and open formats. Recent years have seen an emphasis on transparency in government and public agencies, precipitated from the federal level – a 2009 Office of Management and Budget memo encouraged transparency and prompted local governments to develop open data portals.

Publishing data helps transit agencies meet this transparency goal, which can positively impact public perception. Two of the transit agencies interviewed identified transparency as a reason for sharing data.

Increased Awareness of Transit Services

In addition to transparency, data sharing can serve to publicize the transit agency, and may even encourage citizen engagement (Kassen, 2013). Increased awareness of transit services was identified as one of the benefits of GTFS data sharing, in a recent study (Schweiger, 2015). In one of the transit agency interviews, the interviewee discussed how their open data spurred engaging online content and helped to “build the agency brand”.

Data Reciprocity

In some cases, transit agencies may receive external data in exchange for the data they share. For example, several transit agencies receive data from the travel apps that their GTFS feeds supply (more on this in Section 4.3). The UITP guidance suggests that transit agencies demand data reciprocity in more instances. On a related note, some transit agencies receive data from private mobility partners through service agreements. Some of these cases are document in Chapter 5. However, in the documented cases, the transit agencies do not share their data with the private provider, though there is one instance in which a transit agency and a transportation network company (TNC) provide data to a common third-party consulting firm.

Monetary Gain

As discussed, several transit agencies expressed concerns over the risk of negative perception of data sales. They felt the public would not support the idea of the transit agency profiting off of data on the individuals who use the system. In addition, several transit agencies mentioned that they could not sell data because they are required to provide it to anyone who requests it with a public records request. One transit agency noted that they can charge requesters for the effort required to fulfill these requests, though they generally do not.

Several transit agencies discussed the monetary benefits achieved by using their data to increase advertising revenue. One transit agency noted that they are in the process of estimating the value of advertising in their system, including data-driven, targeted advertising. Two others described how they used ridership data to price space in their system. According to Transport for London’s (TfL’s) 2017-2018 Advertising Report, 20% of UK’s outdoor advertising is owned by TfL. The report also notes that in 2017, TfL provided customer segmentation data to advertisers. This depersonalized and grouped data from smart ticketing was overlaid with Experian Mosaic demographic groups to
help prove to advertisers that they are reaching their target audiences. According to the report, advertising revenues in the fiscal year 2017/2018 were £152.1 million (Transport for London, 2018). With on-going research on and growth of location-based ads, there may be increasing latitude for transit agencies to generate revenue through advertising.

In one transit agency interview, the interviewee noted that their transit agency had occasionally sold bespoke analysis to clients. This analysis consists of specially-requested analysis that would otherwise not be performed by the transit agency. This type of model avoids the privacy risks of sharing data and may alleviate the public perception risks of profiting off of data that is perceived of as a public good. Another potential fund generator discussed in the transit agency interviews was the potential for transit agencies to sell the data infrastructure expertise they developed to share data, particularly the development of APIs that feed large volumes of real-time data to developers.

**Impacts on Customers**

Ultimately, perhaps the most significant benefit that transit agencies consider when sharing data is its potential to positively impact customers. Travel apps that help customers plan public transit trips and alert them to when buses and trains are arriving can save customers time. In some cases, there may be a safety benefit to knowing when transit will arrive. According to a study of open transit data, the primary reason transit agencies have cited for releasing route, schedule, and vehicle location data is to provide customers with more information (Schweiger, 2015). In London, 42 percent of Londoners use mobile phone apps that use information from Transport for London’s open data feed. The report on the value of these open data feeds estimates that the greater certainty and time savings from these apps translates to between £70 and £90 million per year in time savings.

The innovative studies spurred by open data can impact customers as well. Research that helps transit agencies operate more efficiently or plan service better ultimately translates into benefits for public transit customers. Particularly in data sharing models where transit agencies are able to influence external research to target their needs, the impacts of external research on customer may be significant.

**Costs and Effort**

Section 3.2 articulated the many steps in preparing data for sharing. These steps require staff time and often also require contracting with external vendors. In many cases, these steps are required even for internal data use, a factor identified in several transit agency interviews. Interviewees noted that good internal data management practices make data sharing easier. For example, a well-documented internal data repository helps transit agency staff make use of data and also reduces the additional steps require to distribute data. However, several transit agencies noted that much of their data was not collected with analysis in mind, was not stored in a centralized location, and was not documented for external use. As such, preparing data in response to data requests often requires significant effort. In addition, there is additional effort required to conduct privacy and other risk assessments and to develop any licensing agreements necessary.
As discussed in Section 3.2 and in the section on benefits above, standardizing data prior to sharing can produce additional value, for example by encouraging standardized, open source tools, as has been the case with GTFS. However, standardizing data also requires additional effort. Aside from GTFS, there is a lack of standardized formats for transit agency data. But if such standards were developed, transit agencies would have added costs of converting their data into the standard format. The transit agencies interviewed noted that in the case of GTFS there was a clear benefit to putting data into the standardized format, encouraging agencies to dedicate effort to data standardization. With the wide-spread adoption of GTFS, vendors adapted to produce data in the standard format, reducing the burden on transit agencies. While there are efforts underway to produce standards for other transit ITS data, transit agencies expressed concerns that the lack of a clear use case for the standardized data may limit the adoption of these standards.

Challenges and Needs

In the quick-changing data management and sharing environment, transit agencies identified a variety of needs, many of which are also reflected in the literature. These needs are categorized in two groups. Some challenges and needs are internal – these are new protocols, organizational structures, and other changes that are required within the transit agency. At the same time, transit agencies are looking for external guidance and even regulation to govern these internal changes. Transit agencies expressed frustration at the challenge of working out every detail of a data sharing agreement internally. They recognize the potential efficiency that could be gained through standardized protocols and policies. They also see regional or federal policies as a potential mechanism to encourage more cooperation from vendors and private mobility and data providers.

Internal Data Management Structure and Protocols

The majority of the transit agencies interviewed identified a lack of coherent organizational structure for managing data internally as well as for data sharing. Interviewees noted that data was collected and stored across a variety of divisions or groups within the transit agency and responsibilities for data sharing were therefore also spread across staff in different areas of the organization. For example, maintenance staff collect and manage maintenance data, operational staff have operations data, other data is housed by the agency’s information technology (IT) department, and planners access yet another set of ridership and route and schedule data. These data silos present challenges for internal use of data as well as for external sharing of data. Several transit agencies described these challenges in terms of both personnel and organizational needs – the need for a centralized data management staff person or group and technical needs – the need for a centralized data repository and catalog.

These needs are related, as a centralized data repository and catalog requires dedicated staff to develop and maintain. These data-focused staff could also take responsibility for other needs identified by the transit agencies interviewed: the development of formal data sharing policies and protocols, including standard data licensing agreements and an established method for evaluating privacy risks.

As noted in the literature, developing these capabilities among transit agency staff likely requires staff training, particularly in small and medium-sized agencies (Lawson, 2014). As an additional challenge, staff turnover can make it difficult to ensure that progress in data management is
sustainable. Establishing a staff member or team that is dedicated to data management is likely the first step in addressing these challenges, as outlined in Figure 7.

External Data Policies and Standards

Additionally, this internal development process would benefit from external guidance and policies. While all transit agencies identified internal improvements needed for data sharing, they also expressed interest in more external support. This support consists of guidance and regulation. Transit agencies recognize that many data sharing challenges are shared across public agencies. Rather than devoting resources to solve these challenges individually, they seek external guidance around topics including:

- handling sensitive or private data, including when small values need to be suppressed, and what precautions need to be taken to avoid re-identification risks
- Which data licenses should be used
- Appropriate level of data documentation

Transit agencies are also looking to external organizations for the development of data standards. And several of the transit agencies interviewed noted that there may need to be a regulatory push to encourage the adoption of new data standards, particularly to require private vendors to comply (Lawson, 2016).
4.3 Deeper Dive: Sharing Route, Schedule, and Vehicle Location Data and the Customer Experience

Route, schedule, and vehicle location data are among the most commonly shared types of transit data. Ten of the eleven transit agencies interviewed shared this data and according TCRP Synthesis 115, most transit agencies share this data and do so free-of-cost (Schweiger, 2015). This data is typically shared in standardized GTFS and GTFS-RT formats, and this information has a clear value to customers looking to plan transit trips and find out when transit vehicles will arrive.

Across the United States and abroad, private app developers have created apps that use GTFS and GTFS-RT feeds to provide information to customers. In London alone, there are 600 apps powered by public transit open data feeds, which are used by 42% of Londoners (Deloitte, 2017). According to TCRP Synthesis 115, approximately 40% of respondents to an APTA survey have developers using their open data. For large transit agencies, 68% reported that developers use their data. (Schweiger, 2015).

“Riders interact with these apps multiple times daily, making open data the most important customer communication channel agencies offer to the public.”

(TransitCenter, 2018)

These apps have become a key component of how customers interact with transit systems. In many cases, customers are much more likely to receive information about transit services from private apps than directly from the transit agency.

Many transit agencies publish GTFS and GTFS-RT feeds using a developer API, which in turn is used by third party apps. Some transit agencies provide a list of vetted apps on their websites. Even with this safeguard, this hands-off data sharing model counts on developers to provide the best information and user experience to customers and customers to find the best app through the app review process or word-of-mouth.

There are drawbacks to this approach, many of which were highlighted by transit agencies in the interviews conducted. First, some apps may not provide accurate information. Even if all apps use the same data feeds, the algorithms they use to suggest routes and predict vehicle arrivals vary. One transit agency interviewed commented that one commonly used transportation app in their service area often provided inaccurate predictions, likely due to the algorithm used. This is a problem, as transit agencies want to ensure that customers access the most reliable information available.

Transit agencies are also concerned with context. Many apps, including Transit App and Google Maps present information about the cost and travel time of Uber and Lyft on the same screen as
transit information. At least one of the transit agencies interviewed expressed concerns that this presentation of information may encourage users to choose TNCs over public transit.

A hands-off approach to sharing GTFS and GTFS-RT data also means that data on transit customers who use transportation apps is collected by the app developers rather than by the transit agency. As discussed in recent UITP guidance, this can be strategically risky for transit agencies, as they are missing out on information about their customers. This information is accumulating instead to private developers. There is at least one example in which a private app developer, Citymapper, piloted bus service, though in that example, the pilot was in cooperation with the public transit agency, Transport for London.

Increasingly, transit agencies are seeking to take more control over the way their GTFS and GTFS-RT data is presented to transit customers. Given the outsized impact this information may have on the way customers perceive the system, transit agencies are following models that enable them to control some of the information presented. In addition, transit agencies are developing models in which they receive data collected by transportation apps using a variety of mechanisms to facilitate this data transfer.

Figure 8 shows four different models for using route, schedule and vehicle location data to provide customer information that were reported in the transit agency interviews. Transit agencies can control the customer information received through an app in a variety of ways. Many of the transit agencies interviewed simply publish route, schedule and vehicle location data and allow app developers to use it in their apps, following the hands-off approach described above. However, four of the transit agencies interviewed have followed different approaches. The easiest and cheapest option, followed by one transit agency interviewed, is to endorse an existing app. A more resource-intensive option, which one of the transit agencies is in the process of completing, is to commission an app. Finally, for ultimate control, some transit agencies develop apps in house, including two of the transit agencies interviewed. This requires staff with specific technical skills to develop and maintain the app. It is important to note that in the cases observed and documented, those transit agencies that endorse, commission, or develop an app continue to provide the data openly and to allow other apps to use the data.

**Develop an App In-House**

Two of the transit agencies interviewed have their own app, developed in-house. This model gives the transit agency maximum control over the contents of the app and the data extracted from it. As long as the transit agency continues to put resources into maintaining and updating the app, they have flexibility to adapt the app over time, as data and customer needs change.

Such a model is generally only feasible for large transit agencies with sufficient IT and technical staff to devote to the effort. However, even some of the larger transit agencies interviewed opted out of developing their own app because of the specialized and fast-changing nature of app development.
**Commission an App**

An alternative for transit agencies that do not want to develop an app in-house is to contract out the app development. Unlike most transportation apps, which are developed by private app developers without input or funding from transit agencies, a commissioned app is built to specifications defined by the transit agency and is paid for by the transit agency. One of the transit agencies interviewed had released a request for proposals seeking a vendor to develop an app meeting the transit agency’s requirements.

The selected vendor will build an app that displays information as directed by the transit agency. Not only will the transit agency have control over the way information is presented, but they anticipate being able to provide more custom information such as detours and real-time alerts and service change notifications. The RFP also specifies performance requirements including a minimum level of prediction accuracy and a limit on the “ghost” bus and train incidence rate.

In addition, the RFP requires that the vendor provide data from the app to the transit agency. Specifically, the RFP asks for an analytics platform for the transit agency to use and that data on operations, predictions accuracy, customer preferences, and travel patterns be provided. The RFP includes a sample agreement stating that upon expiration or termination of the software as a service (SaaS) agreement, data will be transferred to the agency and destroyed by the contractor. The agreement also states that the contractor must have data security measures in place and must maintain data privacy.

In the interview, the transit agency identified several potential benefits from the data collected by the app. They see it as a way to get better data on origins and destinations than they currently
collect from surveys (this particular agency has limited access to smart card data). They also hope to learn more about the alternatives to transit that their customers select. They expect to use this information to plan service that better meets the needs of their customers. They even anticipate that they may share some of the data with research partners to gain more insight from it.

A major drawback of this approach is the cost. In addition, the transit agency cannot guarantee that customers will use the app they commission, as other apps will still be available. Also, one argument against a custom app, whether it is developed in-house or commissioned, is that it works against the trend of universal transit planning tools. Apps like Google Maps and Transit App that are available across a large number of transit markets standardize the transit experience for visitors who ride transit in different cities. Standardization of apps may make transit easier to use and actually encourage transit use.

**Endorse an App in Return for Data**

For transit agencies that neither want to develop their own apps nor pay for a custom app, another model exists. In Boston, Transit App, which the MBTA officially endorsed for real-time transit information and travel planning, provides the MBTA with data on app users’ searches and locations (Enwemeka, 2016). The MBTA made this agreement after releasing a non-funded RFP asking app developers to submit bids competing for endorsement by the agency. The transit agency evaluated bids to select the best app to recommend to its users. The RFP also required that the selected app share data with the transit agency. Thus, as in the commissioned app model, this model enables the transit agency to draw insights from app data.

Compared to an in-house app or a commissioned app, this model is much less expensive. However, as in the other cases, there is no guarantee that customers will use the endorsed app. In addition, in this model, the transit agency has less control over the presentation of information in the app. Rather than specifying these components explicitly, they select the best of the option available.

**Getting More Value from Third Party Apps**

There may be other ways for transit agencies to get more from third party apps, in return for the data they share. In Tampa, Florida, a pilot program embedded Open311, a service that allows users to report issues to the local government and transit agencies, in the open source OneBusAway app (Barbeau, 2018B).

In transit agency interviews, transit agencies also discussed two additional information types that could be presented in customer-facing apps: crowding information and fare information. As data and processing methods improve, the possibility to provide reliable crowding information in real-time is increasing. While some transit agencies are concerned that reporting on crowded trains and buses may discourage public transit use, others see this as an important way to increase customer knowledge and improve their experience. Fare information is part of the GTFS standard but is not supplied by all transit agencies. Wang (2014) noted that the existing GTFS standard is insufficient in the way that it describes fares and proposed an extension to GTFS to model complexities in fare structure such as time of day variance, distance-based, free transfers. At least one of the transit agencies interviewed was working to include fare information in its public information feeds and ultimately in transportation apps.
ACCESSING EXTERNAL DATA SOURCES

In addition to getting value from sharing their own data, transit agencies see benefits in accessing external data sets. Mobility data from other sources can help transit agencies understand how people in their service areas get around, which can inform transit planning strategies.

This chapter highlights external data sets that relate to mobility, including those that were identified as data sets of interest by transit agencies in the interviews. The second part of the chapter discusses models for accessing these data sources that are being pursued by transit agencies, including those interviewed.

5.1 External Data Types and Transit Agency Uses

There are a wide variety of data sources that could have relevance to transit agencies, including for example, financial data and social media data, such as twitter. While these data sets may be beneficial to transit agencies (spending patterns can reveal customers’ movements or trip purposes; twitter can be mined for tweets about public transit disruptions), this section describes three classes of data that most directly measure travel patterns: trace data from cellphones and other GPS-enabled devices, data from transportation apps, and data from other mobility providers. It is also worth noting that this section focuses on private data sources. There are also external, publicly-available data sources that transit agencies access on a regular basis, such as census data, weather data, and GIS data from cities, states, and regional agencies.

Cellphone and GPS Speed and Trace Data

Cellphone connection data is collected by cellular service companies, while apps that request users’ locations, including fitness, navigation, social media, and dating apps, also collect data on peoples’ movements. Companies also collect GPS data from in-vehicle navigation and fleet management systems. These data sources are aggregated by analytics companies who sell speed and origin-destination insights, inferred from the data. The transit agencies interviewed were interested in this type of data for a variety of purposes. Motivations for accessing this data include:

- Enabling transit agencies to evaluate overall demand patterns in their service area and determine how to better meet peoples’ needs.
- Enabling more detailed road speed data to inform bus operations and route alignment decisions, and to improve bus arrival predictions.
- Enabling analysis of how people behave when incidents or other disruptions prevent them from using the transit system.

One transit agency interviewed was in the process of acquiring cellphone data to better understand demand patterns in their system, and another had already purchased cellphone data.
Transportation App Data

One particular type of cellphone data is data from transportation planning apps. These apps include navigation apps, such as Google Maps and Waze, and apps such as Transit App and NextBus that provide information on transit vehicle arrivals, and collect information including:

- records for each session, including beginning and ending coordinates and time stamps
- placemarks – stored home and work locations
- carshare, bikeshare, and TNC bookings (if available through the app)
- trip planning routes and stops searched, favorite routes

Data from these apps enables an additional layer of insight over other location data from smartphone apps. Because an analyst can identify when and where a user looks at transit information for a particular location or route and then how they behave after (whether they take transit, book an alternate mode, or do not travel at all), this app data allows transit agencies to better understand their customers’ decision-making processes.

Data from Other Mobility Providers

Transit agencies are very interested in the travel alternatives that transit passengers have since that is a major determinant of transit demand. The transit agencies interviewed were interested in TNC, scooter, carshare, and bikeshare data. One of the transit agencies interviewed had already used bikeshare data to understand public transit’s competitiveness with other modes. However, the small user base for the bikeshare system made it difficult to draw conclusions.

In general, private mobility providers share some data publicly. For example, several bikeshare systems have released data on trip history [https://www.capitalbikeshare.com/system-data](https://www.capitalbikeshare.com/system-data) and [https://www.citibikenyc.com/system-data](https://www.citibikenyc.com/system-data) and Uber provides Uber Movement data, which shows zone to zone travel times based on Uber driver data. However, most TNC companies are hesitant to share demand data publicly.

5.2 Models to Access External Data

As discussed in Section 4.3, some transit agencies are leveraging their GTFS and GTFS-RT data, as well as their capacity to endorse an app for their customers to use, to access the data that transportation apps collect. In addition to this, there are several methods that transit agencies are using to access private data sources. The following are examples that were identified either through the literature review or in interviews.
Buying Data

The simplest model is buying data outright. Transport for London has documented its pilot project, purchasing data from O2, a cell service provider, on users of the O2 network (Transport for London, 2019). They are purchasing depersonalized, aggregated data that they will use to better understand demand patterns, particularly for non-transit users, including cyclists, pedestrians, and drivers. They note that this data provides a better, more convenient alternative to roadside interview surveys. Another transit agency interviewed was in the process of purchasing cellphone data to be used for strategic planning purposes.

“Buying anonymous and aggregated mobile phone event data is a practical alternative to [roadside interview surveys]. It will provide data on trip patterns in a cost efficient, safe way without inconveniencing customers or compromising their privacy.” (Transport for London, 2019)

Accessing Data Through Mobility Service Partnerships

Many transit agencies have partnerships with TNCs. A 2018 study cited 29 examples in the United States (Schweitzerman, 2018). A Federal Transit Administration (FTA) program, the Mobility-on-Demand (MOD) Sandbox Program provides funding and oversight for many of these partnerships and includes a data sharing requirement. Data sharing happens in a variety of ways. This section describes three examples from three different transit agencies that were interviewed. In one case, the transit agency has taken an evolving approach to its partnerships, gaining access to more data over time. In another case, the transit agency negotiated a detailed agreement with their partner (Via, a private mobility provider and competitor of Uber and Lyft) for specific data items and levels of aggregation. In the third case, the transit agency and its private partner (Uber) provide data to a third party who conducts analysis combining the two data sets.

Evolving Access to Private Data

In this example, the transit agency interviewed has a partnership with Uber to provide first and last mile trips, feeding the transit system and to provide door-to-door rides for members of its transportation disadvantaged program. This transit agency has been at the forefront of transit agency partnership with TNCs. Initially, they received only aggregated data from Uber on the subsidized first and last mile rides, but with each contract amendment, they have negotiated to receive more data. Still, they do not receive trip level data.

The first and last mile program was the transit agency’s first partnership with Uber, and one of the first such programs in the country. They were focused on establishing the partnership initially and have devoted effort to accessing data in subsequent steps.

For the transportation disadvantaged program, the transit agency has access to data through the Uber for Business platform and receives full trip-level origin-destination data as of October 2018.
They use the data to track users of the program as well as program response time. They would like to use this data to better plan their transit services. They see the evolution in data transfer for both programs as a sign of their experience. They understand the importance of trip level data and have learned that they can negotiate for it successfully.

When private companies provide data to public agencies, the data can then be requested through public records request, a factor that concerns some private companies. To address this, companies such as Uber generally require that requests are sent to them for review. This transit agency noted that while they notify Uber of public records requests, so far, they had provided the data without dispute.

While overall their experience has been positive, the transit agency noted some data challenges. For example, Uber and taxi data are provided in different formats and require significant effort to join.

**Custom Agreement**

In contrast with the incremental approach described above, another transit agency interviewed has taken steps to carefully negotiate a detailed data agreement upfront. For their FTA MOD-Sandbox-funded project, the transit agency had originally planned to partner with Lyft, who had verbally agreed to share data. In the scoping process, however, the transit agency determined that they would not be able to reach the kind of data agreement they wanted with Lyft.

Instead, they went through an informal procurement process seeking a mobility provider that was willing to share data, and ultimately selected Via as a partner. Once they selected Via, the transit agency continued with a careful and time-consuming process to iron out the details of their data sharing agreement. The agreement includes:

- A list of variables that Via will share and the level of granularity for each one
- Designation of which researchers would have access to the data and where it would be housed.
- Designation of what data would be made public. The transit agency has committed to sharing aggregated data with municipal operators and city partners. In addition, there will be public reports on the pilot. Disaggregate data will not be made public.
- Data ownership: The transit agency will not own the disaggregated data but rather will have access to it for five years. They will be able to access aggregate data indefinitely.

In the negotiation process, Via expressed concerns over competitors accessing the data and over the privacy of their users.

**Using a Third Party**

A partnership between Uber and transit agencies in Cincinnati includes a “first-of-its-kind study” in partnership with transit agencies SORTA and TANK, in which a transit consultant will combine Uber and transit agency data to draw insights that can inform strategic transit planning in the region (Salzberg, 2018; Schweiterman et al., 2018).
In this case, Uber hired and is paying for a consulting firm to analyze how people move in the city, including their use of rideshare and transit. The consultant has been in communication with the transit agencies about their needs and will publish a report on their findings based on the two data sources.

**Challenges with Using External Data**

The first challenge is engaging with external partners and finding a partner who will cooperate, particularly if transit agencies seek a cost neutral solution. Transit agencies are curious about the value of their data to private companies. For example, can public transit data be combined with cellphone data to help identify which mode people are using?

Transit agencies are also worried about data quality from external partners. If they use data from an external provider, transit agencies are interested in knowing about the data source, including factors such as

- sample size
- data cleaning process
- data latency
- processing methodology/ algorithms used (if any)

Ultimately, some of the transit agencies interviewed are looking to have more control over relationships with private companies and to expect more in return from data they share.
KEY FINDINGS AND NEXT STEPS

The transit agency interviews and literature and information review conducted to date reveal several key findings.

- **Transit agencies collect a wide variety of data on transit passengers, much of which can generate value through sharing.** Transit agencies collect data on passengers, including fare and bank card transaction data, Wi-Fi and Bluetooth data, video data, and passenger count data, all of which shows how the transit agency’s network is being used. This data is often of interest to journalists, real estate developers, and researchers. While data on individual passengers can present privacy risks, publishing this data promotes transparency and sharing it with researchers can generate insights and innovation that are beneficial to the transit agency.

- **Transit agencies also have data on the transit system, including route, schedule, and vehicle location data, which is commonly shared and contributes to customer information.** In addition to data on passengers, transit agencies store incident data, route and schedule data, vehicle location data, system and maintenance data, staffing and operations data, and financial data. Sharing these data types can also promote transparency and generate insightful research. Additionally, route, schedule, and vehicle location data is frequently shared by transit agencies (ten of eleven transit agencies interviewed share this data publicly). Private developers routinely use this data in customer-facing apps that help transit passengers plan their routes and find out when transit vehicles will arrive at stops and stations. Some of the transit agencies interviewed receive data from these travel planning apps.

- **Collecting, cleaning, processing, documenting and cataloguing data requires significant effort.** Several of the transit agencies interviewed noted the significant technical needs and effort required to prepare data for sharing. In some cases, data must be processed to match the level of aggregation requested, to make the data accessible to non-technical audiences, or to protect individuals’ privacy. Those transit agencies that had developed procedures for processing and cataloguing data found that this saved time responding both to public and internal data requests.

- **Transit agencies share some data openly and share other data sets directly with partner institutions or individuals through private data sharing agreements.** The most common type of open data, among the transit agencies interviewed, was route, schedule, and vehicle location data, but transit agencies also share ridership, on-time performance, survey data, and financial data publicly on their websites. In addition, all of the transit agencies interviewed respond to public records requests for data. Several of the transit agencies have established data sharing relationships with research institutions and reported on beneficial insights gained through these relationships.

- **Transit agencies access external data sets, either by purchasing data or leveraging a mobility services partnership.** The transit agencies interviewed acknowledged the challenges of negotiating data sharing agreements with private mobility providers, even when they have reached a service agreement. Despite these challenges, transit agencies see
value in these data sources, and at least two of the transit agencies interviewed have negotiated access to data from TNCs.

- **The key factors transit agencies consider when deciding what data to share and whom to share it with whom include risks, costs, and benefits. They also must account for any laws or regulations that apply.** Risks of data sharing include privacy, security, misuse, and strategic risks. The transit agencies interviewed mentioned all of these risks, with privacy risks described most frequently. Transit agencies are also aware of the cost of sharing data, noting the time required to respond to data requests and to prepare data to be shared publicly. Transit agencies weigh these risks and costs against the benefits they have observed from sharing data, including innovation, cost savings, transparency, data reciprocity, and sometimes monetary gain.

- **Transit agencies identified both internal and external needs to improve their processes for sharing data.** The majority of transit agencies interviewed do not have a centralized data repository or a staff or group dedicated to data management. The transit agencies noted that establishing dedicated staff could help make data management a priority. In addition, they expressed frustration that data is stored across divisions making it difficult to find, use, and share. Transit agencies are looking for external guidance to answer technical questions about the legal context for data sharing and protocols for minimizing privacy and other risks. They also see the need to establish formal frameworks for data sharing, which they believe requires both internal effort and external guidance. The majority of the transit agencies interviewed were supportive of the idea of standards for public transit data types, noting that standards could promote the development of shared tools and other resources. Transit agencies are looking to external organizations for standards creation and adoption.

Given these findings, the next step for this project is to seek additional perspectives that can shape guidance and help fill the needs that transit agencies have specified. The next information gathering phase will answer questions including:

- How do private companies perceive public transit data? What kind of use cases might they have for the data?
- What are the legal issues that transit agencies should consider when making decisions about data sharing now and in the future? (See more detailed questions regarding laws and regulations in Chapter 2)
- What kind of privacy guidance should transit agencies follow?
- What lessons can be learned from the private sector and other public sector agencies? What kind of models do they use to share and access data? What factors do they consider when making decisions about data sharing? What kind of protocols do they use to minimize privacy and other risks?

The additional information gathering will expand this report to cover the topics introduced more comprehensively, including the addition of lessons learned from other sectors. In addition, the information in this report and from the next information gathering phase (Task 4) will be synthesized into guidance for transit agencies (Task 5). This guidance will be informed both by the transit agency experience (as understood through the transit agency interviews and literature review) and by lessons gleaned from other sectors.
BIBLIOGRAPHY


Antrim, Aaron, and Sean J. Barbeau. (2013). The many uses of GTFS data–opening the door to transit and multimodal applications. Location-Aware Information Systems Laboratory at the University of South Florida 4.


Energy Information Administration. “Smart meter deployments continue to rise.” 1 November 2012.


APPENDIX

I. Transit Agency Interview Guide

1. Has your transit agency shared, traded, or sold any kind of transit agency data?
Sharing data includes open data initiatives and private data sharing agreements (for example, with a research institute). Trading refers to an in-kind arrangement (for example, providing data in exchange for services or advertising). Selling involves exchanging data for monetary compensation.

If yes:

a. What data was shared, traded, or sold?
b. How was it shared, traded, or sold?
c. Who was it shared or traded with, or sold to?
d. Can you share the data agreement with us (or any related documentation)?
e. How did your transit agency identify and engage the partner who received the data?
f. Did your transit agency have to clean, process, or standardize the data?
g. What benefit(s) did your transit agency receive from sharing, trading, or selling data?
h. Can you quantify the benefit in monetary terms?
i. Were there negative consequences of sharing, trading, or selling data?
j. What (if any) barriers did your transit agency encounter in the process of sharing, trading, or selling data? For example, consider privacy concerns, technological barriers, or public perception challenges.
k. Is there any information or tool that you think would have made the process of sharing, trading, or selling data easier?
l. What advice would you give a transit agency considering a similar agreement?

2. (In addition to the experience just described), has your transit agency considered sharing, trading, or selling any kind of transit agency data?

If yes:

a. What barriers did your transit agency encounter that have prevented you from sharing, trading, or selling data at this time? For example, consider lack of partners, privacy concerns, technological barriers, security concerns, or public perception challenges.
3. For transit data types that your transit agency has not considered sharing, or trading, or selling, what barriers to sharing, trading, and selling exist?

For example, consider lack of partners, privacy concerns, technological barriers, or public perception challenges.

*Example Data Sources*

- Fare Collection Data
- Automated Vehicle Location Data
- Route and Schedule Data
- Vehicle Maintenance Data
- Station Data
- Survey data
- Automatic passenger counter/load data
- Safety data
- Wi-Fi Data
- App Data (for example from a trip planner or fare purchasing app)
- Other

4. For transit data types that your transit agency has not considered sharing, or trading, or selling, is there any information, tool, or structure that would allow your transit agency to overcome these barriers?

This could include information on partners or tools to connect with them, tools or structures for data standardization, organizations for data warehousing, etc.

5. What changes do you anticipate in the future that could alter transit data sharing, trading, and selling?

6. Are there external data sources that your transit agency would be interested in gaining access to? Explain the value these data sources would provide.

External data sources could include private mobility provider data, Wi-Fi data, GPS probe data, and app data.

7. What value can transit data provide to external entities?

External entities could include public agencies, private companies (e.g., app developers, mobility providers, advertisers), and researchers.

8. Are there questions you have for these external data providers regarding access to their data?

9. What questions would you want to be answered in a guide on transit data sharing, trading, and selling?

10. Is there anyone else that you know of who we should talk to for this project?

This could be someone else at your transit agency, at another transit agency, someone who consults with transit agencies, or some other subject matter expert.
II. Private Sector Interview Guide

I. DATA SHARING, BUYING, SELLING

Q1a Does your business have any existing agreements with public sector entities to buy, share or sell data?

   If yes, interviewer will continue with b-f

Q1b What data or data product does your business buy/sell/share to (from) public sector entities?

Q1c What is the data format?

Q1d Are there obstacles to buying/sharing/selling because of agency data format?

Q1e Who was it shared with or sold to?

   Interviewer will prompt the following categories:
   cities or city agencies
   transit agencies
   metropolitan level agencies
   state level agencies
   federal level agencies
   NGOs
   Other

Q1f Could you share the data agreement with us (or any related documentation)?

   If respondent is unable to share details of the agreement, interview will ask “if you could share the name of the public sector agency with whom the agreement is signed”.

Q2a Would your business be interested in buying data from transit agencies?

   On a scale of one to ten, one being no interest at all, how would you rate your interest in buying data from transit agencies?

Q2b Would your business be interested in sharing data with transit agencies?

   On a scale of 1 to 10, one being no interest at all, how would you rate your interest in sharing data with transit agencies?
Q3 What types of data currently collected by transit agencies may have value to your business?

*Interviewer will prompt with the following categories:*

- Automated Vehicle Location Data
- Route and Schedule Data
- Vehicle Maintenance Data
- Station Data
- Survey Data
- Anonymous passenger counter/load Data
- Safety Data
- Wi-Fi Data
- App Data (for example from a trip planner or fare purchasing app)
- Other

Q4 What types of data not currently collected by transit agencies, as discussed in the previous question, may be useful to your business?

Q5a Would data from transit agencies be more valuable if it were compiled from multiple agencies (as opposed to one agency) within one region (city, region, or state)?

Q5b Would data from transit agencies be more valuable if it were compiled from multiple regions (as opposed to one city or region)?

Q6a Would transit data be more valuable to your business if it were compiled with a national or international data standard, such as GTFS?

On a scale of one to ten, ten being the most valuable, how would you rate the value of standardized transit data?

Q7a Would transit data be more valuable to your business if it were processed, but not necessarily in a national or international standard, as opposed to raw data?

On a scale of one to ten, ten being the most valuable, how would you rate the value of processed transit data, as opposed to raw data?
Q8 How does your business currently engage with other private sector data businesses?  
*Interviewer will prompt different data transaction models: buying, selling, sharing*

Q9 How many business clients/partners do you currently engage in buying, sharing or selling data?

Q10 What are some of the obstacles to your business buying, sharing or selling data to (from) public sector entities?  
*Interviewer will prompt with the following categories:*
- regulatory hurdles,
- privacy or security restraints,
- data granularity,
- data not processed,
- data not aggregated or anonymized,
- lack of public sector capacity or knowledge,
- other

Q11 What is the most common form of data exchange that your business engages in?  
*Interviewer will prompt with the following categories:*
- buying/selling of data,
- selling or trading propriety data streams,
- trading data in exchange for DaaS platform use,
- receiving data in exchange for DaaS platform use,
- other.

Q12 What changes do you anticipate in the future that could alter the practice of data sharing and selling for your business?
III. Public Sector Interview Guide

I. DATA SHARING OR SELLING

1. Do you have an agreement to share or purchase mobility data from private partners?

2. Do you share or sell data to a private partner?
   
   If yes, interviewer will continue with questions 3-7

3. Who is the private partner?

4. What data is shared or sold?

5. In what format is the data shared, or sold?
   
   □ Raw data
   □ Data visualization platforms
   □ Data analytical results in aggregated forms
   □ Software with data analytical results
   □ Other, please explain ____________________

6. If you have an agreement to sell or purchase data, how did you decide on the price?

7. Can you share the data agreement with us?

II. DATA POLICIES

8. Is the data you collect available to the public?

9. How does the public currently access and use data from your agency? How do you anticipate that changing in the future?

10. What (if any) barriers exist to sharing or selling data to (from) the private sector?
11. Is private sector data that you receive or purchase available to the public? In what form?

III. MOBILITY ON DEMAND SANDBOX/SMART CITY CHALLENGE PROJECT OVERVIEW

12. What city department or agency is conducting the project?

13. What are the goals of the project?

14. What private entities have you partnered with to meet the project goals?

IV. DATA COLLECTION

15. What kinds of transit data does your department or agency collect?

16. Do the private entities you’ve partnered with collect mobility data?

17. What data do private partners most often request?
IV. Utility Industry Interview Guide

1. Has your utility/organization shared, traded, or sold any kind of utility data?

Sharing data includes open data initiatives and private data sharing agreements (for example, with a research institute). Trading refers to an in-kind arrangement (for example, providing data in exchange for services or advertising). Selling exchanges data for monetary compensation.

If yes, interviewer will ask these follow up questions:

   a. What data was shared, traded or sold? What criteria did your utility/organization use to define time intervals, time period, and frequency of providing this data?

   b. How was it shared, traded or sold? What are the benefits to your method/platform?

   c. Who was it shared or traded with, or sold to? If smart meter data is shared – how is individual customer data used compared to aggregated/anonymized data used (are uses different, is the value of these types of data different)?

   d. How did your utility/organization identify and engage the data user/buyer?

   e. How does your organization process and standardize data? What do you recommend to transit agencies, to data sharing standards?

   f. What benefit(s) did your utility/organization receive from sharing, trading, or selling data?

   g. Can you quantify the benefit in monetary terms? If priced, how did your utility/organization determine pricing for the data?

   h. Were there negative consequences of sharing, trading, or selling data?

   i. What (if any) barriers did your utility encounter in the process of sharing, trading, or selling data? For example, consider privacy concerns, technological barriers, public perception challenges, expertise and capital that needed to be developed.

   j. Is there any information or tool that you think would have made the process of sharing, trading, or selling data easier?

   k. What advice would you give a transit agency considering a similar agreement, considering they have many of the same challenges that an electric utility has in managing data?

2. Is there anyone else that you know of who we should talk to for this project?

This could be someone else at your utility/organization, at another utility/organization, a partner, or a subject expert.