Getting Started with Open Data

A Guide for Transportation Agencies

Sarah M. Kaufman
Rudin Center for Transportation Policy and Management
Robert F. Wagner Graduate School of Public Service
New York University

5/1/2012

Getting Started with Open Data

A Guide for Transportation Agencies

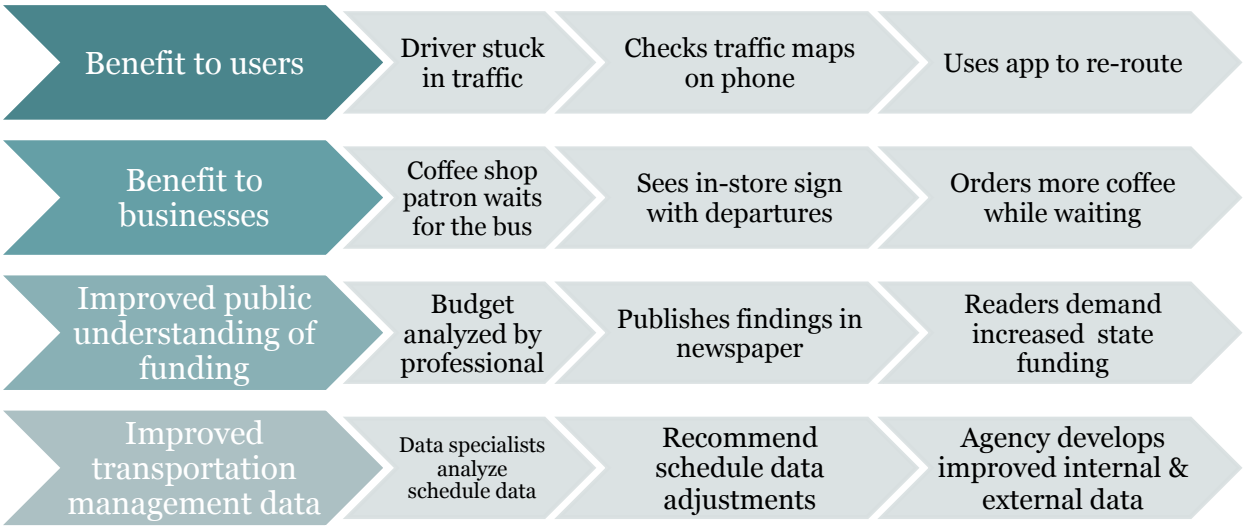
This guide is designed for transportation agencies: to open and maintain data, overcome potential obstacles, and create a relationship with users. Please note that this is a living document; your comments are welcome on the Google Doc version.

What is "Open?"

To be truly open, an organization must make available to the public internal data in a format usable by both interested individuals and application programmers. The opening of data should generate improved communications between transportation organizations and their customers, resulting in improved travel and services. The benefits of opening up data include more efficient travel (with an enhanced ability to find optimal routes while on the go), a greater understanding of finance/administration (helping to possibly promote improved funding), and crowdsourced analysis capabilities (potentially helping detect schedule improvements or errors in stop locations/names, for instance) See the graphic below for examples of data uses. The data, which would typically include sets like schedules, routes, budgetary information, ridership numbers, traffic numbers and road conditions, should be released in both historic and real-time for both analysis and prediction. The data should be released closest to its most original format, barring security-sensitive inputs, and regardless of potentially negative concerns (which are typically neutralized by positives in other data).

Being open involves sharing information for optimized travel, management, and future improvements.

Typical Transportation Data Benefits



Costs of Opening Data

Because open data uses existing internal data, the costs of generating it are modest. However, there are costs, such as:

- Converting data to mainstream formats
- Web service for hosting data
- Personnel time to update and maintain data as needed
- Personnel time to liaise with data users

These costs vary, depending upon the size and scale of the transportation services covered. However, subsequent sections in this guide will discuss these responsibilities more fully. It should be noted that these costs are lower than any internal attempt to create applications internally; as written by the Boston Globe about the Massachusetts Department of Transportation's data releases, "This approach is a smart 21st-century alternative to hiring some consultant who develops inelegant software at exorbitant costs."[1] It should be noted that the costs involved should not prevent openness: for public transportation agencies, data was developed using public funds, and should be accessible for public use and analysis. The returns on these investments are manifold, reaching large numbers of people with modest effort.

The benefits of improved travel outweigh the maintenance costs; organizations must think holistically about their missions, like providing mobility resources and the tools to use them, rather than by departmental line budgets, in order to embrace data openness fully.

Open Data in Use

Transportation data is used worldwide for policy analysis and smartphone applications. In the United States, 228 transit agencies are sharing their schedule data[2] (nearly 500 worldwide[3]). Of the half-million applications in the Apple App Store, many thousands are transit-related. Real-time transit data is available in dozens of locations via the NextBus system, and others via Google's new standard in four US cities (Boston, Portland, OR, San Diego, and San Francisco).[4,5] In a recent Transit Cooperative Research Program study, of 28 transit agencies surveyed, 13 relied on third-party applications to disseminate real-time information, but only three agencies relied on internally-developed applications.[6] Those 13 agencies have embraced the potential of open data to inform customers through any medium; they cite the benefits of releasing data as:

[1] "MBTA: App Judgment," *The Boston Globe*, Editorial, September 5, 2009. http://www.boston.com/bostonglobe/editorial_opinion/editorials/articles/2009/09/05/mbta_app_judgment/

[2] www.citygoround.org

[3] Roush, Wade. "Google Transit: How (and Why) the Search Giant is Remapping Public Transportation," XConomy, 2/21/12. http://www.xconomy.com/san-francisco/2012/02/21/google-transit-a-search-giant-remaps-public-transportation/

[4] NextBus Agency Selector. http://www.nextbus.com/predictor/agencySelector.jsp

[5] Roush.

[6] TCRP Synthesis 91. "Use and Deployment of Mobile Device Technology for Real-Time Transit Information," p 27. http://onlinepubs.trb.org/onlinepubs/tcrp/tcrp_syn_91.pdf

- Free development of mobile applications,
- Increased ridership,
- Improved customer service,
- Time saved by agencies in developing customized applications,
- More accurate applications, and
- Positive image for agencies.[7]

In addition, using the data, developers can create applications for users in foreign languages, for those with vision disabilities, and highly localized information for particular neighborhoods. The following table shows the number of applications and ridership in several U.S. cities.

Transit Apps in High-Ridership U.S. Cities[8]

Transit agency	New York MTA	Chicago CTA	Boston MBTA	Washington WMATA	Portland TriMet
Avg. weekday ridership[9]	8,487,642	1,640,000	1,300,000	1,093,112	318,500
Number of apps[10]	68	25	43	11	45
Ratio (app/riders)	1/124,818	1/65,600	1/30,232	1/99,374	1/7,078
First year of data release	2010	2009	2009	2010	2006

[7] "Simplifying the Open Transit Data Debate: A Comprehensive Guide to Providing Real-time Information to Your Passengers," *Fleet Beat*. February 8, 2010. http://www.mentoreng.com/blog/index.php/2010/02/simplifying-the-open-transit-data-debate-a-comprehensive-guide-to-providing-real-time-information-to-your-passengers/

[8] Based on the table in "Transit transparency: Open data in action," with additional facts from agency websites. http://policybythenumbers.blogspot.com/2012/01/transit-transparency-open-data-in.html

[9] 2011; all modes

[10] Featured on agency's own websites as of April 2012; excludes duplicate apps for separate platforms

Transportation Data Standards: A Primer

The chart below describes common data standards and file formats used by transportation agencies for a variety of purposes (data standards are composed of multiple files, arranged in a particular way, in the formats listed below):

	Champion	Where it's used	Applicable data sets	Examples	More information[11]
Data Standards					
GTFS	Google	Worldwide	Schedule data	Train line schedule	https://developers.google.com/transit/gtfs/
GTFS-realtime	Google	Select US & European cities	Real-time data	"Train arriving in 3 min"	https://developers.google.com/transit/gtfs-realtime/
SIRI	European Committee for Standardization	European cities	Real-time data	"Train arriving in 3 min"	http://bustime.mta.info/wiki/Developers/SIRIIntro
TransXchange	UK Gov	UK Buses	Bus schedules & data	Bus route schedule	http://www.dft.gov.uk/transxchange/
DATEX 2	European Commission	European Cities	Traffic data & Management	Delays on Route 4	http://www.datex2.eu/content/datex-background
File Formats					
CSV	Many	Worldwide	Data tables	Historic on-time data	http://www.ehow.com/how_5091077_use-csv-files.html
TXT	Many	Worldwide	Text	Textual information	http://en.wikipedia.org/wiki/Text_file
GIS	Many	Worldwide	Geographic mapping	Subway station entrances	http://en.wikipedia.org/wiki/GIS_file_formats
KML	Google	Worldwide	Google Maps & Earth	GIS road outlines	https://developers.google.com/kml/documentation/
XML	Many	Worldwide	Large data sets	Traffic numbers	http://www.w3schools.com/xml/xml_whatis.asp

Note that files in the formats above are used both within the data standards (for example, a GTFS data set is actually a series of specific TXT files), and on their own (to convey text or information separately, like budgetary information in a CSV file). This is a preliminary list and does not include all data standards or file formats.

[11] Recommended resources; not necessarily official

It's essential to choose a file format that is both easily convertible within your transportation agency and useful to the majority of readers and data developers (that includes converting Microsoft Excel spreadsheets to CSV, and Word documents to TXT). Striking that balance may require using your second-choice data format to accommodate the most users in your area. (Not sure what they'll use? Post the question online, or hold a workshop to collaborate on data releases.[12]) Be sure to avoid using PDF files, as these make extracting data most difficult.

A Note on Google's Role

Google plays a critical role in transit and mapping data. The company created the tools and environment for widespread information for transit users with GTFS (General – formerly Google – Transit Feed Specification) in 2006.[13] The standard made it simple for transit agencies to open their data to riders, helping them ride the system more efficiently; now 228 U.S. transit agencies use the format. However, some consider GTFS to be a necessary evil, due to its lack of ability to specify agency needs (for example, limits to publishing "exception trips," like holiday schedules). Furthermore, with the potential for Google to impose charges for use of its maps in applications, many transit agencies are wary of investing in the standard. However, GTFS is so widely-used that it is the default in most communities, and has become the *de facto* data standard. If a new standard emerges, it will need to make porting the data over relatively simple.

Other data champions are not corporations, but standards bodies, such as ANSI (American National Standards Institute) and CEN (European Committee for Standardization), which champions SIRI. These independent bodies are useful reference points for understanding industry standards.

[12] For a useful list of all file formats, visit: http://en.wikipedia.org/wiki/List_of_file_formats

[13] Roush.

How to Release Transportation Data

The following is a step-by-step process of releasing data.

Step 1: Find your data

What data is readily available? Most agencies have schedule and GIS locations already in internal databases. Aim to release these data sets, at a minimum:

1. Schedules (GTFS)
2. Routes (GTFS)
3. Infrastructure locations, including stations, roadways and landmarks (GIS)

For enhanced transparency, these data sets should be released soon, if not immediately:

1. Budgetary data (CSV)
2. Performance data (CSV)
3. Ridership Data (CSV)

Step 2: Convert data

Data will need to be converted from HASTUS or other internal formats into the more widely usable formats. For GTFS, see the instructions from Google. If your agency lacks the resources or know-how to convert data as needed, you may wish to use a third-party application, such as Timetable Publisher.

Step 3: Test your output

Test the GTFS feed using Google's GTFS Validator, which will find incorrect conversions and schedule anomalies. In addition, seek out several friendly application developers and request that they test out the data (many will be thankful for an early look).

Step 4: Write up a License Agreement

The license agreement for developers and analysts using your data should cover at least some of these elements:

1. Use and placement of copyrighted logos and images
2. Right to use the agency's data
3. Agency's right to alter data without notice or liability
4. Non-guarantee of data availability or timeliness

5. Liability limitations for missing or incorrect data
6. Indemnity from technical malfunctions due to users' use of data
7. Indemnity from legal actions against data users

For an exemplary license agreement, see that provided by Massachusetts Department of Transportation (MassDOT).

MassDOT also provides a Relationship Principles guide, essentially a statement of respect between the Agency and its data users.

Step 5: Publish and Publicize

It's time to release the data. On your website, post the following items:

- All GTFS, GIS and other files, compressed if larger than 10 MB (with date last uploaded and next expected update), for users to download
- License Agreement
- Glossary of agency terms (similar to the one provided by TriMet)
- Link to developer community (see Step 7)
- Links to adjoining transportation providers' data sites

Congratulations! You have opened your data.

Once the site is up and running, partner with Google's Transit Partners group to put your data on Google Maps.

Finally, publicize the data and developer group on the agency website, using social media, and reaching out to local news outlets.

Step 6 (continuous): Update and modify as needed

The data should be updated as needed, both due to information changes and user requests for different data fields or terms. This process is ongoing, and should be considered regularly (monthly, at a minimum).

When publishing data, you may wish to consider third-party API hosts, such as Socrata (used by NYC and other government agencies). These services host your data sets in the cloud, releasing your agency from needing to maintain adequate server space for numerous data pulls, automate data updates, and provide usage analytics and technical assistance. They are particularly useful for providing real-time data when your agency's infrastructure is unable to handle the requests.

Visit Socrata or see a list of several recommended providers here.

Step 7: Create and maintain a dialogue

It is essential to keep up a conversation with data users for the following purposes:

- Announcements of updates, modifications, etc.
- Technical support
- Feedback on data anomalies
- Requests for future data sets
- Explanation of jargon and definitions

These conversations are typically best conducted through online forums, like Google or Yahoo Groups. However, your agency should also reach out to data users in person to foster more advanced discussions and promote collaboration between developers to produce even better products. Resources like Meetup.com can help you plan the gathering and reach out to appropriate attendees. Data user comments and findings can be extremely valuable to your agency, in terms of technical updates and operations planning, so be sure to invite your agency's data creators.

Events like hackathons and barcamps, in which attendees develop software at the event in a contest setting, are popular and highly recommended for your data releases.

Finally, consider incentivizing certain data uses. For example, if your agency would like to see an app in a secondary language, a contest with that objective may pose many solutions, and will likely highlight a need that developers are happy to learn about.

Step 8: What more can you do?

What other data can you release? It's worth considering, at least monthly, your long-term goals for data releases. Perhaps it's an extremely large data set that would take too long to convert. In that case, consider enlisting developer who might volunteer to ready the data for public consumption. Developers are often interested in receiving nominal credit. If the data is seen as posing a security concern, bring it to the attention of your agency's security forces for their input on how it can be altered for public release. It's always productive to think about future data release goals, which will often stem from developer requests.

Common Concerns

Ideas to consider as you and your agency prepare for data releases:

What are the costs of releasing data?

Costs vary by agency and volume of data output, but should be considered in terms of these considerations:

- Initial costs: Building data section of site (with adequate server space); publicity
- Intermittent costs: Labor – converting new data into appropriate format; attending events
- Continuous costs: Labor – making necessary data fixes and updates; relationship-building with data users

In a Transit Cooperative Research Program Report, agencies surveyed reported that staff time was required primarily from these departments: Information Technology, Marketing/Communications, Customer Service and Operations.[14] It should be noted that only four agencies ventured to estimate monthly labor hours, and those numbers ranged from 4 to 200, showing the vague distinction between data opening and other tasks, and the flexibility of time spent.

Why can't I charge money for the data?

You may be a public agency, funded by the government, so the data is already publicly-owned. Like weather information, it belongs to the public and we're all better off having it easily available, since it will likely reduce website congestion, minimize conversion efforts for new technology releases, and enable the integration of data into places people already go.

Furthermore, it is illegal in the US to copyright facts, of which most transportation data consists.[15] You'll find savings in other areas, like less customer information staffing, and reduced call center needs.

What if a vendor owns our trip-planning software?

Check your contract for any notes on data releases; there's a chance your vendor will convert the data for you. If you'll enter into a new contract soon, negotiate to either have your vendor generate the GTFS for you, or ensure no conflict as you release data.

[14] TCRP Synthesis 91, p. 31

[15] "Copyright in General," U.S. Copyright Office. http://www.copyright.gov/help/faq/faq-general.html

What if app developers post incorrect information to their users?

First, open data is only one channel for releasing transportation information, which should also be displayed on your website (always), sent to television and radio stations, and posted on multiple social media channels. Users will stop using apps with bad information, and the market will account for these less informative apps with lower ratings and bad reviews in their stores.

If some data users are truly irresponsible, you can prevent them from accessing real-time data by providing it only to registered users.

If people get their information elsewhere, they won't come to my site anymore.

Most people get their traffic and transit information from local television and radio, according to a Pew study.[16] While nearly 20% of survey respondents got information from local TV and 9% from the internet (other than local government sites), only 1% went to their local government. It's best to push information through multiple channels, including open data, to reach the most customers.

I can't respond to requests 24/7

Manage expectations: if you're available for assistance Monday through Friday, 9-5, post that with your contact information. It's always helpful to check in off-hours in case of emergencies, but set a precedent of business hours-only, and users will respect that. Reply to emails and return phone calls during your announced available hours.

I'm having technical issues

Work with your IT staff to check filenames on the production server along with webpage links. And of course, remember to back up your data in multiple locations, and/or the cloud. Be sure to change passwords frequently and limit access to the web server. Also, check in with developers and peer transportation agencies for input on the issues.

I don't know how to respond to this request for data

Some data sets are not released for political, technical or other concerns. If you would like to release the data, host a meeting to discuss possible solutions or compromises, like partial releases. For technical issues, seek out trusted individuals from the developer community who

[16] "How People Learn About Their Local Community," Pew Research Center, January 2011. http://pewresearch.org/docs/?DocID=140

are interested in figuring out solutions; they will be interested in the credit and the insider perspective, and are often remarkably smart.

Additional Resources

Case studies

- TriMet in Portland, OR
- Transport for London, UK
- MTA in New York, NY
- MBTA in Boston, MA

Takeaways

1. Remember that the well-informed traveler is calmer and moves more efficiently
2. Be open: Pursue clarity, authority and accessibility to propagate your information.
3. Be the authoritative voice on multiple channels.
4. Information is a two-way street: absorb suggestions and corrections, which will often prove extremely valuable.
5. Build and maintain relationships with the public, and many concerns will solve themselves through collaboration.

Useful Links

Selected transportation agency developer sites:

- TriMet Transit – Portland, OR
- MBTA Transit – Boston, MA
- MTA Transit – New York, NY
- CTA Transit – Chicago, IL
- WMATA Transit – Washington, DC
- Transport for London Real-Time Transit Data – London, UK
- 511 Traffic & Transit – San Francisco, CA
- NYC DOT Real-Time Traffic Speed Data – New York, NY
- LA DOT Traffic Data – Los Angeles, CA
- GTFS Exchange Transportation Agency List

GTFS Tools:

- GTFS Specification Intro (Google)
- How to Provide Open Data (from GTFS Exchange)
- MBTA GTFS Primer (PDF)
- GTFS Feed Validator (Google)
- Open Data Manual (loose basis for the format of this document)
- Video: A case for open data in transit

Discussion Groups:

- Google Transit Developers Discussion Group
- GTFS Specification Discussion Group
- Proposed GTFS Specification Changes Group
- MTA Developer Resources Group

Data Consultants:

- Socrata
- Mashery
- OpenPlans

This report was sponsored by the University Transportation Research Center at City University of New York.

Thank you to our friends at OpenPlans for their input.