Making Open Transportation Data Useful and Accessible: Recommendations for Good Practices in Open Data Standards Management.

Elizabeth Sall*
UrbanLabs LLC
4405 4th Ave NE Seattle WA 98105
919 302 0265
easall@gmail.com

Lisa Zorn
Metropolitan Transportation Commission
375 Beale Street, Suite 800 San Francisco CA 98105
415 778 6644
lzorn@mtc.ca.gov

Drew Cooper
San Francisco County Transportation Authority
1455 Market Street 22nd FL San Francisco CA 94103
415 522 4800
drew.cooper@sfcta.org

Bhargava Sana
San Francisco County Transportation Authority
1455 Market Street 22nd FL San Francisco CA 94103
415 522 4800
bhargava.sana@sfcta.org

Stefan Coe
Puget Sound Regional Council
1011 Western Ave #500 Seattle WA 98104
206 464 7090
scoe@psrc.org

Word Count: 6,773 + 250 = 7,023

Submitted to the 2017 Transportation Research Board Annual Meeting in response to the Standing Committee on Library and Information Science for Transportation's Call For Papers: *Innovative Transportation Information & Data Tools and Practices*

Submitted August 1, 2016

ABSTRACT

Effective data standards are critical to ensuring that open data is useful and accessible to its intended audience. However, historical standards making organizations are using processes that aren't agile enough, accessible, or relevant to the variety and quantity of data that is being introduced from the multitude of recent open government data policies and pushes. This paper examines a variety of methods for creating and managing open data standards to ascertain best practices. After evaluating the applicability of these best practices to a real life standards-developing need, the project team concluded that there were significant leadership needs in the data standards for transportation analysis arena.

INTRODUCTION

While open data policies and advocacy has resulted in an explosion in the number of available datasets across a variety of sectors, some groups in the open data movement have moved beyond an initial push to get as much data online as quickly as possible to increase *access*, towards measuring success by how open data is *used* to support answering questions to solve a particular problem (*1*). Going one step further, researchers have been investigating not just if open data is used, but *by whom* (*2*) and have suggested the creation of a new digital-divide between those who have the skills and infrastructure to make use of the data for their benefit and those who don't. These researchers have suggested that in addition to free and open data, that quality documentation, technology access, and technical assistance must also be free and open in order to empower those not already empowered. These issues can be broken into supply- and demand-side issues, with this paper focusing on supplying the data in such a way that it is truly *open* to the most potential users.

Most open data policies and manifestos published by local and federal governments as well as advocacy groups list factors in making the open data supply *accessible* (*3*). Some of these are straightforward and well defined such as requirements for machine readable formats and permissive use licenses. Less well defined or mature are mandates for meta-data and the use of *standards maintained by a standards organization* (*4-5)*. The Office of Management and Budget's Open Data Policy (created in response to President Obama's Executive Order (*6*)) points to a "common-core metadata schema" that is further defined on its website. A *data standard* is a set of rules on the format and meaning of data to facilitate its sharing and exchange. The Open Knowledge Foundation, an organization devoted to promoting the free exchange of information defines an *open data standard* as one that is free from licensing restrictions and developed to be vendor-neutral (*7)*. After giving a little bit of background on who makes standards and how we might be able to evaluate them, this paper analyzes several case studies of data standard development and management, recommends a set of best practices from them, and evaluates how implementable the best practices are with a new data standard creation project.

Background: Who Makes Data Standards?

So who makes all of these standards that these recommendations discuss, and how do we make sure they are developed and managed in a way that will make sure they remain useful to the appropriate audiences? Starting fromt the top-down, there is the International Standards Organization (ISO), the American National Standards Institute (ANSI), a variety of other "accredited" standards development organizations (SDOs), "non-accredited" official standards development organizations, companies, and ad-hoc, or loose standards. The Federal Government has had limited, but strategic involvement in standards-making per *OMB Circular A-119: Federal Participation in the Development and Use of Voluntary Consensus Standards and in Conformity Assessment Activities* (*8*) and clarified in the 2012 OMB Memo M-12-08 *Principles for Federal Engagement in Standards Activities to Address National Priorities (9)* which

reaffirms that the development of standards be left to the private sectors **but for** instances of national priority identified in law where government may take on the role of the catalyst to help progress a market-based solution.

The most recognized body for standards is ISO, a non-governmental body that develops standards within over 250 different technical committees. Currently, the only standards related to the transportation sector are related to vehicle communication, intelligent transportation systems, and shipping. ISO has four key principles in standards development: responsive to the market need, based on global expert opinion, participation from multiple stakeholders, and based on a consensus (*10*).

ANSI is a member organization of ISO, but has the specific mission of progressing the US economy by promoting and progressing standards. ANSI accredits the processes of over 200 standards developing organizations (SDOs) that have created over 10,000 standards. ANSI requires that standards be developed by consensus of those affected by them, in a public manner, and with due process. (*11*)

Almost all of the accredited SDOs (*12*) are industry-specific groups or professional organizations such as the American Society of Civil Engineers. All of the two-dozen ANSI-accredited SDOs that the team reviewed as a part of this project conducted their activities via static web-postings of PDFs or in a non-public webspace, accessible by request. While technically fulfilling the definition of of "open", they are not conducive strategies to encourage dynamic dialogue or friendly challenges that would be the hallmark of an agile, open standard.

The federal government does sponsor a number of standards-creation or endorsement activities in order to advance items of strategic importance. This includes the Federal Geographic Data Committee (*13*) which both makes and endorses standards.

Finally, there is no shortage of "open" data standards, successful and not, developed at companies (e.g. GTFS at Google) or by individual projects or unaccredited organizations (e.g. Open Street Map).

METHODOLOGY

The approach to developing recommendations regarding open data standards is based on evaluating a set of case studies both in and adjacent to the transportation industry to glean a set of best-practices. This paper evaluates how implementable these best practices are within the context of a project from which we derive a series of recommendations and conclusions.

In order to identify characteristics and practices of successful data standards, the team first identified a handful of measures that would separate out successful from unsuccessful data standards and then evaluated a handful of examples from the transportation industry and beyond.

Measuring the effectiveness of a data standard could be done in a myriad of ways. One way, is to measure the *market share*, or the degree to which the standard **is** used when it **could** be used. Another, is how *compatible* the standard is with other popular data schema and standards that operate in the same ecosystem. *Rigor*, whether a standard is clear and well defined, is important because otherwise isn't much use and could result in real damage. We have infamously sent rockets on the wrong trajectories because of non-rigorous data standards. The next set of methods for measuring a useful data standard are likely to be more qualitative in nature. *User satisfaction*, or the degree of enthusiasm with which people use the standard can be used to distinguish a much despised, but entrenched standard from a sparsely used, but beloved one. *Approachability,* or the level of information and technical pre-requisites, can help address concerns about increasing the digital divide.

Depending on what a standards is used for and its *audience,* some of these measures will be more important than the other. For the purposes of this paper, we define the following audience types:

- Technical: completely backend and never intended for public consumption and use. Market share, compatibility, and rigor are the most important measures here. Example: Vehicle data, USB Specs.
- Technical Consumer: approachable to people with no specific industry training, but general technical abilities. Approachability becomes more and rigor slightly less important measures. User satisfaction becomes more important because people could easily defect to another standard of their choosing which would cause market share and interoperability to decline. Example: GTFS, HTML.
- Public Consumer: hobbyists with no particular training can easily understand and use. Measures other than approachability are much less important as the datasets are usually simple enough and there is little likelihood that they need to interface rigorously with other data. Example: Street Tree List

Because there is less rigor required for public consumer-oriented data standards, the discussion and case studies will focus on technical and technical consumer-oriented data standards.

CASE STUDIES

The following cases span industries, audiences, and levels of success and touch upon not just the development of the data standard, but its management. Some of the cases pertain to specific data standards while others identify how an organization responsible for multiple data standards operates.

Universal Traffic Data Format (UTDF) - Not Open Enough

Trafficware created the UTDF to import and export data to and from its Synchro traffic engineering software (*14*). Synchro and the format are widely used by traffic engineers and the format has seen support from a variety of other open source projects such as DTALite (15). However widely used, it is not publically managed or even documented and is thus not an "open" format in the common use of the term.

TransXML - Lack of Persistent Management

The TransXML standard was developed as cooperative research program project NCHRP 20-64 and published as NCHRP Report 576 (*16*). While the specifications developed as a part of this project are technically open, there does not appear to be widespread use of any of the formats put forward. There are three primary reasons that this specification failed to be popular. First, it was developed using a typical contracting relationship rather than a truly collaborative process involving the actual users of the standard. Second, there is not any evidence of continued support to embrace the standard: the last update to the TransXML project website was in 2006, shortly before the official NCHRP project concluded. Finally, the standard itself wasn't very accessible. It was buried in a large, less searchable PDF report with no code validator, example data, or copy-pasteable text.

General Transit Feed Specification (GTFS) – Outgrowing its Original Purpose

GTFS (17) was created by and is managed by Google Inc., which invested a fair amount of resources in making it free and very easy to convert existing proprietary transit service formats to GTFS, to make it easy for Google to use the GTFS data in their maps application. Google used a very accessible file format

that could be opened up and read by just about everybody with a computer (CSV), and a vocabulary that made it extremely human-readable at the expense of computationally efficient (e.g., "bus" rather than some integer code that means bus, and "Mission St/24th St" rather than some code). Their format is well documented on their website and there is a Feed Validator that allows users to view the files in their maps application to check for errors.

Despite GTFS being very ubiquitous, its user satisfaction is not as high as it could be. One reason could be that while declaring itself "not set in stone", there is very little that Google does to actively manage the format to respond to new and growing needs.

In a sense, GTFS is a victim of its own success. Its ubiquity and availability has opened it up to new and growing audiences who need more and more from the standard. And Google, this being outside of their main line of business, has not adequately responded to the demand. Changes to the specification are requested through an email listserv without an obvious release structure or version control system outside of an on-going change-log. While GTFS has declared itself to always remain backwards-compatible (a double-edged sword), version control is perhaps paramount in a good data standard. Oddly, Google does have its GTFS Feed Validator under version control, but the Feed Validator does not use GTFS version (because there is none) as an input.

Open Street Map (OSM) - Too much help, not enough central power

OSM's XML-based format uses a wikimedia-flavored wiki for documentation, which is both extensive, easy for editors to update, and easy for editors and users alike to view the history/changelog of the page (*18)*. The data vocabulary is well documented with examples. The wiki notes the various pros and cons of using an XML format: while it is fairly human and very machine readable, it is not space-efficient. As a space-saving, but illegible alternative, OSM also provides a Protocolbuffer Binary Format (PBF). There are two main deficiencies with the OSM data standard itself: (1) its lack of interoperability with existing standards or data vocabularies; (2) technocratic nature of the membership has led to, despite their best intentions, an "ungrokable" data set by people with a normal level of technical skill.

The OSM data format is managed by the OSM Foundation, a bare-bones, mainly volunteer effort. Despite their best efforts, the Foundation's underwhelming and confusing primary support is often trumped by superfluous and conflicting secondary documentation, often provided by a for profit company. The OSM Foundation allows for a well-documented, formal public proposal process for adding tags and relationships, but it is clunky to navigate and the open member voting allows for people to reject things with unhelpful comments such as "history is just plain wrong". It would be more helpful to have a smaller, dedicated group of people who are vested in the standard to be the primary stewards of any changes. Finally, while these changes seem to be much more dynamic and responsive to user needs than GTFS, there is no version control of the standard to speak of.

Traffic Data Exchange Format - Fragile but exciting

A consortium of not-for-profit entities (The World Bank) and companies who had a history of collaborating together on open transit data using GTFS sought to create an open data standard that would allow them to display traffic data in map tiles (19). While well documented with examples and managed and available on GitHub, the specification itself is too specific and hasn't yet considered a variety of use cases. It is also fragile as it is overly dependent on specific revision numbers in OSM data, which exposes a flaw not just in this format, but in OSM's as well. While the collaborative effort to create a

data standard for transportation performance measures is promising, it needs some more thought and investment in order to take root.

OpenTrails

OpenTrails is a project incubated by Code for America to define a geojson-based data structure for recreational facilities (20). It is new, but has a fair amount of traction and a good consortium of partners which gives a good indication that it will soon have substantial market share over the existing spread of various GIS formats residing on people's desktops and an existing federal trail standard that is too narrowly defined and buried deep in government-controlled PDF documents. OpenTrails is well documented via its website in an approachable format and allows the public to comment and make suggestions to its specification via a google document. Unfortunately, while these comments are likely read and incorporated to some degree, all decisions about the specification are made in a temporally-sorted listserv, making it difficult to look up a change-request for a particular field. There is also no way to compare versions of the standard dynamically or even via a changelog.

One useful tactic that the OpenTrails team used was to have a defined "Request for Comment" period between version 1.0 and 1.1 and then a freeze on development for a period of time after version 1.1 was released to allow users update their data. OpenTrails, similar to GTFS, also publishes a data validator and example data. However, they are confusingly maintained in the same Git repository as the main OpenTrails website.

W3C - Necessary for experts; ungrokable for newcomers.

The World Wide Web Consortium (W3C), the organization that standardizes the bulk of the soft internet, develops standards within working groups. While most of the W3C work takes place within a mediaWiki instance, each working group undertakes the task of discussing and iteratively developing the standard differently, but there has been a trend recently to shift the development to a public GitHub repository. As an example, the Spatial Data on the Web Working Group (*21*) maintains their normal MediaWiki site, but also has a GitHub repository (*22*) where their standards are actually worked out and documented. W3C maintains an "official site" for each standard along with links to the previous or new versions if applicable. They number the release of each standard (e.g., HTML 5.0, 5.1, etc) but use dates for small changes made between such as HTML 5.1 from September 29th 2015. Perhaps most interesting is their use of a strict framework, the Resource Description Framework (RDF)(23), for defining relationships and ontology and a language, Turtle, that is capable of defining the RDF in plain-text. The biggest drawbacks from the W3C process are (A) the variety in ways in which the standards are developed makes it difficult to understand how to participate in the process in a meaningful way, and (B) the process and the standards themselves are designed for experts. They are complicated to comprehend.

In addition to its work on standards, W3C has developed a draft *Data on the Web Best Practices* based on a number of case studies researched by the Data on the Web Working Group. Best practices pertaining to data standards include recommendations to use multiple, open, machine-readable, standardized data formats to promote interoperability, and to reuse existing data vocabularies when possible.

Internet Engineering Steering Group (IESG) - Useful and Stylish Documentation for Expert Technicians

IESG manages the internet's engineering standards process, which dates back to the days of DARPAnet. The process for developing internet standards is documented by its own standard: RFC2026(*24*).

Standards are developed through the "Request for Comment" (better known as RFC) process. All RFCs are maintained on www.rfc-editor.org and are assigned a serial number. Not all RFCs are standards, but those that are go through an evolution from *proposed -> draft -> standard*. Standards each have serial numbers (e.g. STD 77) and new RFCs can update or make old standards obsolete. So the STANDARD has a number and the RFC number is changed to reflect changes to that standard.

RFC-editor maintains a rolling list of status-changes. The problem is that if you are using a canonical reference to the RFC, you might not ever know that anything has changed or moved forward.

There is a very popular RFC "style guide" (*25*), which is used for defining many standards across many disciplines, not just IESG. IESG also allows another type of RFC called "best current practice" (BCP), which are not standards, but define the coalescence around a certain way of doing things in practice.

Frictionless Data

Frictionless Data is a project of the Open Knowledge Foundation, a not-for-profit dedicated to making knowledge accessible and making sure people are able to use it (*26*). One of their key contributions is the definition of a Data Package including a meta-data standard. One of the most interesting things about these standards is that they are developed and published on the web using a GitHub repository that seamlessly creates a Jekyll-based webpage based on an RFC-style specification that lives in a version-controlled markdown file. Because the entire operation is in GitHub, you can easily access a change-log and do key differences. Moreover, it is very human-readable and easy to interact with using GitHub's built-in markdown renderer, issue tracking system, and commenting system. It is also easy to use when downloaded *from* GitHub, since it just uses ASCII text.

This approach is nice for several reasons. First, it forces any change to be noticed. Because you are using version control software, you can't 'cheat' and slip in something else to a previous version. It is ASCII-based, which allows for easy viewing of differences across versions of the standard. It is human-readable and allows for more formatting than the pure ASCII view of RFCs. It has a seamless translation between the version-controlled ASCII and the front-end webpage which reduces possible error or misalignment. Finally, it leverages GitHub's toolset which will allows you to subscribe to a repository which will let you keep abreast of any changes.

DISCUSSION

These case studies have a broad range of success because of or in spite of their origin story and continued management strategy. Despite the broad number of "best practices" in open government data management that point to using defined, open, and interoperable standards, there is little detail aside from broad ideals that pertains to the development and management of these standards. Additionally, because the vast amount of data collected at various levels of government does **not** likely already have a relevant standard to point to, there needs to be a large increase in the number of standards successfully developed and maintained in order to meet this need.

The primary takeaways for technical standards from a review of these case studies are:

1. Developed and managed by a trusted source with permanence, customer fcus, and with sufficient user involvement
2. Leverage existing data vocabularies
3. Right-size the standard (and its management) to the audience
4. Evolve the standard at right pace and using rigorous methods

5. Limit unnecessary tools and libraries
6. Diligent documentation of standard and process
7. Balance flexibility while limiting vocabulary dispersion
8. Structure and tools that limit and catch errors
9. Promote your standard to make sure industry knows it is there

The next section evaluates how implementable these recommendations are by trying to implement them when designing a new data standard based on a project need.

IMPLEMENTATION EVALUATION

When three public agencies, the Metropolitan Transportation Commission (MTC), Puget Sound Regional Council (PSRC), and San Francisco County Transportation Authority (SFCTA) decided to implement a user-ready version of some software developed at a University, one of the first tasks was to develop a set of interoperable data standards that could ensure that all three agencies (plus others) would be able to use the software. The software, Fast-Trips (*27*), does person-based dynamic transit passenger assignment and requires passenger demand and transit networks as inputs.

This section evaluates how this interagency team considered the case study findings to develop both the GTFS-PLUS transit network (*28*) and the dyno-demand demand (*29*) data standards and evaluates the degree to which the case studies findings are implementable. These standards were primarily developed for a purely *technical* audience, but envisioned to be useful to a broader, *technical consumer* audience thus accessibility was highly valued.

Developed and Managed by a Trusted Source with Permanence, Customer Focus, and with Sufficient User Involvement

The first problem the Agency Team faced was that there was a need for a standard "as soon as possible," but there was no relevant, existing SDO with the technical domain expertise within which to officially conduct the business of standard development. The Transportation Research Board (TRB) has relevant domain expertise among its volunteers and has previously directed standards-development activity, but the timeline for this was years, not months. The American Society of Civil Engineers is the closest SDO to this domain, but their process was not conducive to the timeline or technical needs of the project. On the transit networks side, the team considered trying to make changes to the official GTFS specification, but that process as previously noted is ad-hoc and many of the variables that the project needed did not fit within the requirements that Google has put forth for new variable names.

As a workaround to these issues, the team developed the data standards publically under the auspices of the "Open Source Planning Data Standards" GitHub handle and worked to fully document the standard beyond the needs of the project.

Findings:

- Need more SDOs (accredited or not) to cover areas of expertise where open data is expected to be released, but not too many such that it isn't clear who is responsible.
- Need a directory of SDOs by domain and a standardized process for "involving appropriate users" so that standards aren't re-developed.
- Standards creation activities need to be responsive to the timeline of project needs: weeks, not years.

Workaround:

- If you need to develop the standard within the context of a project without a home organization, separate the standard as much as possible from the project and prepare to be flexible and evolve the standard to other users' inputs.

Leverage Existing Data Vocabularies

Given the general industry coalescence around GTFS, the team decided to start with its standard and data vocabulary in order to maximize the overlap in required data and leverage existing tools. While GTFS contains a significant amount of data, it doesn't have all the detail or structural relationships required for our project. However, the changes required would both take months to years to make within the existing GTFS management process as well as break the rules for GTFS to be backwards compatible.

At the outset, the team decided to make everything that was mandatory in GTFS mandatory in the new standard even if it wasn't needed to run Fast-Trips. Starting with GTFS meant an implicit adoption of a CSV-based format with a header line containing case-sensitive variable names. In addition to file type, in order to utilize existing GTFS validator and visualization tools, the existing GTFS files needed to stay intact and not add unexpected fields. Therefore, the team adopted a transit network standard that had additional files with supporting information.

Generally speaking, for every GTFS file, there is an additional file with the additional fields as well as a identifier key to link it back to the GTFS information. For example, **`stops.txt`** would be a standard GTFS file, but and **`stops_ft.txt`** contains a variable **`stop_id`** to link back to **`stops.txt`** as well as the additional information needed for Fast-Trips such as the presence of shelter, lighting, seating, etc.

Finding:

- Existing data standards that the industry has coalesced around may not be perfect.

Workaround:

- Don't throw the baby out with the bathwater. Reuse existing data vocabulary and software by "extending" the standard rather than creating a new one.

Right-Size the Standard (and Its Management) to the Audience

In the Fast-Trips case, the team was primarily developing this set of standards for a purely technical audience of professionals in the travel analysis industry with the technical public a secondary audience. That said, the in-depth and rigorous standards development processes found at W3C, ISO, and the Open Street Map Foundation would likely scare away all but the most technical of the technical consumers. Nobody is likely to read a several-hundred-page specification document or read through a document written in the Turtle language to evaluate if the right variable ontology has been used. Thus, the team tried to strike a balance between documentation rigor and approachability and limited the introduction of new tools, skills, and terminology that people in the travel analysis industry didn't already use.

Finding:

- Rigorous and overly detailed processes and standards documentation would likely deter all but the most technical and determined individuals in our professional niche.

Workaround:

- Use tools and terminology already familiar to the industry.

Evolve the Standard at Right Pace and Using Rigorous Methods

During the first few months after an initial draft, the standard was in constant flux as the team put it through the paces of use in the real world. This flexibility in the rapid evolution was necessary in order to get a standard that worked for the purposes of the project but would be completely untenable to a broader audience who would be trying to develop tools that could also interact with the standard.

In order to mitigate this, the team adopted semantic versioning (http://semver.org/) where anything before version 1.0.0 should be considered in pre-release and subject to change at any time. Once version 1.0.0 is realeased, any changes that break backwards compatibility should be contained to annual or sparser events. Implementing strict version releases also allows rigor in defining requirements from a software perspective (e.g., this software requires data be defined using standard version 1.2.0 or higher).

In order to enforce rigorous version control and leverage their existing cooperative development tools, the team decided to put the standard on GitHub as Markdown-flavored ASCII text. The git version control software that GitHub is based on enforces that every version is stored and allows for easy comparison of differences between versions.

The collaboration tools available on GitHub such as issue-tracking (see Figure 1) and resolution allowed the team to raise issues with the standard, discuss fixes, and quickly implement resolutions in the standard itself. GitHub offers substantial benefits over MediaWiki, which requires substantial initial setup to achieve the structure already found in W3C and Open Street Map Foundation. GitHub is also free for public projects whereas MediaWiki requires a server that would incur ongoing maintenance costs.

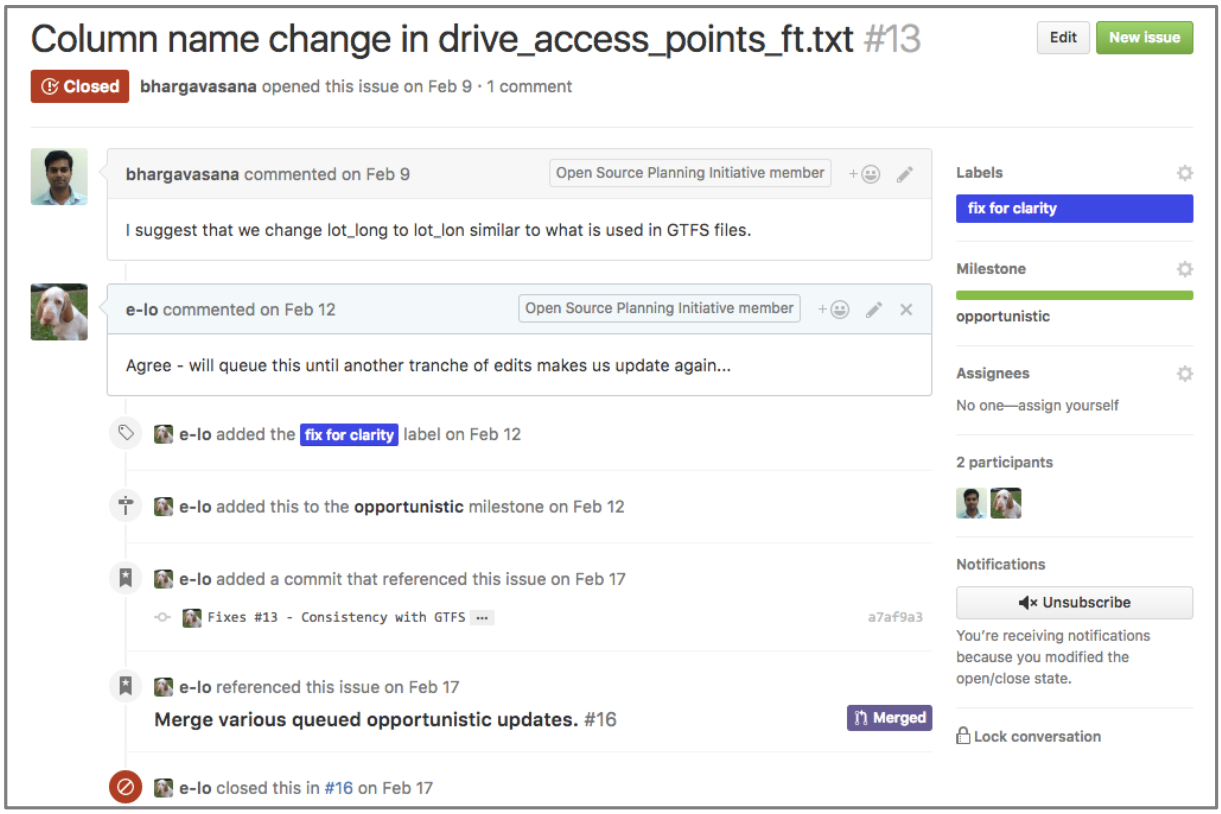

Figure 1 Example of Issue Management in GitHub

Findings:

- Use semantic versioning and be explicit about pre-releases using version number less than 1.0.0

- GitHub provides appropriate built-in tools for writing, managing, and collaborating on data standards without the initial setup time and continued cost of using MediaWiki.

Limit Unnecessary Tools and Libraries

When considering file formats, the Agency Team seriously considered HDF5 which would offer a highly compressible format for the large demand files and SQLITE which has a substantial number of built in query capabilities. However, it was telling that all three agencies didn't have any one of these tools installed in their standard computing environment and thus the team adopted a CSV-based standard that could be read and edited in a normal text editor if desired, or easily read in using standard libraries to almost every computer programming environment. The CSV-based format has the added benefit of being able to version control the data itself.

Diligent Documentation of Standard and Process

Each standard is documented using RFC-style text in a GitHub repository. In the repository base is a README.md file that has the following sections:

- Version
- Date last updated
- Date created
- Authors
- Changelog
- High level Known Issues
- Files that MAY or MUST be contained in the data to comply with the specification, with links to descriptions of those individual files.

Within a subfolder /files are separate ASCII-based markdown files describing each of the mandatory or optional files using RFC-style text describing the format type, and required and optional variables.

Balance Flexibility while Limiting Vocabulary Dispersion

The whole point of having standards is to come up with a set of expectations about the information that will be provided. However, even between the three participating agencies, there were often three different methods for approaching problems which required differences in the required data. An example of this is park and ride lots and whether they should be included as a singular lot, a list of lots, or not as an input at all with the travel demand. The team managed this particular issue by including the park and ride lot as an optional field which can contain a list of lots, including a list that just contains a single lot.

Another issue was coming up with a common vocabulary to talk about the same thing that was recommended, but not required. For example, mode names could be light-rail or light_rail. Our standard suggests (but does not require) that you use light_rail so that hopefully, eventually, everybody will use the same vocabulary, but it won't prevent you from using it if it doesn't work for your application.

A final strategy in accommodating a multitude of needs was to make only the very basic required data "required" as part of the standard and everything else optional. This is a double-edged sword because while it will make the standard accessible to many agencies who want to use it and don't have the data to support it (not every agency has information about every escalator or bus stop bench), it will create a very inconsistent level of data.

Findings:

- Small differences in workflows can sometimes require significant differences in data structures

- Flexibility in mandatory fields will result in wide variety of data being available and limit the scope of interoperability

Workaround:

- Specify different fields as mandatory for different applications. i.e. dyno-demand, platinum-level could mean that all mandatory and optional fields are filled and gives you access to the broadest number of uses.

Structure and Tools that Limit and Catch Errors

Limiting the number of times something is duplicated will reduce the opportunity for conflict and limit the number of places that data needs to be edited. The only place that data should be duplicated is when it serves as a key between files. Sometimes, as in the case of dyno-demand, this means separating out data into key component files: person, household, and trips so as to not necessitate duplicating person and household attributes across every row in the trips file. While this requires a bit of manipulation on the software side, it reduces file sizes and the opportunity for error.

Limiting the propagation of errors was another key reason to decide to allow plain-worded string types for variables like travel mode. Using number-string lookups for common words like modes, vehicle types, and trip purposes would save a small amount of space and input read time. However, if one opens a file and sees "`1, 88`", one is much less likely catch an error compared to reading "`commute, pogo stick`" which would immediately alerts one that there could be an error.

Finally, the data standard should include a validator that reads in the data and lists both fatal and potential errors. Fatal errors means that the data does not comply with the minimum standards. Potential errors are a list of things that you should consider before proceeding. These might include unreasonably fast or slow buses on the GTFS-PLUS side, or values-of-time outside normal parameters on the demand side. While the Agency Team did not write its own validators within the standard (instead using Google's GTFS validator), it is something that should be considered in the future.

Promote Your Standard to Make Sure Industry Knows It Is There

Lastly, no data standard usefully exists in isolation. In order to let people know what the team was working on, it was broadcast to the relevant user groups like the Travel Model Improvement Program Listserv as well as published on our project blog (http://fast-trips.mtc.ca.gov) and presented at the Transportation Research Board Innovations in Travel Modeling Conference and at informal conferences such as Transportation Camp. The team has also worked with other agencies interested in using the standard. That said, the standards need to have a home and team behind it that last longer than the length of this particular research project, and thus a SDO for the travel analysis industry must be developed in the short term with collective buy-in from the broad industry.

CONCLUSIONS

Based on the review of case studies and the Agency Team's experience in developing and managing their own data standard, the following conclusions have been reached.

Leadership Needed

Open data standards are an important and required component of making open government data useful and accessible, but are currently more of an afterthought in the process. Rather, the data standard should pre-date the data release. This currently isn't happening for a significant amount of data in the transportation sector because there is a lack of official coalescence around a standards development

organization that has standards development in its facility, mission, and funding. Rather, there are several ad-hoc standards developed, each without significant market share and limited use cases and several others managed by corporate interests. While the government has purposefully (and rightfully) decided to take a backseat role in standards development, there is a role for a third party leader in modern data standards creation and management that needs to be funded, empowered, and filled. It should be emphasized that having strong leadership is not at odds with having a user-driven process. Rather, strong leadership is required in order to facilitate a user-driven process that is accessible to the appropriate audiences

Tools for Open Standards Development Are Now Easy to Use and Accessible

The Agency Team found GitHub to be an ideal tool to manage the development and documentation of a data standard and much preferable to the start-up effort, cost, and continued maintenance of MediaWiki. GitHub (and other tools) are free to use for public projects and obviate any excuses for conducting matters behind closed doors/email groups/ or in indecipherable PDF document dumps.

Using Best Practices Will Increase the Usability and Accessibility of a Data Standard

This paper outlined a set of data standards best practices based on various case studies. The standard itself should use existing data vocabularies to promote interoperability, limit unnecessary tools and libraries, be thoroughly documented, balance flexibility and unneeded dispersion, and limit data duplication.

New standards should be developed and managed by a trusted source with permanence, customer focus and with user involvement. The standards developer should be aware of evolving the standard too fast or too slow and use rigorous version control and validation methods to limit errors. Finally, a data standard needs to be promoted and supported in order for the right audiences to know to use it and understand how.

ACKNOWLEDGEMENTS

The authors would like to thank the Federal Highway Program SHRP2 Implementation Assistance Program which funded this work as well as the Fast-Trips Implementation Management Team and project manager: David Ory, Billy Charlton, Joe Castiglione, Mark Simonson, and Diana Dorinson. Finally, the authors would like to thank Mark Hickman and Alireza Khani, the original Fast-Trips designers.

REFERENCES

1. Susha, Iryna, Åke Grönlund, and Marijn Janssen. “Organizational Measures to Stimulate User Engagement with Open Data.” Transforming Government: People, Process and Policy 9, no. 2 (May 18, 2015): 181–206. doi:10.1108/tg-05-2014-0016.
2. Kaasenbrood, Maaike, Anneke Zuiderwijk, Marijn Janssen, Martin de Jong, and Nitesh Bharosa. “Exploring Factors Influencing the Adoption of Open Government Data by Private Organisations.” International Journal of Public Administration in the Digital Age 2, no 2 (2015): 92. Doi: 10.4018/ijpada.2015040105.
3. Burwell, Sylvia, Steven VanRoekel, Todd Park, and Dominic J. Mancini. “M-13-13 – Memorandum for Heads of Executive Departments and Agencies: Open Data Policy – Managing Information as an Asset.” https://project-open-data.cio.gov/policy-memo/.
4. U.S. Senate. 114th Congress. S.B. 2852, *OPEN Government Data Act*. Washington. Introduced April 26, 2016. https://www.congress.gov/bill/114th-congress/senate-bill/2852/all-info.

5. U.S. House. 114th Congress. H.B. 5051, *OPEN Government Data Act.* Washington. Introduced April 26th, 2016. https://www.congress.gov/bill/114th-congress/house-bill/5051/
6. Executive Order of May 9, 2013, Making Open and Machine Readable the New Default for Government Information. https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-.
7. "Open Data Handbook," Open Knowledge Foundation, http://opendatahandbook.org/guide/en/.
8. Office of Management and Budget, "Draft Revision OMB Circular A-119: Federal Participation in the Development and Use of Voluntary Consensus Standards and in Conformity Assessment Activities." https://www.whitehouse.gov/sites/default/files/omb/inforeg/revised_circular_a-119_as_of_1_22.pdf
9. Chopra, Aneesh, MIriam Sapiro, and Cass R. Sunstein, "M-12-08 Memorandum for the Heads of Executive Departments and Agencies: Principles for Federal Engagement in Standards Activities to Address National Priorities." January 17, 2012. https://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-08.pdf.
10. "International Standards Organization Website," http://www.iso.org/iso/home/standards_development.htm.
11. "American National Standards Institute," https://www.ansi.org/about_ansi/introduction/introduction.aspx.
12. "American National Standards Institute List of Accredited Standards Development Organizations," https://share.ansi.org/Shared%20Documents/Standards%20Activities/American%20National%20Standards/ANSI%20Accredited%20Standards%20Developers/JUNE16ASD.pdf.
13. "Federal Geographic Data Committee Website," http://www.fgdc.gov/.
14. "Trafficware Website," http://www.trafficware.com/transferring-data-using-utdf.html.
15. "DTALite Release Notes," https://github.com/xzhou99/dtalite_software_release.
16. Ziering, Eric, Frances harrison, and Paul Scarponcini, "NCHRP Report 576; XML Schemas for Exchange of Transportation Data." Transportation Research Board, 2007: http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_576.pdf.
17. "General Transit Feed Specification Website," https://developers.google.com/transit/gtfs/.
18. "Open Street Map Data Standard," http://wiki.openstreetmap.org/wiki/OSM_XML.
19. "Traffic Data Exchange Website," https://github.com/opentraffic/traffic-data-exchange-format.
20. "Open Trail Specification," http://archive.codeforamerica.org/specifications/trails/spec.html.
21. "W3C Spatial Data on the Web Working Group," https://www.w3.org/2015/spatial/wiki/Main_Page.
22. "W3C Spatial Data on the Web Working Group Github Repository," https://github.com/w3c/sdw/.
23. "W3C Resource Description Framework," https://www.w3.org/TR/rdf11-concepts/.
24. "IETF Request for Change 2026," https://www.ietf.org/rfc/rfc2026.txt.
25. "IETF Request for Change 7322," https://www.rfc-editor.org/rfc/rfc7322.txt.
26. "Frictionless Data Website," http://frictionlessdata.io/.
27. "Fast-Trips Software," https://github.com/MetropolitanTransportationCommission/fast-trips.
28. "GTFS-PLUS Transit Data Standard," https://github.com/osplanning-data-standards/GTFS-PLUS.
29. "Dyno-demand Data Standard," https://github.com/osplanning-data-standards/dyno-demand.