TONY LAIDIG

QUANTITATIVE CONCEPTS FOR ’GOOD’ TRANSIT SCHEDULE DATA

Disclaimer.

This guide is the result of a career's worth of experience with transit data. In the process, the author has worked with data from dozens of transportation agencies, of varying sizes, both in the USA and internationally. The information within should not be considered the result of experience with any one particular entity, and where not cited, the opinions expressed are of the primary author alone and not of any other individual, company, agency, government, or interstellar federation.

Introduction

> “The [tracking] program was put on hold in the fall of 2007, after [transit agency] announced that software problems were causing customers to receive inaccurate information”
>
> “One problem, developers said, is that some of the information [transit agency] provides developers for its bus system was missing data on route destinations. At other times, developers said, ’weird’ formatting or perhaps a software bug makes the name of a bus route — such as the 10A — appear as the time, as in 10 a.m”
>
> “Unfortunately, in its current incarnation, the system cannot differentiate when it is displaying schedule data due to real-time data being unavailable.”
>
> “[S]tops are labeled on the map, but all stations are a blue “M” with no regard to their corresponding lines. Indicators of stations with transfer possibilities are noticeably absent; nor can you [...] search for specific stations.”

Every person tasked with implementing an IT project dreads encountering quotes like those above. How then, in an era of where many technologies are off-the-shelf and systems engineering is prevalent, do these problems keep happening? The answer is the unconsidered effect of bad or marginal data on what may be otherwise good projects.

This document is intended to provide language for a dialog between parties involved in transit data creation and usage through practical qualitative concepts, with the intention of making currently intangible risks more obvious.The following groups may find it useful:

Intangible risks have a high probability of occurring, but are not identificatied and thus unmitigated.

- Project Managers, Consultants, and Business Analysts who are tasked with estimating resources for a project using available and newly created data sources.

- Technologists, in consultation with their project team and stakeholders, who are looking for a guidance in measuring available data quality, potential gaps, and developing solutions to bridge those gaps.

- Analysts, who may need a simple framework to describe issues with data to its maintainers and owners.

A PACT for Good Data

Many agencies have existing, and new projects are often thought up believing that existing data will serve new purposes without major rework. Sadly, is often not the case. This chapter presents four crucial qualitatitive concepts that are important in describing the qualities of a system's data. These are Precision, Accuracy, Concurrency, and Transparency, or, in manager speak, PACT. By reviewing the quality of the datasets involved according to these four concepts, project teams can address possible concerns before beginning development, with the aim of reducing errors during the implementation stage. The quantification of these concepts is not addressed. The exmples presented use an ubiquitous dataset in almost every transit data project, a dataset of stops.

Precision and Accuracy

Statisticians and Geospatial scientists use two terms primarily to describe the quality of a dataset— *accuracy* and *precison*. The *accuracy* of a dataset is the degree of closeness that a record or observation matches the real-world value. *Precision* refers to the level of refinement available in the dataset to store said record. A stop dataset provides a simple case for differing levels of precision and accuracy, and these two concepts are described in the margin.

A dataset can be accurate but not precise, precise but not accurate, neither, or both. Some organizations conduct expensive, highly detailed stop surveys which are both precise and accurate at the time of their creation, but these lose accuracy over time. It is not uncommon for the very first survey to be considered only as part of a capital project, and the organization in question believes that the work is done. This, however, is often a folly. Due to changes on the ground, accuracy of a stop dataset degrades over time if it is not updated. Streets are paved and closed, businesses open and close, and in some areas entire neighborhoods appear overnight.

Precision of an example stop dataset, in ascending precision:

- A stop is at Main and 1st Street.
- A stop is at Main and 1st Street, Northbound.
- A stop is at Main and 1st Street, Northbound, NE corner.
- A stop is at Main and 1st Street, Northbound, NE corner, with coordinates X, Y.
- A stop is at Main and 1st Street, Northbound, NE corner, with coordinates X.XXXXX, Y.YYYYY, served by Route 101.

Accuracy of the example stop dataset is dependent various factors, including:

- Notification of changes
  - e.g. "The stop at Main and 1st was removed due to construction"
- Master database update frequency
  - e.g. "Stop changes are entered into the database within 1 week."
- Frequency of exports and releases
  - e.g. 'GTFS is generated at least every 4 months."

Concurrency

While precision and accuracy are important when considering a project that only uses, or heavily relies upon one dataset. The elephant in the room in projects that draw upon data from multiple systems is the *concurrency* of that data. Concurrency, in this case, refers to the ability to join together, reliably, the data from one or many systems.[1] A medium to large-size public transport operation may have an IT workflow of many stages, covered by different departments or even different organizations or agencies, public and private. The less that the systems involved are concurrent, the more difficult building a new system upon those projects will be. The design of concurrent systems often entails finding reliable techniques for coordinating their execution, data exchange, and execution scheduling to minimize confusion and maximise utility[2]. Data from systems originally designed to work independently may need some finesse to properly match.

The Enemies of Concurrency

While it is difficult to quantify concurrency, it is possible to suggest characteristics of 'weaker' data sources that should be validated before relying upon them. Some of these are:

- Possibility of unsanitized / unvalidated data entry
  - If an operator can enter data that is invalid or non-sensical, then the system that uses it must take this possibility into account. Policies may need to be put in place to ensure correct entry.
- Long update or synchronization intervals
  - A dataset that models frequently changing datapoints (e.g. bus stops) that is only updated every several years is likely to have errors.
- "Static includes"
  - Systems may rely on one particular file as an input not coming from the same repository as other souces; it may be either as an export from an otherwise unconnected system, or a hand-updated file. If policies for updating the file in question (e.g. on a regular basis or simultaneously with other exports), the data in question may fall out of concurrency.
- No / limited versioning

[1] Systems may not even be concurrent with themselves. Consider a large organization that builds their schedule in phases; one area may be ready for production while the rest of the network is being worked on. For example, Routes 1 and 2 stop at 1st and Main St. Route 1 has its schedule generated before Route 2. When Route 2 is scheduled, the scheduler has been notified that there will be long term construction at 1st and Main and buses cannot stop there; she then removes the stop for Route 2 only. When data is export from the scheduling system, Route 1 will still be recorded as stopping at 1st and Main, even though the network has been changed.

[2]

  - Datasets should clearly include the times they were updated so that when two datasets are joined they can be sufficiently linked.

Transparancy

Hopefully, when working with data from an existing system, there is documentation on what that data means. Often, however, when the data has only been used for only one purpose or within one organization or department, documentation is not complete. *Transparency* refers to the degree to which the data conforms to the documentation provided.

- Poor and/or obsolete documentation
  - Systems may be with good documentation, but as time goes on minor changes may be made to both the data format and the usage of that data. Documentation is not always updated in accordance. Well-intentioned users may extend the existing data structure beyond the original design by adding additional information that does not strictly conform to original specifications.This information may or may not be passed down through word of mouth, and is often internalized to the point that when asked, the data owners suddenly recall the difference.
  - Organizations often take advantage of 'turn-key' systems that include an openly-accessible database component often add additional functionality 'around the edges.' It is the author's experience that these additions are rarely documented.
- Special cases
  - IT groups may tailor the same application for different users performing the same task within the same organization, often based upon varying user preferences.
- Excessive room for interpretation and/or lack of usage conventions.
  - If a data model allows for the same data can be stored in two places, it is often the case that different users see the data differently. For example, if a stop dataset model has fields for both STOP_ID (intented to be a normalized key) and STOP_CODE (intended to be a short, but human-readable descriptor) and users fail to understand this difference, users at may choose to populate these fields in ways that are not consistent across the dataset. A very basic example is presented in the margin.

* One operating division may prefer that their route numbers begin with a leading zero (e.g 01 vs 1), while another may not. Beware special cases such as this one.

STOP_ID	STOP_CODE
1	1001
2	1002
3	3
4	4

- Ambiguous, Undefined and/or complex key structure
  - a ’Primary Key’ uniquely defines the characteristics of each record in a dataset, and has to consist of characteristics that are unique. If the key structure for a dataset is not clearly defined or not enforced, making the proper relations within the dataset is more difficult and concurrency may be affected.
- Data intended for human visual consumption.
  - Input data should be machine readable, not just human readable. If a system has previously been used as the final point in a data flow with the sole aim of output for human interpretation, it is likely that in places data does not accurately conform to its specification. This, when combined with unvalidated entries, can be especially confusing during implementation.
- Unclear methods of interpolation.
  - When data from one system is interpolated by a "black box" process, the results are often unclear.

A common case of this problem is the interpolation of stop times from AVL systems. There are several methods for determining the time when a bus passed by a particular stop. If a system only presents one time, was it the arrival or the departure time? Was there any interpolation involved in calculating the time, or is it simply the time of the position nearest to the stop in question?

- Specifications that do not agree with established practice or common nomenclature
  - Developers are notorious for not reading documentation. If a format specification happens to borrow a term from the common nomenclature and use that term either in a more loose or strict fashion, it is likely that the generators of the format are using the field according to their definition.

A *Vehicle Block,* in common transit usage, is vehicle’s schedule that begins and ends at a depot. GTFS provides optional block_id information, specifically for when "a passenger can transfer from one trip to the next just by staying in the vehicle." It is not uncommon that GTFS feeds use the common definition, rather than conform strictly to the GTFS specification.

Changes: They are a’comin.

Given the complex nature of many transit systems and the differing requirements for PACT based on the type of project, the question is not if data will need changing, but when and how. When implementing a complex system depending touching party, it is often easier for all parties when the data producers and users agree on a policy on what data should be updated and at what moment. One framework that may be useful in quantifying the value of updates to data is to define impacts from service changes over three dimensions:

- (temporal) duration, the amount of time that the change will be in effect;
- (spatial) extent, the area over which the change affects; and

- (human) magnitude, the number of customers/passengers that are affected.

As with the proper level of PACT, there is no universal policy for data changes, and one should be developed on a project by project basis. The data owners should define the relevant impact classifications (short versus long duration, small versus large area, etc), and the proper policy will be specific to the organization's needs. One of many possible strategies is outlined below.

Temporal	Spatial	Human	Policy
Very Short	Small	Small	Do nothing.
Very Short	Large	Large	Create a "Service Alert."
Long	Small	Small	Update data with next scheduled update.
Long	Large	Large	Update schedule data ASAP.

Conclusion: Why does this all matter?

As new uses for existing data are required, the need for additional precision, accuracy, concurrency, and transparancy may increase with them. While data may exist, it is important to take into consideration the data's PACT given it's previous uses. Most large transit organizations have had a rough progression in systems along the lines of the following.

- Scheduling: Early scheduling systems focused primarily on driver schedules over geographic schedules, using a simple schematic network between them. Geography or intermediate stops were not taken into account.
- Detailed Scheduling: Modern scheduling systems include geographic information for paths and stops and has the ability to generate passing times for each stop, which allows for detailed GTFS creation.
- Computer Aided Dispatching: In order to dispatch effectively, locations of terminals and timepoints must be accurate. If a timepoint is on the wrong side of the intersection or block, the CAD system will produce incorrect schedule adherence in that vicinity.
- Static Trip Planning: Passengers need to know where to board and alight the bus.
- Real-Time Passenger Information: The location of every stop must be within reason, and every revenue trip must be accounted for.
- Reporting and Analysis: Reporting on service delivery requires the behind-the-scenes data (e.g. runs and blocks) to be correct, above and beyond the passenger information use case.
- On Board Announcements: Announcements made too late because of incorrect stop locations or for stops that do not exist render the system useless.
- Real-Time Trip Planning: Planning trips in real time, sometimes hours in advance, requires the greatest level of PACT from all of

the above– systems must concur for real-time information in the future to be valid.

- The “Connected Vehicle:” While this is still being defined, it wil surely require more PACT from existing datasets.

The progression of technology leads to greater needs from data, and often technology can grow before data is updated. Understanding this is a key to delivering both good data and good projects.