The Digital Matatu Project:
Using Cell Phones to Create an Open Source Data for Nairobi's Semi-Formal Bus System

Abstract:
In many of the world’s growing cities semi-formal buses form the basis of the public transit system. Yet little open and standardized data exist on these systems. The Digital Matatus project in Nairobi, Kenya set out to test whether mobile technology could be used to collect this essential data and translate it into widely used transit data standard (GTFS) to provide anyone with the ability to expand on and benefit from the data. The results of the work in Nairobi show that mobile technologies, particularly cell phones which are increasingly prevalent in developing countries, can be used to collect and deliver data in a modified GTFS format for semi-formal transit and hence provide wider access to transit information. Perhaps more importantly through our work in Nairobi we were able to identify the benefits and technical needs for developing a data set on semi-formal transit. The work therefore illustrates; 1) How General Transit Feed Specification (GTFS) can be adapted to semi-formal systems and used by others semi-formal systems. 2) There is demand from technologists as well as transport communities for data on semi-formal transit. 3) That releasing the data openly in a standardized format (GTFS) can help to encourage development of transportation applications. 4) That including a wide transit community during the data development can create a community of users and mechanisms for institutionalizing a process of data updating and sharing. The engagement strategies our research team developed around the data collection process in Nairobi became just as important as the resulting data it produced.

Keywords:
mobile data collection; cell phones; open data; GTFS; African transit; informality; GTFS; semi-formal transit; paratransit; Nairobi; Kenya; informal transit; data standards; international

1. Introduction
For millions in the developing world, citywide transportation options are often limited to semi-formal networks of bus or minibus, largely run by hundreds of diverse operators. Often referred to as paratransit, these systems constitute the backbone of mass transit for the majority of citizens in the rapidly growing cities of Africa, Asia and Latin America (Cervero 2007, Behrens et al. 2011, Guillen et al. 2012). System-wide maps of station locations, routes, fares, schedules, operating calendars and other key information for these systems are simply not available to the public for the vast majority of routes around the world. Lack of data makes it hard for users to know how to navigate these systems but more importantly creates limitations for transit planners who need this data for the development of transit models (Thakur and Sharma 2009, Barcelo et al. 2010). This is in stark contrast to cities with formal, planned transport systems where such
information is not only expected of operators, but is increasingly being integrated with new technology to make better transport planning decisions in real time (Catala 2011, Lee-Gosselin and Buliung 2012, Sussman 2005, Kramers 2014).

Semi-formal bus networks are composed of many private actors which, like taxis, operate for profit and are owned either by the drivers themselves or by businesses of varying size (Cervero 2007, Guillen et al. 2012). Vehicle size and capacity can vary widely, from small cars to full size buses (Zhang, et al. 2013). Unlike regular taxis, these paratransit bus systems often follow set routes with designated stops, much like formal transit systems (Cervero 2007). They provide essential transportation infrastructure to developing cities by providing mobility to residents especially the urban poor and lower middle class who often cannot afford other means of transport (Zhang, et al. 2013). While helping to fill a transportation gap, paratransit systems, have some drawbacks including contributing to traffic congestion, crashes, and environmental pollution (Cervero 2000) as well as unreliability and safety concerns (Author, 2014). However, the ability to provide transit where the government has not, offers some advantages, including demand responsiveness and flexibility as well as local ownership (Mutongi 2006, Woolf and Joubert 2013).

With the spread of mobile phones with geo-locative abilities and improvements in information and communication technologies, new possibilities are emerging to collect paratransit data by individuals at a dramatically lower cost. Key questions emerge out of these new technological developments: Can data be captured on paratransit systems outside formal institutional frameworks? What is the best way to collect such data? What data structure makes the most sense for the storage and distribution of paratransit data? Can the data be distributed using the GTFS (General Transit Feed Specification) which is largely used by more formal transit agencies to facilitate transit routing applications? Does a new data standard need to be developed that better captures the unique, often informal, aspects of paratransit systems? Who will use the data and for what purposes?

Our research team set out to answer some of those questions by testing whether the geo-locative capabilities of mobile technology could be used to collect data on a semi-formal transit system and determine if that data could be translated to the GTFS data standard allowing it to be more widely used by the larger transit and technology communities. We used Nairobi, Kenya as our case study and set out to analyze the city’s semi-formal bus system called matatus. Our research team collected basic route data from scratch using mobile devices. The data was then converted into GTFS, a standard widely used by transportation routing applications. The team worked with

---

1 In Nairobi, a number of technology entrepreneurs were starting to develop transit applications before we started our work but were generally not perceived as part of the “transit community” of planners, regulators, operators, insurers, mechanics, drivers, and passengers”. That is now changing, and we from now on include technology entrepreneurs in our category of “transit community”.

2 “Matatu” is the Kenyan name for the semiformal buses of varying size.
the GTFS community to develop changes to the GTFS format to accommodate the differences in the way the matatus and other semi-formal systems operate. The research team also worked with Nairobi’s transit and technology community to inform them about the data. Ultimately the local technology entrepreneurs extended the value of the data by creating mobile routing applications. The transit community and the government used the data to develop new transport plans for the city. By sharing the work with a broader set of actors as it was being developed, we helped to generate a local and global community around using GTFS data for semi-formal transit.

2.0 Theory Literature Framing

2.1 Leveraging Mobile Devices to Collect Transit Data

One the biggest issues for studying and modeling transport is acquiring data to accurately represent these systems (Herrara et al 2010). The prevalence of mobile devices that include GPS positioning has produced research on the possibility of using the data generated by these devices to collect critical transport data. Many of these studies have shown cell phones can help to model transit flows by actively collecting GPS data (Caceres et al 2007, Choi and Jang 2000, Herrara et al 2010, Wang et al 2010). Other studies have looked at how the GPS data stored by cell phone providers can be used to model traffic flows, including those in developing countries (Author, 2006, Gonzalez et al. 2008, Caceres et al. 2012, Talbot 2013, Wakefield 2013). Other projects look at how transit riders could crowd source transit vehicle locations in real-time using people’s cell phones (Thiagarajan et al. 2010). Many formal transit agencies globally, are actively collecting GPS data from devices they install on their vehicles (Farzin 2008). However, studies that look at public transportation data collection often focus on formal systems rather than semi-formal buses (Farzin 2008). Acquiring cell phone records from telecommunications companies is one key way to access mobility data but it is often extremely difficult to access (Gonzalez et al. 2008). Experiments, in which cell phone users are actively collecting and contributing data through their mobile devices have become more successful as the data is owned by the collector and can be shared. Our research team wanted to see if we could apply this kind of methodology to semi-formal transit.

2.2 Data Availability: Semi-Formal Bus Systems

When our research team started our project in 2012, we did not know of any organizations using mobile devices to generate data on semi-formal bus systems. However, as our work progressed we discovered a handful of initiatives which sought to develop semi-formal transit data using mobile devices that were working in parallel to ours. A team at the World Bank with support from AusAid, worked with the Philippines Department of Transport and Communications and other transport-related agencies in Manila to set up a Transportation Information System which includes an open database containing basic service information for the myriad of public transport modes in the city (World Bank and Ausaid 2014). The World Bank also supported a project in Mexico City with the Department of Transport (SETRAVI or Secretaría de Transportes y
Vialidad del Distrito Federal) and is conducting similar work in three Chinese cities (World Bank and Ausaid 2014, Eros et al. 2014). The MIT based team (Urban Launchpad) has collected data, although not initially in GTFS format, for the bus system in Dhaka (Ching et al 2012, Zebras et al. 2014). In each case the groups developing the project created various mobile tools to collect routing and stop data.

The informal and flexible nature of paratransit systems make them highly variable and erratic, which presents a serious challenge to data collection (Guillen, et al. 2012). Governments are often reluctant to collect data on these systems as they find them too “chaotic” or too complex to address. Some government and industry actors collude and mutually benefit from the lack of transparency of data in these systems (Cervero and Golub 2007, Kimei 2014, Author 2015, Republic of Kenya 2009). When government agencies do in fact collect data, they often hire consultants that do not always share the data (Author, 2014). This is the case in the Mexico City, which recently collected data on their formal and semi-formal system, but only released data on their formal system to the public (Eros 2014).

Operators of semi-formal transit sometimes collect analog data on their system to help manage their services, but this data is rarely standardized or shared across transit operators or with the public. Many semi-formal transit operators do not see an immediate benefit from providing the data or do not have the means to collect it. The informal and often unsanctioned nature of these operations may lead some owners to keep their activities hidden from government oversight. The operators who do collect data on their systems do so to maximize efficiency and profit (Eros 2014). This means that even if data is collected, it is usually analog, incomplete, unstandardized and private, and therefore unavailable for comprehensive transportation planning or the development of user centered transit information.

While the recent initiatives to collect data on semi-formal bus systems marks a change with the past, few cities in the developing world are currently generating or sharing transit data in a standardized format, such as GTFS. A review of the GTFS Exchange, a web based platform for sharing GTFS transit data started and maintained by Jehiah Czebotar, shows that of around 766 agencies producing feeds only four were in Africa including our own and two were mostly rail feeds not bus data although of course, more feeds may exist that are not on the exchange. ³

2.3 GTFS (General Transit Feed Standard) Open Standard for More Accessible Data

³ The others were Tunisia Society National Des Chemins De Fer (Tunisian railways) Gautrain Management Agency (Pretoria and Joburg, South Africa)- rail and they have tried to add some bus My Citi (Cape Town some bus data from their integrated bus rapid transit system). Accessed August 24, 2014.
The General Transit Feed Specification (GTFS), was first developed in 2005 for Portland's TriMet transit agency in conjunction with Google to provide a way for transit agencies to standardize their data for use with trip routing software, such as Open Trip Planner and eventually Google Maps (McHugh 2013). The standard was implemented in Google Maps in 2006 and adopted by transit agencies across the US who wanted to provide their users with better access to route and schedule information. This simple standardized data format consists of a series of text files collected in a ZIP file. Each file models a particular aspect of transit information, much of which is relational: stops, shapes, routes, trips, stop times, and other schedule data - the guidelines are found on the Google developer site (Google Developers 2014).

By 2007, many formal transit agencies adopted the GTFS standard to share their data, even if their transit data was originally collected in a different standard. They did so because they wanted their transit routing information to be viewed in Google Maps (McHugh 2013, Wong 2013). This process has created a worldwide standard for openly sharing transit data. Transit agencies who develop data with the standard often freely share that information by posting the data on the GTFS Exchange.

The openness and sharing of GTFS data has encouraged its use for transit applications beyond trip planning including tools to improve transit operations and planning overall (Catala 2011, Lee-Gosselin et al. 2012). A Brisbane study used GTFS data along with smart card data to create travel paths of passengers on their transit (Tao 2014). Another study used GTFS data in Auckland (New Zealand), Vancouver (Canada), and Portland (Oregon, USA) to develop a model that would allow public transport agencies to assess and benchmark different services (Hadas 2013). Before the release of GTFS data, this type of analysis and assessment was hard to achieve because of the varying data standards that existed across rail, bus, and subway routes used by many public transit systems (Hadas 2013). Open Trip Planner, a tool originally developed for GTFS routing, has created a plug-in that allows users to determine accessibility of transit. The plug-in has been widely applied including determining accessibility to transit directly after Hurricane Sandy in New York City (Byrd et al. 2012, Wong 2013). Overall, transit planners are realizing that GTFS can be used for applications beyond trip planning and are starting to use this data to model transit in new ways (Cartala 2011).

3.0 Nairobi Context

3.1 Nairobi’s Technology Community

Nairobi, Kenya provides a good case study for the use of mobile phones to collect transit data because of the city’s widespread use of cellphones and its strong technology community. Over the last decade mobile technology use has exploded in developing countries, and Kenya has become a center for some of these developments in Africa (Aker et al. 2010). The number of mobile connections rose from 30.4 million in 2012 to 31.2 million in 2013, and Kenya’s current
mobile phone penetration rate is 74.9%, above the average for Sub-Saharan Africa (Kenya National Bureau of Statistics 2014). The low cost of handsets and texting along with their high utility facilitates the rapid spread of cellphone use. This rapid expansion of mobile use in Kenya is evidenced by the success of the M-Pesa a mobile banking service which nearly 2 year after starting in 2007 has 8.5 million Kenyan user and US $3.7 billion (Equivalent to 10 percent of Kenya’s GDP) had been transferred over the system (Safaricom 2009, Mbiti and Weil 2011).

Nairobi has a thriving technology community and higher cellphone use than the rest of the country. It is home to the iHub, an innovation and technology space developed to encourage and support technology entrepreneurs in Nairobi by creating a shared community of learning (Hershman 2012). Ushahidi, a crisis mapping tool now used worldwide, was developed in Nairobi as a response to the 2007 election crisis, and this innovation around mapping is ongoing. In 2013 IBM launched a research lab in Nairobi in collaboration with the Ministry of Information, Communication and Technology (ICT) through the Kenya ICT Board. The lab is meant to focus on applied research, solving problems “relevant to Africa and contribute to the building of a science and technology base for the continent.” (McLeed 2013). Much interest and experimentation around the use of mobile technology has focused on health, economic development, and humanitarian response. The application of mobile technology to the many problems in transportation appears to be just beginning.

3.2 Nairobi Semi-Formal Transit (Matatu) System

Nairobi’s semi-formal bus (matatu) network comprises over 135 routes, which according to the 2009 census, serves a population of well over 3.1 million within the metropolitan area. Matatus act as the main motorized public transport for the majority of city inhabitants even though they are privately run and operated (Salon and Aligula 2012). In Nairobi, the matatu network developed in reaction to the poor funding and management of the municipal public transport systems and helped fill a gap in service (Mutongi 2006, Author 2015) In contrast to other infrastructure, the vehicles are locally owned and involve large numbers of small businesses, from the operators themselves, the drivers, to the mechanics tasked with repairing the vehicles (Mutongi 2006). Matatus largely run on “official” routes, usually remnants of the former bus network. However, as the city expands and as new roadways are constructed, new routes are developed by the operators, which are not often officially sanctioned. The service does not generally have fixed schedules and fares. Drivers often take detours to avoid traffic or traffic police and sometimes take the liberty of improvising stops. Currently, approximately 9,554 matatus and buses serve the Nairobi region on approximately 135 routes (Transport Licensing Board 2012).

4.0 Data Collection Team and Methodology
Over the course of 2012-2013, data on Nairobi’s matatu system was successfully generated by our research collective which consists of an inter-disciplinary team from three universities (two US and one Kenyan) and one US design firm. Our Kenyan partners at the University of Nairobi led the data collection process with a team of five students who performed most of the field work. Student road on the buses and collected route name, stop name, undesignated and designated stop information. In areas where the matatus were too dangerous for the student to ride, students followed the matatus in cars. Data collection occurred over the course of one year, however the process was periodically stopped to test data quality and retool our collection software. The final collection took roughly 6 months during Nairobi’s summer which is relatively dry. Routes often needed to be surveyed multiple times to ensure that we obtained the most consistent route. Routes can change because of construction, avoidance of police, school opening and closing. Once collected, the data was validated using the Google GTFS validator.

The data collection process involved identifying existing routes, developing and testing mobile GPS enabled tools to collect the data, creating a unique coding structure to allow the data to be formatted in GTFS, generating a methodology for data collection in the field, translating the data into GTFS, interfacing with Nairobi’s transit community, and ultimately releasing the data by posting it on GTFS Exchange website and holding a public launch. This is the first time that these routes have been mapped in a comprehensive manner, as well as the first time paratransit data in Nairobi has been fully integrated into the GTFS standard. Details of the development process are below. Figure 1 provides a flow-chart of the overall research methods, data collection and development process.

**FIGURE 1 GOES HERE**

Figure 1 : Flow-chart showing research strategy and process. One can see that engaging the community with the data was just as important as developing the data itself. This process helped the transit community trust the data and therefore use it.

4.1 Identifying the Current Routes

The first step in this work involved finding and collecting existing data on routes. We obtained government data in the form of word files and found it to be incomplete, outdated, and often inaccurate. Route changes are often developed by the matatu industry, not the government, in response to new demand and possibilities opened up by new road construction. These changes are often not recorded in the government files. It should be noted that the Kenyan National Transport and Safety Board recently started moving towards publishing official and unofficial changes to matatus routes as well as new matatu licences in the *Kenya Gazette*, the official government publication. This may help in updating the data moving forward.
The research team discovered a paper based map created in 2010 by Kenya Buzz, a Nairobi based media company, which created the map for commuters. The map had a small print release and was not obtainable through purchase or download at the time of the study. The data used to develop the map was never released and was not digital. “Living in Nairobi” published a highly stylized route map in 2012 after we had started our work, but they also did not publish any of the data collected to create the map, and have not maintained the map. Panga Safari, formally Matatus Online, developed a private matatu route database that covered parts of the city, but did not include standardized routing information or consistent stop documentation, making it difficult to upgrade this data to a standardized format such as GTFS. The database has been expanded, and can be searched through a web interface, but the backend data was never made public. We also identified and reviewed existing data collection projects performed by entrepreneurs for business or social reasons. Many were incomplete, included major errors, or employed inconsistent methodologies and data structures that made the data impossible to combine or format into GTFS. In addition, only a few were willing to share their data. In order to develop our own consistent and standardized data format, it was clear that we would need to create it from scratch.

4.2 Tools Used to Collect Data and Method of Collection

The team began by testing various Android smartphone-based data collection tools, eventually focusing on using MyTracks, a basic GPS tracking system for mobile devices developed by Google. Much later in the collection cycle, we tested TransitWand, which was released near the completion of our data collection process. GPS units were used as back-up so that we could compare the accuracy of the multiple forms of data collection (See Figure 2 & 3). Through testing we found that standard GPS units and the mobile applications on Android phones had similar accuracy. However, mobile phones sometime took longer to lock in on GPS satellite signal and could lose connection more regularly. The Mytracks app allowed for easier digital collection of meta-data (for example, the name of a stop and current passenger counts could easily be recorded). Collection with GPS units needed a paper recording that cross referenced waypoint numbers, which was then digitized and joined to the GPS data later on.

The biggest challenges in using the cell phone based data collection applications included extremely limited battery life, slow speeds of affordable android phones, the risk of losing a high value android phones to theft in a matatu where security is a constant problem (which did happen unfortunately), and the small screen size that made digital data entry more time consuming—

---

6 Last Accessed 10/9/2014 (http://www.matatuonline.com/)
particularly in situations where frequent stops were made. Ultimately, we found that mobile phones were effective tools for data collection, although new phone applications could be developed to streamline information processing while in the field and for the eventual conversion to GTFS.

As we were engaged in this data collection process, we discovered that an open source web and mobile app TransitWand was released by the consulting company Conveyal for a similar GTFS data collection project in Mexico City, involving the World Bank and the Department of Transport for Mexico City. We tested TransitWand in Nairobi and found that this tool, which is built for transit data collection, resolved some of the drawbacks of earlier apps. When compared to data produced by MyTracks and GPS units, TransitWand generated cleaner data because the software automatically snapped location data to roads. However, because the tool was still in Beta development, the ability to directly export to GTFS was not fully operational. This made post-production work of TransitWand data more time consuming and cumbersome than the other applications used. While TransitWand will be very useful if it is developed further, the team decided to complete the data set using MyTracks.

**FIGURE 2 and FIGURE 3 – GO HERE**

![Figure 2: Left: Image of student collecting data using and android cell phone and a GPS unit as back-up. Image Credit Adam White. Figure 3 Right: Image of data mapped in Open Street Maps an open source routing application.](image-url)

While data was largely collected onboard matatus, on particularly dangerous routes as well as the busier, more complicated routes, to ensure security data was also collected in private vehicles that followed matatus. Safety became an important issue to address with our methodology. Theft and crime is common on matatu routes, with some traversing particularly dangerous sections of the city. However, travel onboard the matatu vehicles themselves allows data collectors the opportunity to engage with drivers and passengers. It is also more affordable and scalable than the alternative of employing a tracking car. On the other hand, the use of a private car allowed the data collector to observe multiple vehicles and to identify stops based on patterns of vehicles. The control of the vehicle also allowed extra time to take notes, capture more metadata and make sure that a stop or notation was not missed.

After testing several different tools and processes used for the data collection, we devised a standard protocol and methodology for creating route, stop and shape data to fit the GTFS coding structure (See Appendix A). In all cases, data collectors would ride a route, either in the car or matatu, use the data collection tool to generate latitude and longitude points along the route, and record all of the stops as well as specific coding information we developed for each route, stop and shape which was essential for the GTFS protocol. (See Appendix A).
While many paratransit systems involve some stopping at varied locations depending on customer demand, regular and central stops as well as large terminals exist. Students identified stops based on their personal knowledge, information from frequent users of these routes, visual notation (signs, shelters etc.), and if necessary, confirmation from discussion with matatu crews or group of commuters on the route. In many cases stops were identified as either designated - meaning that a government agency had established a location for stops- or undesigned-those locations which the matatu operators established based user demand.\(^7\) (See Appendix A) Adding this additional data to the GTFS file could be a useful tool for the city should it move to formalize many of the more heavily used undesignated stops. (See Figure 4)

**FIGURE 4 GOES HERE**

Figure 4: Image of the designated and undesignated stops along one matatu route in Nairobi. D - marks stops designated by the city. U - marks undesignated stops. There are many more undesignated stops which in some cases contribute to the traffic congestion problem in Nairobi, although the problem is also of road design that does not account for the needs of matatus.

### 5.0 GTFS Formatting for Semi-Formal Transit:

Once the essential data on the routes, stops, and shape of the routes was collected, we started the work of translating the data into the GTFS standard. The GTFS data format assumes that the system is part of a formal transit agency, but Nairobi’s matatu systems is managed and operated very differently from one organised by a formal transit authority. GTFS data assumes that the transit agency has developed a unique identification system for routes and stops. We needed to develop this unique identification system from scratch (See Appendix A). The specification also assumes there are standard schedule and fares, standard vehicle types, scheduled services outages, and that transit agencies are maintaining the data overall. Since many of these assumptions do not exist for paratransit, we needed to make modifications to fit our data to the GTFS format.

The GTFS standard uses trips, their stop times, and frequencies as the framework for the relational database that makes up the standard. Given that matatus have loosely set schedules we had to generate rough estimates for departure and frequency of trips from the main terminus at peak and off-peak periods as well as the stop times (a matatus generally leaves a stop every two minutes). While this does not approach true accuracy, it serves as a reasonable estimation.

\(^7\) We used three criteria to identify designated stops: 1) physical infrastructure (pullout from the road, bus shed or bus stop, a sign that the stop is “matatu and bus crew organized”) 2) evidence of approval from Nairobi City Council (now Nairobi City County) or 3) evidence of approval by being noted in official road maps. However, as the city government has not been actively planning and designating official stops, the majority of stops remain informal and undesignated. We therefore collected both the designated and undesignated matatu stops and coded them in the stop ID data file. (See Appendix A)
Matatus do not have standard fares as the fares are largely demand driven. For instance, when it rains in Nairobi, fares can triple. There are also cases of predatory fares that are lower than usual to lock out competitors. Note that besides this variability, fares are unexpectedly expensive, with poorer residents often unable to afford them (Salon and Gulyani 2010). Most casual laborers in Nairobi earn about Kshs 1000 per week which translates to about 100 to 150 shs per day or about 1 dollar per day while fares average 100 shs per day for return journeys hence low income earners choose to walk unless there is an emergency. Fare information was optional in the GTFS format; therefore we decided not to populate this file as it would be impossible to develop it in a standardized way.

GTFS requires an “Agency” file, usually a transit agency. Given that the data was developed for the hundreds of “agencies” that operate matatus, the research team was initially listed as the agency producing the data. The matatu system is fragmented and complex, therefore a neutral and technically capable institution collecting the data can ensure quality and uniformity. Ideally, this function should eventually be taken over by a government agency such as Kenya Institute for Public Policy Analysis (KIPPRA) that has a steady budget allocation for updating the data along with a strong mandate to make it openly available. KIPPRA has expressed interest in maintaining the data and the methodology which will be refined in a next phase focusing on streamlined and user friendly systems and tools for updating. More recently, the government has made moves to create a Nairobi Metropolitan Area Transit Authority which will have clear responsibilities around transit including data gathering. The data, methodologies and tools developed through this work, along with the expertise KIPPRA has gained through our collaboration, will be helpful in kick-starting the data and transit planning work of this new agency.

5.1 Changing the GTFS Standard for Semi-Formal Transit:

As the previous discussion shows semi-formal transit systems operate differently from traditional buses and the research team wanted a way to indicate that difference in the GTFS data format. Changing this specification in GTFS is particularly important for hybrid transit systems that include both formal and semi-formal systems, because it would allow for more accurate transfer and routing between the two systems and would also allow planners to analyze the dynamics between the two transit types.

Our team sought to actively address the changes needed to GTFS for use with semi-formal transit. With support from the Rockefeller Foundation and the World Bank Open Transport Initiative we convened a conference that brought together participants involved in developing the GTFS standards itself, research teams focusing on developing GTFS for semi-formal transit (This included members of the team in Mexico City, Manila, and Dhaka), and members of the paratransit community in the United States who are struggling with similar issues with using the
GTFS standard. The GTFS standard is particularly interesting in that it has never been formalized by any agency or multi-lateral body. It has become a de facto standard though adoption by growing number of users globally who want their data to appear on Google maps, which uses the standard. Our team hopes that modifications to GTFS for paratransit would encourage more creation of GTFS data for paratransit and similarly be adopted through widespread use.

Conference participants proposed and approved a change to the GTFS format. The group added a “continuous_stops” field to the stop_times and routes table indicating that the route and its stops do not follow normal bus transit behavior, but rather it is possible to board or debark from a transit vehicle at any point along the vehicle’s path of travel. The field can have the following non-negative integer values; 0 or blank - Normal stop behavior along entire route (default). 1 - continuous stopping behavior along entire route. If 1 is specified, a valid shape file must be identified for the route, in order to indicate the complete path of travel for each trip. This makes shape files, which are optional for formal agencies, more essential for the development of GTFS for semi-formal transit.

6.0 Working with the Transport Community in Nairobi

As we proceeded with the process of data collection, we also partnered with the Kenya Institute for Public Policy Analysis (KIPPRA), Kenya’s primary government think tank tasked with transport data analytics and modeling. The research team held two workshops where we invited technologists, various government transportation offices, policy analysts and transport operators to discuss the project. The workshops were held to obtain early feedback from potential users and to inform members of the transit and technology community about the data collection process so that could trust the data we collected. In the process, the team assisted Laban Okune, who used the data to improve his award-winning Ma3Route mobile app which shares real-time, crowd-sourced matatu and traffic data between users. We also worked with Jeremy Gordon of Flashcast who developed a routing program called Sonar using the data and who also shared data with us. We facilitated the use of the data by UN-Habitat/Institute for Transportation and Development Policy (ITDP) who found the data useful as they began a Bus Rapid Transit Service Plan for the city. They also shared further data with us, creating more and more circulation of valuable data. Overall, the workshops helped to inform potential users about the data and its format and gain feedback. The transparency in which we discussed the development of the data at these workshops helped to create users of the data and an ethos of sharing (Author 2014).

The research team also developed a workshop with the matatu drivers and operators organization to provide feedback on the maps we developed from the data. The matatu drivers provided

---

8 We acknowledge the role of Holly Krambeck who leads an Open Transit Initiative at the World Bank in convening and facilitating this conference.
helpful information about routes and stop names that were missing from our data-set. They were also excited by the development of the paper map (See Figure 5) which allowed them for the first time to see the extent of the transport system. After seeing the map the matatus drivers began a conversation about the development of new routes in underserved areas. A similar conversation with officials from the National Transport and Safety Authority (NTSA) helped them recognize that many of the non-official routes make sense. The map helped generate an overdue conversation on transit routing and planning with Nairobi’s transport community.

**FIGURE 5 INSERT HERE**

Figure 5: Matatu map used during our focus group with matatu drivers and owners. Nairobi’s city government made it the official matatu map when the final edited version was released in January 2013.

In order to further disseminate information about the data the research team hosted a transit hackathon. Over the weekend of 25-26 January, 2014 eight teams of up to four university students each participated in the collaborative programming competition. The teams came up with a number of interesting ideas using the data, from route planning apps to apps designed to integrate with the local Beba Pay system (an automatic card payment system that subsequently failed). One application functioned as the backend for a group ridesharing program, effectively enabling partygoers to “crowdsource” a matatu ride home late at night. Another would alert drivers of “blackspots” or notoriously crash prone areas. The winner of the hackathon was a young programmer who devised Ktransit, a program which created an API to access the GTFS data more easily by translating the series of comma-delimited data into information that can be captured and used by other programs.

### 7.0 Discussion and Conclusion:

In many cities with paratransit, basic transport data often does not exist. This project demonstrated that with a dedicated team from the university using mobile technology, it is possible to create valuable data in such cities. Further, it showed how to transform that data into a GTFS format. The research team also found that the GTFS format is a very useful framework for paratransit data collection, because it can be used with planning software developed for the format. We also discovered, however, that specific changes are needed to the GTFS standard to accommodate the nature of paratransit. Through our work we identified growing numbers of teams working in different parts of the world that are developing GTFS data for semi-formal transit systems. Together this work can help spread the use of the data and facilitate cross city comparative studies in how these systems function. This collaborative work will provide needed data to cities that may otherwise be left behind in the growing technology revolution in transportation (Townsend 2014).
Our tests of existing technology for mobile geographic data collection, including My Tracks and Transit Wand show that these tools are adaptable for GTFS data collection and with some modifications these tools could better facilitate the collection of GTFS data for paratranist systems. Data collectors found it cumbersome to enter in the metadata necessary for the GTFS format while in the field. Future research should address changes to the tools to assist with the data collection process. Data storage and export from the tools made it difficult to translate the raw data into the formatted text files GTFS needed. Future work should also look at the ability to develop data collection tools that automate GTFS formatting. Transit Wand developers hoped to do that, but as a consultant company they would need a project that specifically tasked them with that work. This points to the need to find a mechanism for more investment in some of these tools and open data and innovation they enable. Future research should also look at the possibility of developing crowd sourced data collection tools which have largely been applied to more formal transit systems (Thiagarajan et al. 2010). Creating new tools that facilitate data collection processes in the field and the ability to generate GTFS data on the fly would help the development of data on semi-formal transit systems.

While the GTFS format may not be perfect for paratransit, trying to fit the data into the standard to ensure routing application could be built with the data. We found that the GTFS format does allow the inclusion of additional data points not part of its core, which can be helpful for future modeling and planning of the system. We used this feature to develop additional information on designated and undesignated stops, other data such as ridership stats, passenger pick-up, safety of vehicle could also be collected and would help with transit planning. More importantly, the standardized nature of this data has created the possibility of using plug-ins and programs developed for GTFS (Byrd et al. 2012, Hadas 2013, Wong 2013) to measure transit accessibility and transit flows among other planning applications.

One of the more unlikely findings from the study was our team’s ability to locate other researchers who were simultaneously working on similar projects to develop data on semi-formal transit in the GTFS format. Our research helped to bring this group together, through a conference we convened with the World Bank. Continued development and expansion of this community and sharing of insights, data and tools could help support a new paratransit inclusive GTFS format and encourage the development of transit planning tools for semi-formal transit that use the format. Several changes to the format have already been developed. Future work should address issues that GTFS has with handling scheduling and fares.

The project illustrates that there is a strong demand for semi-formal transit data in rapidly developing cities across the globe where mobile technology is increasingly prevalent. This is evidenced by the fact that openly sharing the data allowed for the development of five mobile phones based matatu routing applications, Ma3Route, Sonar Flashcast, Matatu Maps, Digital Matatus and Transit App which are now in use in Nairobi. However, it is not only the technology community that benefited from the GTFS data. NGOs, such as Institute for Transportation and
Development Policy (ITDP), multilaterals such as UN-Habitat and the World Bank have used the data in their project work for Nairobi and UN-Habitat more recently has tried to replicate this work in Kampala, Uganda. The government is increasingly seeing the benefit to the development of the data. The Nairobi city county government made the data and map developed from it the official transit map for the city (See Figure 5). The ability of the government to accept the data was in large part because of the process of including them in workshops about the data collection and openly sharing that data with the government. By engaging Nairobi’s transit community during the data development we created trust in the accuracy of the data, demand for its use, and as stronger data sharing ethos (Author 2014). Leveraging technologies such as cellphones that are ubiquitous in rapidly urbanizing countries to create data and linking this data to open data architecture, such as GTFS, has the potential to fundamentally transform what has often been a closed, data deficient transport planning process.

References


XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX


Author (2006) XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX


Authors (2014)
Xxxxxxxxxxxxxxxxxxx

Wong, James. "Leveraging the General Transit Feed Specification for Efficient Transit Analysis."


APPENDIX A: ROUTE CODING STRUCTURE DEVELOPED FOR GTFS

Nairobi’s transit routes largely fall along the major road corridors. The team gave each corridor a numeric identifier and used that as the basis for the unique identification system developed for the GTFS data (See Figure #1) A different alpha-numeric identification code was then developed for the routes, stops, schedules and shape files that are part of the GTFS data structure. The codes included metadata about each data point collected, to help maintain knowledge gained about the system during the data collection process. The coding structure methodology was developed in a way that would easily allow new stops and routes to be added to the data over time. The development of the identification system is detailed below.

Appendix Figure 1:
The above image shows how we broke down Nairobi’s matatu system into a series of corridors.
Appendix Figure 2: Here you can see Nairobi’s main corridors.

Routes Coding Structure:

In the GTFS file structure the routes file includes a unique identifier for the route. It also includes route short and long name as well as a description of the type of route (bus, rail, subway, etc). Each matatu route falls along a major corridor. The first digit in the route unique identifier is the number representing that corridor (see Figure #1). Then as the matatus route branches off that corridor, it is given a series of numeric identifiers representing its place along the branch. Therefore, the second two digits in the identifier represent the second level branch. The next four digits represent the alpha-numeric characters developed for unique route numbers. The next digit represents whether it was a designated (1) or non-designated (0) route. The final digit represented within it was an inbound (1) or outbound (0) route. See Figure #3 for how the route branching structure works.

Figure #3 : Route coding : The figure below illustrates the route coding based on our protocol.
Route coding: corridor | 1st level branch | 2nd level branch | route no. | gazetted | direction
E.g. route code for route 48 is: 8 | 01 | 01 | 0048 | 1 | 1
Operates from Odeon terminus (in CBD) through Riverside Drive to Kileleshwa (along Waiyaki Way - Corridor 8)

**Stops Coding Structure:**

The GTFS data structure for the stops includes a stop unique identifier, the stop name, latitude and longitude information for each stop, along with the stop type and a determination of whether it had a parent location.

The first digit in the stop unique identifier represented the name of the main corridor. When the stop was designated (1) or undesignated (0). The next digit represented within it was an inbound (1) or outbound (0) route. The next 3 digits were character abbreviations of the stops. For example “WST” for Westlands. (See Figure #4)

Stops coding simplified:
Bus Stop coding: corridor | designation | direction | stop name
Example of a Bus Stop code: 08|1|1|AAA

**Shape Coding Structure:**

In the GTFS data structure the shape file recreates the path of the route. It includes a numeric identifier for the route, and a series of latitude and longitude points and a sequence numbers so the routes can be drawn in various software packages.

The first digit of the shape unique code is the corridor number. The next four digits represents the alpha-numeric characters for the route or the route short name. The next digt represents the origin or what we called Level. 1 - for matatus originating from the main terminus, 2-9 - for matatus not originating from the main terminus. The next number represents the route variation. Many routes vary slightly at the end, and this would indicate that variation. Examples 2 and 3 below show different variations on the same route.

(c) The shape file coding
Shapes coding is made up of:
corridor | route no. | route level | route variation no.

Using two examples to illustrate this coding.
Example 1 – Karen route 5|0024|1|1 - originates at Ambassador to Karen through Langata road
Example 2 – Karen route 5|0024|2|1 - originates at Bomas (Galleria) to Karen through Karen “C”
Example 3 - Karen route 5|0024|2|2 would represent a route originating at Bomas (Galleria) to Karen through Hardy shopping centre (a different variation on the original route)
Figure 1

Data Development

- Identify Existing Data
- Test Tools for Data Collection

- Develop Data Collection Methodology (Including development of Unique Coding Structure)
  - See Appendix A
  - Collect Data
  - Format in GTFS

- Develop Modification to GTFS for Semi-formal Transit (Approved by GTFS community)
- Informal Release of Data
- Development of Paper Map
- Edit Data and Paper Map
- Release Data and Paper Map (January 2014)

Community Engagement

- Workshops with Transit Community
- Workshops with Transit Community

- Conference to Change GTFS data Structure
- Work with Technology Firms Developing Apps with Data (Sonar and Ma3Route)
- Focus Groups/ Data Review: Matatu Operators and Drivers Larger Transport Community Students at Local Universities
- Data Hackathon University and Tech Community

Official Matatu Map City of Nairobi NGOs Use if for Transport Analysis Matatus Operator Establish new Routes Transport Routing Apps Released API Developed With the Data
Figure 5

NAIROBI MATATU ROUTES

From City Center

Outbound

Inbound To City Center

Stop

Terminus

City Center

Junction

Route

A line

B line

C line

D line

E line

F line

G line

H line

I line

J line

K line

L line

M line

N line

O line

P line

Q line

R line

S line

T line

U line

V line

W line

X line

Y line

Z line

**33 lines also to Fedha, Imara Daima, Highway, South

Data Collected by:

Research funded by the Rockefeller Foundation.

Urban Development, Columbia University; School of Computing and Informatics, University of Nairobi; Groupshot.
Nairobi Matatu Stops (As KMZ file)
Click here to download KML File (for GoogleMaps): stops_red.kmz
Click here to download Video: video1_new.mp4
Video of Generation of Map Part 2
Click here to download Video: Video3_new.mp4
Develop mobile application to collect data on Nairobi's semi-formal (Matatus) transit system.

Data collected using GTFS data standard.

Made changes to standard to accommodate semi-formal transit needs.

Data translated to stylized transit map, which became official transit map for the City of Nairobi.

Openly Shared GTFS Data was used by local technology community to develop mobile applications for routing in Nairobi.