SJSU SAN JOSÉ STATE UNIVERSITY

MTI MINETA TRANSPORTATION INSTITUTE

Establishing a Central Archive for Transit Passenger Data

Gregory Newmark, PhD
Hilary Nixon, PhD

MINETA TRANSPORTATION INSTITUTE

LEAD UNIVERSITY OF

Mineta Consortium for Transportation Mobility

Founded in 1991, the Mineta Transportation Institute (MTI), an organized research and training unit in partnership with the Lucas College and Graduate School of Business at San José State University (SJSU), increases mobility for all by improving the safety, efficiency, accessibility, and convenience of our nation's transportation system. Through research, education, workforce development, and technology transfer, we help create a connected world. MTI leads the four-university Mineta Consortium for Transportation Mobility, a Tier 1 University Transportation Center funded by the U.S. Department of Transportation's Office of the Assistant Secretary for Research and Technology (OST-R), the California Department of Transportation (Caltrans), and by private grants and donations.

MTI's transportation policy work is centered on three primary responsibilities:

Research

MTI works to provide policy-oriented research for all levels of government and the private sector to foster the development of optimum surface transportation systems. Research areas include: bicycle and pedestrian issues; financing public and private sector transportation improvements; intermodal connectivity and integration; safety and security of transportation systems; sustainability of transportation systems; transportation / land use / environment; and transportation planning and policy development. Certified Research Associates conduct the research. Certification requires an advanced degree, generally a PhD, a record of academic publications, and professional references. Research projects culminate in a peer-reviewed publication, available on TransWeb, the MTI website (http://transweb.sjsu.edu).

Education

The Institute supports education programs for students seeking a career in the development and operation of surface transportation systems. MTI, through San José State University, offers an AACSB-accredited Master of Science in Transportation Management and graduate certificates in Transportation Management, Transportation Security, and High-Speed Rail Management that serve to prepare the nation's transportation managers for the 21st century. With the active assistance of the California Department of Transportation (Caltrans), MTI delivers its classes over a state-of-the-art videoconference network throughout the state of California and via webcasting beyond, allowing working transportation professionals to pursue an advanced degree regardless of their location. To meet the needs of employers seeking a diverse workforce, MTI's education program promotes enrollment to under-represented groups.

Information and Technology Transfer

MTI utilizes a diverse array of dissemination methods and media to ensure research results reach those responsible for managing change. These methods include publication, seminars, workshops, websites, social media, webinars, and other technology transfer mechanisms. Additionally, MTI promotes the availability of completed research to professional organizations and journals and works to integrate the research findings into the graduate education program. MTI's extensive collection of transportation- related publications is integrated into San José State University's world-class Martin Luther King, Jr. Library.

REPORT WP 18-04

ESTABLISHING A CENTRAL ARCHIVE FOR TRANSIT PASSENGER DATA

Gregory Newmark, PhD
Hilary Nixon, PhD

May 2018

A publication of
Mineta Transportation Institute
Created by Congress in 1991

College of Business
San José State University
San José, CA 95192-0219

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. WP 18-04	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Establishing a Central Archive for Transit Passenger Data		**5. Report Date** May 2018 **6. Performing Organization Code**
7. Authors Gregory Newmark, PhD and Hilary Nixon, PhD		**8. Performing Organization Report** CA-MTI-1631
9. Performing Organization Name and Address Mineta Transportation Institute College of Business San José State University San José, CA 95192-0219		**10. Work Unit No.** **11. Contract or Grant No.** 69A3551747127
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology University Transportation Centers Program 1200 New Jersey Avenue, SE Washington, DC 20590		**13. Type of Report and Period Covered** Final Report **14. Sponsoring Agency Code**
15. Supplemental Notes		

16. Abstract

This report describes the rationale, background, establishing organization, and future steps of CATPAD, the Central Archive for Transit Passenger Data.

The Central Archive for Transit Passenger Data is a repository that collects, indexes, archives, and makes available online the transit survey instruments, data, and reports collected across the country. This resource is unique in its focus on the disaggregated information of individual transit users – information that is critical for a range of transportation planning analyses. In addition, where available, CATPAD contains aggregated information, such as station boardings and service and fare schedules, to provide key context for the disaggregate person-level data.

The Central Archive for Transit Passenger Data seeks to overcome the current impediments to accessing transit survey data by providing a single, searchable, internet archive to store and disseminate this valuable information. The Central Archive for Transit Passenger Data explicitly aims to expand the public return on the considerable investment made to gather transit passenger data. The resource is designed from the start to serve the needs of a range of use cases from transportation planners and policy makers to researchers and community advocates.

The goal of CATPAD is to make useful data available to inform transit decision making at all levels and to foster ongoing refinement of the nation's transit network.

17. Key Words Transit survey; open data; passenger information; repository	**18. Distribution Statement** No restrictions. This document is available to the public through The National Technical Information Service, Springfield, VA 22161		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 28	**22. Price** $15.00

Form DOT F 1700.7 (8-72)

Library of Congress Catalog Card Number:
2018930075

To order this publication, please contact:

Mineta Transportation Institute
College of Business
San José State University
San José, CA 95192-0219

Tel: (408) 924-7560
Fax: (408) 924-7565
Email: mineta-institute@sjsu.edu

transweb.sjsu.edu

052418

ACKNOWLEDGMENTS

The authors thank Jason Clark for programming the CATPAD database and the members of the Transportation Research Board (TRB) Travel Survey Committee, particularly the Transit Survey Subcommittee, for allowing us to share this idea at the TRB Annual Meeting and providing feedback.

The authors thank MTI staff, including Executive Director Karen Philbrick, PhD; Publication Support Manager Joseph Mercado; Executive Administrative Assistant Jill Carter; and Editor Jan Steckel.

TABLE OF CONTENTS

LIST OF FIGURES

EXECUTIVE SUMMARY

This report describes the rationale, background, establishing organization, and future steps of CATPAD, the Central Archive for Transit Passenger Data.

The Central Archive for Transit Passenger Data is a repository that collects, indexes, archives, and makes available online the transit survey instruments, data, and reports collected across the country. This resource is unique in its focus on the disaggregated information of individual transit users – information that is critical for a range of transportation planning analyses. In addition, where available, CATPAD contains aggregated information, such as station boardings and service and fare schedules, to provide key context for the disaggregate person-level data.

The Central Archive for Transit Passenger Data seeks to overcome the current impediments to accessing transit survey data by providing a single, searchable, internet archive to store and disseminate this valuable information. The Central Archive for Transit Passenger Data explicitly aims to expand the public return on the considerable investment made to gather transit passenger data. The resource is designed from the start to serve the needs of a range of use cases from transportation planners and policy makers to researchers and community advocates.

The goal of CATPAD is to make useful data available to inform transit decision making at all levels and to foster ongoing refinement of the nation's transit network.

I. INTRODUCTION

This report presents the development of the Central Archive for Transit Passenger Data (CATPAD). The Central Archive for Transit Passenger Data is a repository that collects, indexes, archives, and makes available online the transit survey instruments, data, and reports collected across the country. This resource is unique in its focus on the disaggregated information of individual transit users – information that is critical for a range of transportation planning analyses.

Transit agencies spend millions of public dollars every year surveying their customers on travel behaviors, demographics, service satisfaction, consumer preferences, etc. These person-level data are essential for evaluating existing services and designing future ones. This practice is so useful that, historically, state and city departments of transportation, Metropolitan Planning Organizations (MPOs), and transit agencies voluntarily invested the time and financial resources to survey their customers. In 2012, the Federal Transit Administration (FTA) began requiring this practice for larger agencies as part of the guidance to ensure compliance with Title VI of the Civil Rights Act, further increasing the demand for disaggregate transit passenger data.

Unfortunately, the considerable public resources spent gathering disaggregated transit passenger data are not yielding their full benefit to society, because the resulting data are rarely made easily available. Without this availability, it is very difficult for researchers, policymakers, community advocates, and the public-at-large to quantify the value of transit, to learn from other systems' experience, to evaluate the equity of service changes, or to identify areas for improvement. The Central Archive for Transit Passenger Data addresses this data-availability problem by creating a single, structured repository to both archive and disseminate this information.

This report describes the motivation for this project, the historical impediments to data sharing, and the specific solutions used in the creation of CATPAD. Finally, this report provides a basic description of the CATPAD database, its organization, and anticipated use cases.

II. BACKGROUND

A perennial problem faced by transportation planners, policymakers, researchers, and community advocates has been accessing the disaggregate passenger survey data collected on the nation's transit systems. While aggregated passenger data such as system-wide passenger boardings and total fare revenues collected are compiled nationally, the detailed individual responses to transit rider surveys typically remain hidden behind the closed doors of an agency.

These data are extremely useful but very difficult to access. A specific motivation for this work is to overcome the challenges experienced by the lead author who, while working for the main regional transit funding and strategic planning agency in a multi-operator environment, found it quite challenging to compile the existing survey data sets from the different operators to explore interagency questions such as transfer patterns, fare product usage, and access modes. This interest in reducing the friction of using passenger transit data, combined with the interest of the Mineta Transportation Institute in creating internet resources to spark research, led to this seed project which provided $5,000 to develop CATPAD, the Central Archive for Transit Passenger Data.

TRANSIT SURVEY DATA

This study focuses on the responses of individual transit users to on- and off-board surveys. The innovation of this project is to preserve, organize, and share this passenger-level information necessary for a range of detailed analyses. The ultimate goal is to collect the raw data tables chronicling the individual responses of users to the survey instrument. These represent the core *transit passenger data* of the Central Archive of Transit Passenger Data. These data points, however, require context. Therefore, at a minimum, this study also seeks to collect the associated survey instrument and survey report, which typically describes the survey background, methodology, and higher-level findings. (Often the instrument can be found within the report or its appendices.) In combination, the rider response data, the survey instrument, and the associated report represent the baseline collection of CATPAD.

To augment and contextualize this baseline information, this study also seeks to collect a specific set of ancillary information that describes the nature of the system and its use at the time of the survey – even though these data are not disaggregated to individual riders. These secondary data include boarding and alighting counts (essential for appropriately expanding individual survey responses to reflect the full user base) and fare structure and service levels (essential for understanding the transit environment facing the survey respondents).

The primary transit passenger data sets (with support from the ancillary sets) provide a wealth of information useful for innumerable service, policy, and modeling analyses – from assessing future infrastructure alternatives to tinkering with route alignments to ensuring environmental justice and many other purposes. Transit survey data are often the only source of disaggregated information on rider socio-demographic characteristics (income, age, ethnicity, race, gender, etc.); behaviors (trip purposes, fare products used,

origin/destinations, etc.); customer satisfaction; and stated preferences. Furthermore, because transit surveys are typically conducted at regular intervals on a single property with relatively similar instruments, they offer a unique opportunity for longitudinal analysis.

Despite these myriad benefits, transportation planners, researchers, and advocates struggle to access passenger survey data from transit systems. This difficulty impedes effective policy analysis.

IMPEDIMENTS TO ACCESSING TRANSIT SURVEY DATA

There are several impediments to accessing transit survey data.

A fundamental barrier is identifying appropriate surveys in the first place. This problem occurs nationwide as well as within a single transit property. Nationwide, there are no centralized lists of transit survey efforts. It is possible to find some information by talking with the few consulting firms who conduct a large portion of the major survey efforts. These consultants are likely to also be aware of the work of their competitors. However, these consultants are unlikely to be aware of smaller in-house surveys which were never bid out, surveys at smaller agencies, or surveys prior to the years in which they were active. The lack of knowledge of what surveys have been done hampers attempts to do comparative work among systems, particularly in smaller towns or rural areas.

Similar impediments occur within a single transit property. Agencies typically do not maintain a comprehensive list of their survey efforts. At best, this knowledge is tied up in the institutional memory of longtime agency employees, but as promotions and retirements accrete over time, the ready recall of several decades of surveys deteriorates. Furthermore, even longtime employees with the portfolio for surveying have trouble listing all the surveys they have worked on. At large agencies this can be due to the high number and variety of projects. At smaller agencies, without dedicated market research personnel, this can be due to the many other competing responsibilities. More commonly, interested parties will need to peruse the agency library (often nothing more than a report storage room) or the libraries of the associated state Departments of Transportation (DOTs) or MPOs to yield a partial list. Transit agencies will often post the report from their most recent survey on the agency website, which may yield references to earlier efforts as well. None of these approaches substitutes for a comprehensive, historical listing of all the agency survey efforts.

A second major impediment is that once a survey of interest has been identified, there are typically no organized archives to access the underlying data. While transportation agencies were among the first to embrace computing technology, particularly for travel demand modeling, the development of structured data storage protocols has been uneven. It can be very difficult to find the transit survey data files, many of which were initially stored on antiquated media, such as floppy disks. This challenge can be compounded if the data were collected by a pre-existing agency, such as a former private operator or a public agency absorbed through a merger.

Conducting a hunt for such data is almost impossible from outside the surveying agency. While it is easier within the agency, it can be difficult for agency personnel, typically focused on the challenges of keeping the system running, to make time to look for data on behalf of an external petitioner. Such requests are understandably not seen as agency priorities. Outside parties are unlikely to be able to get senior management to make the survey data requests priorities unless the petitioner can bring political pressure, such as through effective community organizing or political clout. An extreme option to expedite a search is to make a Freedom of Information Act (FOIA) request, but this approach can be perceived by the agency as unduly antagonistic.

Finally, even when the survey has been identified and the data located, transit agency staff are often reluctant to share this information. As noted above, a large part of this reluctance is due to the time it takes to prepare the files – time that often needs to be spent on more urgent tasks tied to system operation. A second concern is that sharing transit survey data may reveal Personally Identifiable Information (PII) from transit customers to the public – an outcome to be avoided. Other theories as to agency resistance to sharing passenger transit data include beliefs that information is power and needs to be held close, that providing data leaves the agency (or the person who managed the project) open to criticism, that the survey data will be misinterpreted by outside users and create unnecessary trouble for the agency, etc. Certainly, agency culture tempers the willingness to find and share transit survey data. The primary author experienced this variation firsthand while working for Chicago's Regional Transportation Authority (RTA), the funding and oversight agency for the three transit providers, known locally as "service boards." Each of the three service boards responded very differently to essentially identical survey requests, even though the RTA technically funded the surveys themselves. Nationwide trends among public agencies towards open data are beginning to change this culture, and a few agencies, such as Minneapolis's Metro and San Jose's Valley Transit Authority, now openly share the disaggregated transit survey data on their websites.

PRIVACY CONCERNS

As noted above, one reason that transit agencies resist requests to share passenger survey data is concern about sharing PII. The tremendous accessibility of information accompanying the expansion of the internet, social media, and big-data sources (combined with real threats of identity theft) has heightened privacy concerns. Transit agencies are rightfully cautious about sharing information that might directly or through cross-referencing other sources compromise the identity of survey participants. For example, information on a household's income, location, and commuting habits could conceivably invite criminal activity. Such activity might put the transit agency in a difficult legal position – particularly if the original survey had a consent statement promising to limit distribution of the data collected.

Many transit data sets include contact information of respondents, such as names, home addresses, email addresses, and phone numbers. These data fields are occasionally inadvertently released when outside users request transit survey data. To avoid this possible negative outcome, transit agency policies tend towards not releasing information at all. Requests may not be quickly filled if a transit employee needs to take additional time to review the data to scrub these PII fields from the files.

More commonly, agencies do not have an established position on what constitutes PII. While there is broad agreement that individual respondent contact information cannot be shared, less clear-cut consensus surrounds geocoded information. In the past, transit agencies often collected actual addresses of important locations such as home and work. Now, however, many transit properties use tablet computers to allow respondents to directly identify the locations of their trip ends. The broad availability of Geographic Information System (GIS) tools – including the most basic internet mapping programs – allow users to easily connect these latitude and longitude coordinates to actual places. This ability raises concerns for releasing geocoded data sets, even if no explicit contact information or street addresses are included.

An accepted solution to increase anonymity while preserving some useful location information is to systematically recode those points to higher geographies. For example, a home location that is geocoded with specific coordinates may be replaced with the coordinates for a census block group centroid or the American National Standard Institute (ANSI) code for that block group. There is not broad agreement on the appropriate geographic unit to use for such masking. One transit agency may find the block group appropriate, while another requires that point locations be recoded to the census tract or zip code level. The shift to higher-level geographies can hamper the value of the transit survey location data for certain types of research – most notably analysis of non-motorized transit access behaviors. The size of census block groups and tracts are often much larger than typical pedestrian distances to a bus stop. A challenge of the CATPAD process is to balance the loss of location specificity with the agency concern to limit potential release of PII.

OPEN DATA

Over the last several years, the context for data availability has been rapidly changing. The internet offers a perfect distribution method for data. Where in the past, publicly available data sets were physically mailed to users on a floppy disk or CD-ROM, the internet facilitates instant downloading. The interactive nature of the internet also allows for complex data requests to be fully processed without the need for human intervention, saving person-hours for public agencies. Furthermore, many major internet companies have been pursuing business strategies of making data available to the public, most notably Google.

Public agencies and governments that have made data available have been generally pleased with the array of services developed by third-party vendors to make those data useful. For example, in the transit realm, the public sharing of General Transit Feed Specification (GTFS) files (a format initially designed by Google and Portland's Tri-Met transit operator), along with automatic vehicle location (AVL) information, has resulted in a host of real-time transit information applications – which in turn has increased the use and satisfaction with transit. Many public agencies are experimenting with open data portals on their websites where they make internal data directly available to the public. At the same time, and particularly for the Millennial age cohort, expectations that public data will be open are increasing, which creates a shifting environment more supportive of data sharing.

This emerging consensus on the value of data availability provides much of the context for this development of CATPAD. The interest in making public data available has already led to several other transportation data repositories, as discussed in the next section.

III. REFERENCE DATA REPOSITORIES

While disaggregated transit survey data remains largely inaccessible, other types of transportation data have become increasingly available to the public. This section discusses a few of these resources, which have served as inspiration for both the idea and design of CATPAD.

They are an inspiration for the idea of CATPAD because they demonstrate successful and varied efforts to make transportation data available. They serve as an inspiration for the design of CATPAD because they showcase different approaches to information-sharing, providing lessons for the structure of CATPAD. In two cases, these existing databases offer information complementary to CATPAD, and thus affect the field structure of CATPAD to facilitate joining databases. None of these transportation databases provide person-level data from individual passengers responding to transit surveys.

NATIONAL TRANSIT DATABASE (NTD)

In 1974, Congress required public transit operators receiving federal grants through the Urbanized Area Formula Program (5307) or Rural Formula Program (5311) report their financial, operating, and asset condition information. These data are collected by the federal government in a systematic way to allow consistent longitudinal and cross-sectional comparison. In 1995, this resource was renamed the National Transit Database (NTD).

The NTD functions as a searchable repository of operator-level transit data. The database is maintained by the federal government and easily searched for records back to 1991. As of 2017, more than 850 transit operators from urbanized areas submit their data to the NTD via the internet.

The NTD provides an example of the utility of systematic transit data collection as well as a long-running and successful administrative model for gathering and disseminating that information. The model of transit agencies receiving federal grants reporting information back to the federal government provides a precedent for how CATPAD might function in the future. Specifically, transit agencies might be required to upload survey instruments, data, and reports directly into CATPAD.

The NTD also provides critical system-level data that can complement the passenger-level information archived in CATPAD. This linkage is facilitated by the unique 4-digit (now updated to 5-digit) numeric identifier (ID) the NTD assigns to each reporting operator. The Central Archive for Transit Passenger Data incorporates the 5-digit NTD ID for every operator to ensure easy joining to the NTD data tables. One limitation of these codes is that they only exist for transit operators receiving 5307 or 5311 federal funding. Small transit operators, such as those funded by a suburban municipality, may not be federal grant recipients and therefore have no NTD ID. This issue has emerged in the transit surveys collected by the MPO for the Twin Cities region, which extends to several local operators not reporting to the NTD.

NATIONAL TRANSIT MAP (NTM)

In 2016, the federal government released the National Transit Map (NTM), an online resource to combine General Transit Feed Specification (GTFS) data from 196 transit systems into a single resource. General Transit Feed Specification data provide schedule and stop information for fixed-route transit services and are a key input for many data analyses. The NTM provides resources for users to search and select from portions of this large data set to facilitate research and analysis. Unlike with the NTD, transit agencies voluntarily provide GTFS data to the NTM.

The NTM differs from the NTD in three ways relevant to CATPAD. The NTM combines data from several systems into a single layer; the NTM provides geocoded information; and the NTM requests but does not currently require agency participation. A long-term goal for CATPAD is to standardize specific fields from transit surveys to combine data from multiple surveys into a single file. The Metropolitan Transportation Commission in the San Francisco Bay Area is, for example, accomplishing this feat within its management of transit surveys across the more than two dozen operators in its region. In addition to aspatial data, CATPAD also provides geocoded data. Furthermore, since the NTM links the GTFS data to the NTD ID, CATPAD is able to link to the NTM data as well. While CATPAD is designed to incorporate contemporaneous GTFS files for a given system within its structure, these data are not available for older surveys. The NTM will therefore provide a consistent if not perfectly matched resource for linking CATPAD data to current fixed-transit schedules. Should the NTM also archive and provide access to historical GTFS data, it would obviate the need for CATPAD to provide the same functionality natively. Finally, the NTM is demonstrating the success of a voluntary data sharing framework – the approach currently being used by CATPAD.

METROPOLITAN TRAVEL SURVEY ARCHIVE (MTSA)

The University of Minnesota (UMN), with support from the federal government, has developed the Metropolitan Travel Survey Archive (MTSA). This resource collects, indexes, archives, and makes available data collected as part of metropolitan travel surveys conducted by MPOs to build travel demand models. The University of Minnesota reached out to major MPOs to request the data voluntarily and then developed protocols for indexing and storing the surveys that were shared with them.

The Metropolitan Travel Survey Archive is a major inspiration for CATPAD. The Metropolitan Travel Survey Archive represents a researcher-initiated project to preserve valuable but scattered and difficult-to attain-data sets, standardize their structure, and make them openly available on the internet – all key elements that CATPAD is applying to transit survey data. Furthermore, MTSA represents a partnership between a university and local and federal sponsors which is being emulated by CATPAD.

Unlike MTSA, however, CATPAD is focused on transit surveys and not regional travel surveys. The difference has important ramifications. MTSA gathers data from a relatively small set of major metropolitan planning organizations (as well as a few state departments of transportation) with dedicated travel-demand modelers and very large-scale surveying

efforts. The requirements of travel-demand modeling provide an inherent structure to the metropolitan travel surveys that facilitates standardization. Furthermore, the large-scale nature of these efforts typically reduces the number of consultants capable of carrying them out and the frequency with which they occur – typically every ten years (although slightly more for the most active MPOs). By contrast, CATPAD gathers data from many times more agencies with substantially more variance in their in-house technical capacity and survey design and generally higher rates of survey frequency. These distinctions will require more labor effort to populate and maintain CATPAD and more challenge to standardizing data types in comparison to MTSA.

NATIONAL RENEWABLE ENERGY LABORATORY (NREL) TRANSPORTATION SECURE DATA CENTER (TSDC)

As public concerns over PII rose along with the increased geographic precision of travel surveys, the federal government established the Transportation Secure Data Center (TSDC) at the National Renewable Energy Laboratory. The TSDC is similar to MTSA in gathering, archiving, and making metropolitan travel survey data available via the internet; there are, however, some critical differences. The TSDC archives data in a single platform (PostgreSQL), but without an attempt to standardize different survey efforts. The key difference is that the TSDC introduces several security protocols on the use of the fully geocoded data. The TSDC does allow easy download of cleansed data sets where sensitive locations are geocoded to block groups rather than to specific points.) Researchers must apply to use the detailed data by providing their organizational credentials as well as their specific research protocol. Once approved, research can only access and analyze the data via a virtual computer in a secure environment on the TSDC servers. This environment has a preestablished set of available software for analyzing the data. While the disaggregated data can be analyzed within the secure environment, only aggregate results approved by the TSDC can be taken out of the secure environment. Users must prepare files that are reviewed and then made available for download by the TSDC.

The TSDC provides critical access to highly-detailed, recent data sets. It offers an array of analytical tools as well as technical support for users. However, the TSDC, by design, establishes several barriers to easy use and requires a relatively high level of data sophistication. Like MTSA, it is currently limited to metropolitan travel surveys. The Central Archive for Transit Passenger Data follows elements of the TSDC, particularly the conversion of data to a standard format, without standardization of the data elements. Because CATPAD is designed as an open resource, it follows the recoding technique that TSDC applies to its clean data sets.

IV. DATABASE DESIGN

This section presents the organization of CATPAD as a relational database. This organization reflects the anticipated use cases of CATPAD.

ANTICIPATED USE CASES

The Central Archive for Transit Passenger Data is a resource designed from inception to meet the distinct needs of different users, including transit operators, survey consultants, travel demand modelers, advocacy groups, oversight agencies, metropolitan planning organizations, academic researchers, etc. Anticipated use cases of CATPAD include methodological comparison, survey partner exploration, policy analysis, travel behavior research and travel-demand modeling.

Methodological Comparison

It is beneficial for any entity undertaking a transit survey to know how similar efforts have been conducted elsewhere. Such comparison can allow methods to be refined and best practices to emerge. Surveying entities may be interested in refining how questions are worded, what questions are asked, how passengers are intercepted, which formulations improve response rates, etc.

The ability to compare methods may be particularly important for the uptake of innovative approaches, such as tablet-based personal interviews, which diverge substantially from traditional surveying practices. Transit agencies and their funders may prefer to deeply explore a range of approaches (and the resulting data) before making a potentially risky decision to try a new surveying approach.

Survey Partner Exploration

Survey comparison also allows surveying entities to explore the work of others and identify possible project partners. Agencies considering a survey or consultants looking for partners would all benefit from being able to explore the work of their peers. Issues of interest may be how different partners design and implement surveys as well as how they report and analyze the results. Such comparison should lead to a more efficient market for surveyors and higher quality surveys.

Policy Analysis

The detailed information provided in a travel survey allows for a broad range of policy analyses: a transit operator may be interested in the effectiveness of a new route, a funding agency may be concerned about the revenue implications of a new fare regime, an advocacy organization may want to ensure that transit services are improving equity outcomes on disadvantaged populations, or an oversight board may want to track customer satisfaction. While currently these analyses tend to focus on a single survey due to the difficulty of gathering multiple surveys, such analyses could be vastly improved by making either (or both) cross-sectional and longitudinal explorations using multiple data sets.

Travel Behavior Research

Transit passenger survey data are one of the most comprehensive resources for exploring travel behavior research. The detailed and disaggregated nature of survey responses make these data particularly useful for the application of inferential statistical techniques such as econometric analysis. Moreover, transit passenger survey data can directly query respondents on important travel and demographic characteristics, such as trip purposes, ethnicity, income, and policy preferences, that can only be imputed from more passive data sources, like cell phone traces.

ATTRIBUTE CODING

The use cases above structured the indexing structure of CATPAD. For example, a crosscutting need among the use cases is the ability to filter a large sample of surveys according to certain criteria. Accordingly, CATPAD has fields to code for a broad range of attributes.

Basic Information

Several basic attributes are collected that describe the survey. These attributes represent details that are unique to a given survey and not linked to other fields directly. Such attributes include:

- Survey title
- Number of survey respondents
- Survey response rate
- Surveying start date
- Surveying end date
- Descriptive text fields
    - General description
    - Notes on the entry
    - Geocoding description

The entry form on CATPAD is presented below in Figure 1. This screenshot does not show all the entry fields due to space limitations. The remaining fields are below the current screenshot; nonetheless, Figure 1 presents the reader with a sense of the CATPAD form for basic information.

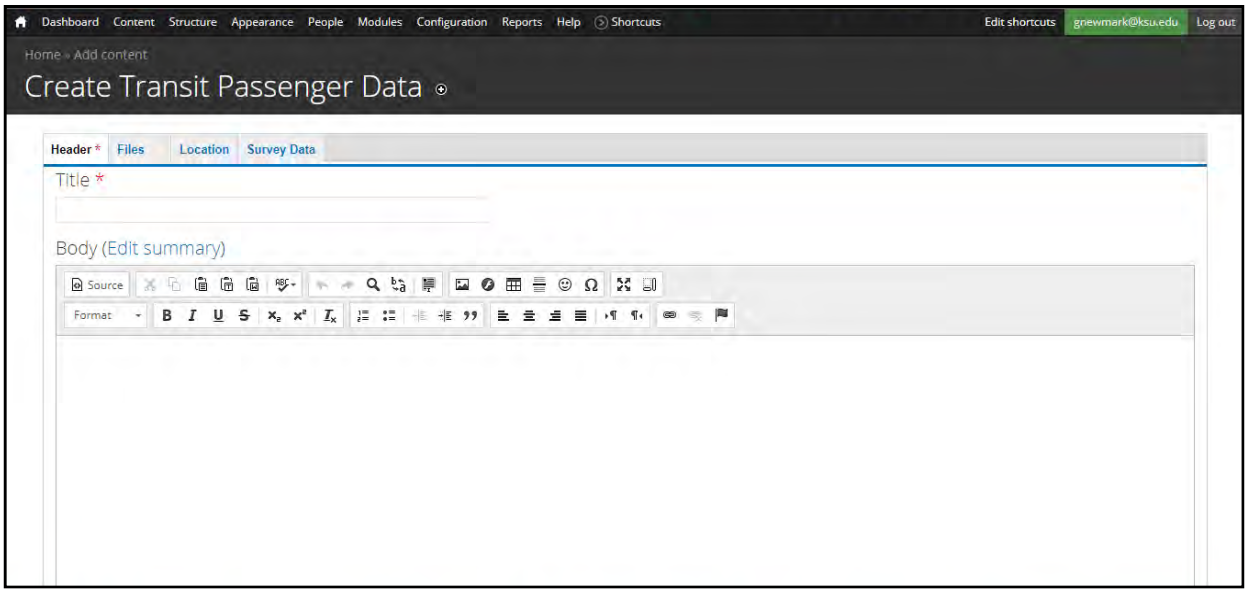

Figure 1. CATPAD Data Entry – Header Tab

Location Information

These data simply identify the country and state/province where the surveying took place. Since many transit properties span state lines, multiple states can be selected. Currently, the only locations available within CATPAD are the US and Canada and the associated states or provinces. These fields are included so that users can sort data by location. For example, a state department of transportation may want to select all the available surveys that have been conducted in their jurisdiction.

Figure 2, below, shows the entry field for the location information. Once the country field is selected, only the associated states or provinces are made available for selection; within a given country, however, multiple states or provinces can be selected, as many regional transit systems cross the boundaries of these jurisdictions. For example, transit services from a single operator in Kansas City serve both Kansas and Missouri residents.

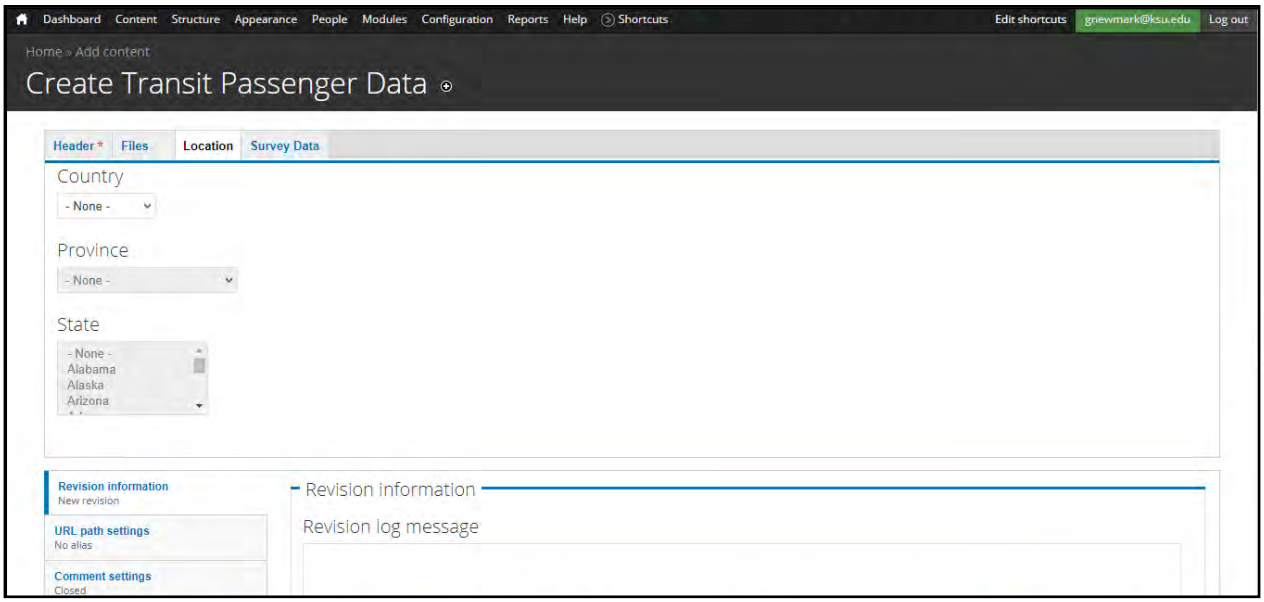

Figure 2. CATPAD Data Entry – Location Tab

Transit Agency

Users can select all the transit agencies covered by a given survey. CATPAD is currently pre-populated with the transit agencies that report to the NTD. (Future CATPAD upgrades will allow users to enter agencies that are not listed in the NTD.) To avoid possible confusion among similarly named agencies, the pre-populated list includes state and NTD codes. Once the relevant transit agencies are selected, all the relevant NTD 5-Digit IDs are recorded as well. This feature allows users to readily link CATPAD's disaggregated information to the aggregated information available in the NTD and NTM resources, enhancing the value of all three resources.

Transit Mode

Users can select all the different transit modes covered by a different survey. The modal codes themselves are drawn from the NTD. CATPAD follows NTD practice in allowing users to select macro-categories of Rail and Non-Rail modes and then allowing the selection of multiple sub-modes as enumerated below:

Rail Modes

- Heavy rail
- Commuter rail
- Light rail
- Streetcar

- Monorail or automated guideway
- Cable car
- Hybrid rail
- Inclined path
- Alaska railroad

Non-Rail Modes

- Bus
- Demand response
- Trolleybus
- Commuter bus
- Ferryboat
- Bus Rapid Transit
- Vanpool
- Publico
- Demand response taxi
- Aerial tramway
- Jitney
- Other

While the user sees the full name, the NTD two-letter modal codes are also included internally to allow linkages to the NTD.

Survey Descriptors

Users can also select a range of options relating to the survey design. The current options are presented below:

Survey Mode

- Self-complete
- Personal interview

Survey Elements

- Travel behavior
- Demographics
- Customer satisfaction
- Stated preference

Survey Technology

- Paper
- Online
- Tablet
- Computer-Assisted Telephone Interviewing (CATI)

Survey Recruitment

- Intercept
- Panel
- Address
- Telephone
- Online

Incentive

- None
- Money
- Gift card
- Transit rides

These options represent the current structure of CATPAD and will likely be augmented with future development.

Additionally, users can select the consultants involved with the survey as well as all the available languages for that survey. Users can add consultants as part of the data entry while the languages are a pre-populated list based on Google translate options.

Figure 3, below, provides the input fields for all of the transit and survey data elements. While this screenshot only shows the topmost portion of the form, the user would scroll down to complete the data entry.

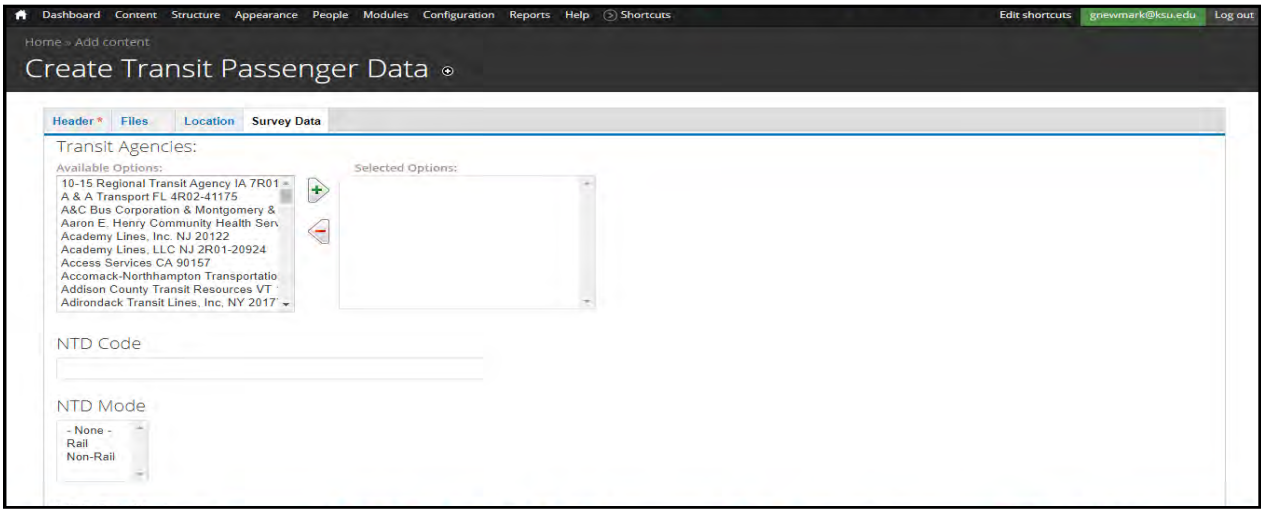

Figure 3. CATPAD Data Entry – Survey Data Tab

DATA FILES

The primary purpose of the attribute coding is to be able to select the appropriate data files for use. Each survey record has slots for different file types with associated file formats. These include:

- Instrument (pdf)
- Data (csv)
- Data dictionary (pdf)
- Report (pdf)
- Presentations (ppt, pptx, pdf)

Preparing these files typically requires some format conversion. These conversions are all completed prior to uploading into CATPAD to ensure consistency for users.

In addition, each survey has slots for data files that, while not including disaggregated passenger transit data, provide key information on the context of the transit system at the time the survey was administered. These include:

- Automatic vehicle location (pdf currently, may switch to zip)
- On-off counts (zip)
- General transit feed specification (zip)

These contextual data files can be more difficult to come by and, particularly for older surveys, may never have been created; however, CATPAD preserves places for their storage to make them available when possible.

Figure 4, below, shows the upper portion of the CATPAD entry form for files. The form limits the file types that can be uploaded and limits the number of files to one, except for presentations (which is off the visible portion of the screen).

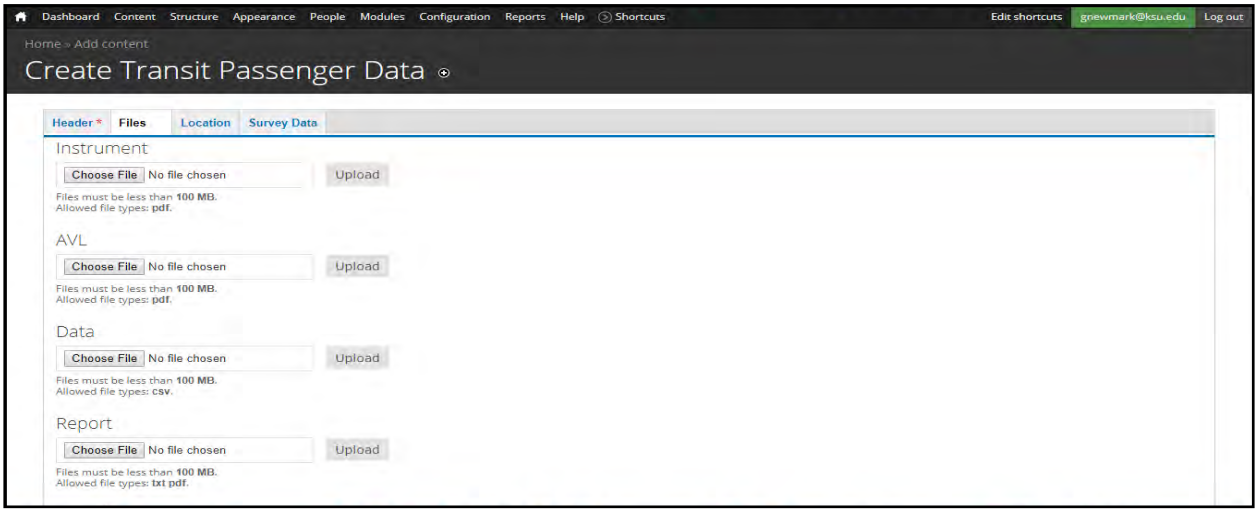

Figure 4. CATPAD Data Entry – Files Tab

DATA SECURITY

Data can only be entered by authorized personnel using password-protected authentication. The Central Archive for Transit Passenger Data is a MySQL database that is hosted by Pantheon as a cloud-based application. Backups are currently manual as the site has not been moved into production mode with a domain name.

To track and minimize any possibility of data loss or corruption, all content revisions are tracked and can be undone. Each time a transit agency or transit passenger data content entry is updated, the previous version remains in the database. This is so that if there are mistakes in the update, or a need to revisit the data that previously existed, one can view all previous revisions and revert to whichever version is desired. When one reverts to a previous version, that version becomes the current version, but all the subsequent regions are preserved as well.

Figure 5, below, demonstrates how the revision information appears on the data entry forms for CATPAD. The bottom portion of the main header tab, where the basic survey information goes, is shown here.

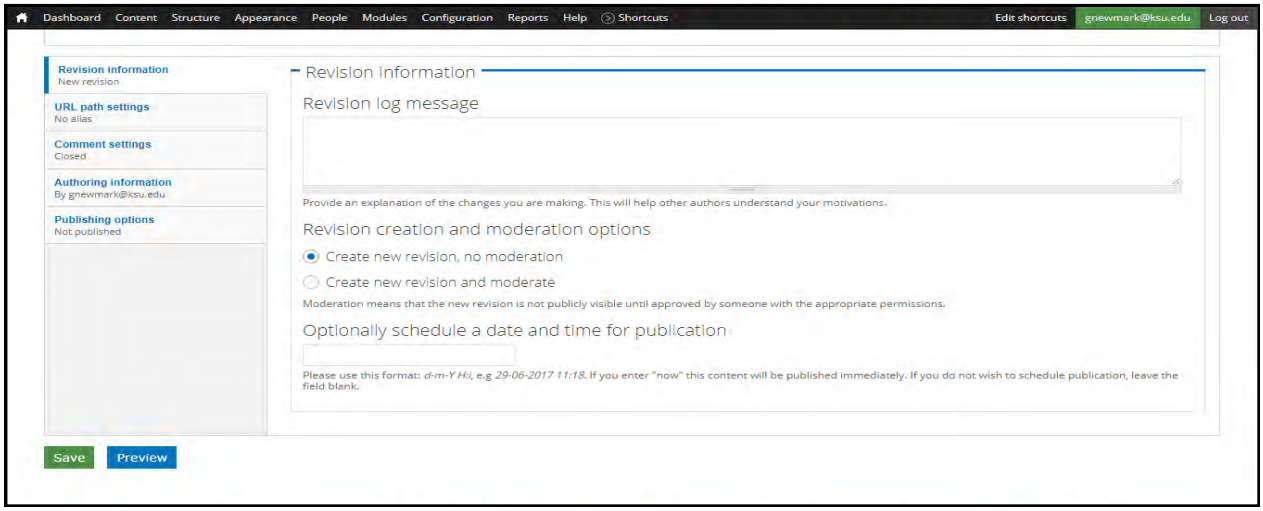

Figure 5. CATPAD Data Entry – Header Tab – Revision History

FRONT-END SEARCH

The functional objective of CATPAD is to make all the records (including their associated data) searchable and openly available. This is currently possible through the front end of the database. Currently, the database represents the basic functionality. It will be further developed as noted below in the Future Steps section.

From the home page, users can do a quick text search of the database, which will return listings based on the titles of the transit Agency or the survey effort. Clicking on the title of the listing will reveal the full content to the user, as shown below in Figure 6.

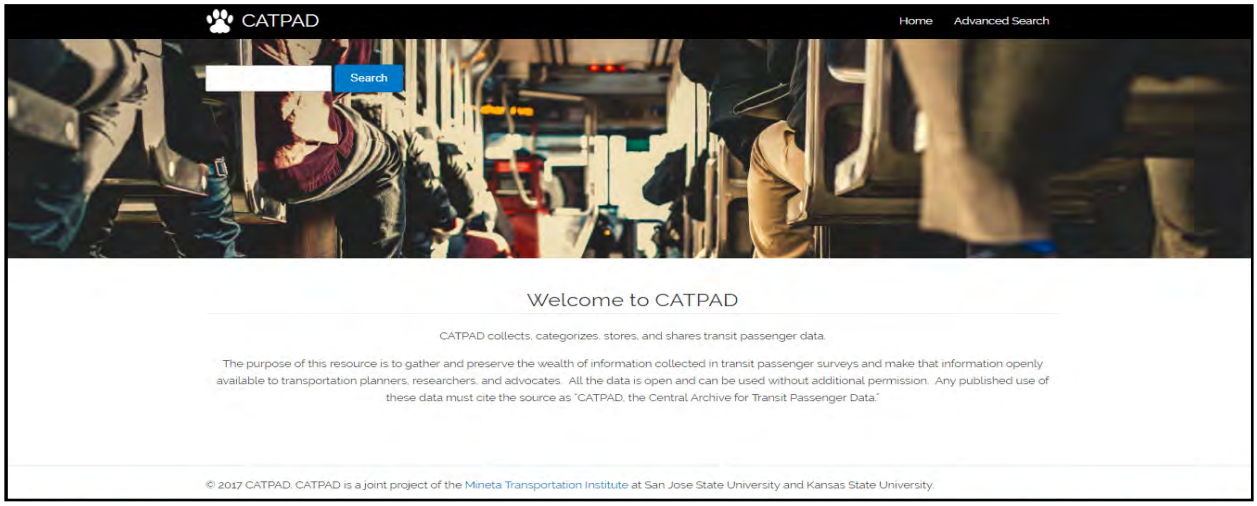

Figure 6. CATPAD Home Page with Quick Search

ADVANCED SEARCH

The Advanced Search feature allows users to filter the results based on a large set of criteria utilizing drop-down menus. The drop-down menus correspond to the fields discussed above and allow users to sort through the entire database in order to zero in on the records of interest.

Figure 7, below, presents a screenshot of the advanced search page.

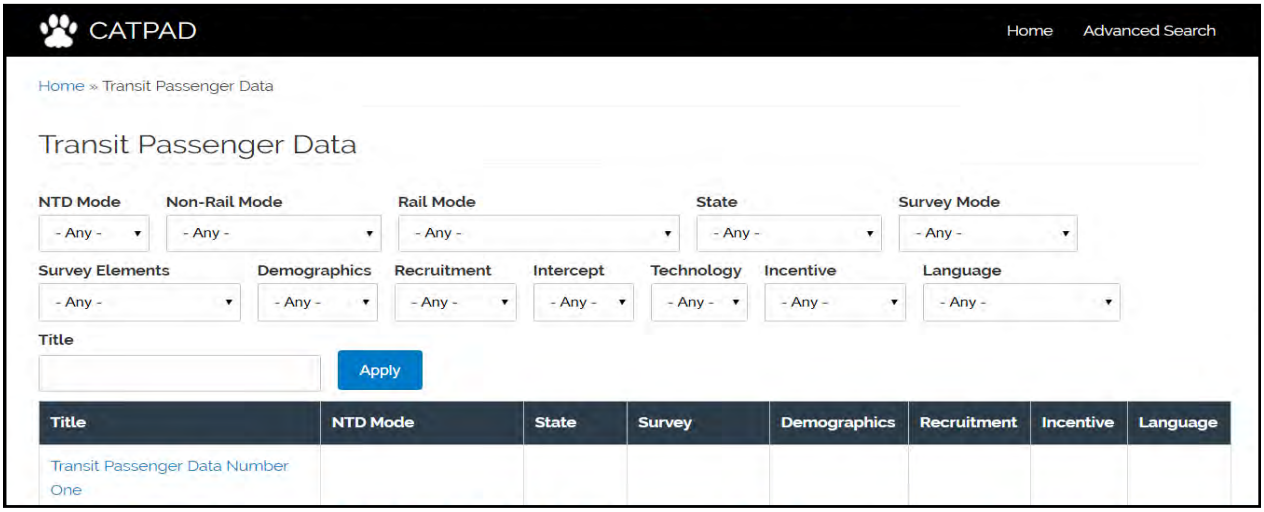

Figure 7. CATPAD Advanced Search

There is also an open text field to allow further narrowing of the data set by title.

V. CATPAD USER SCENARIOS

This section illustrates how several likely users might interface with CATPAD. The purpose of this section is to suggest a range of possible applications.

TRANSIT PEER COMPARISON

The board of a large transit agency wants to compare the customer satisfaction on their commuter rail lines with commuter rail operations elsewhere. Transit agency staffers would respond to the board request by querying CATPAD for all customer satisfaction surveys that included respondents on commuter rail – both are attributes with dedicated field codes. The staffers would download the surveys returned by the query and compare the results. Depending on the board concern, it may be sufficient to compare the aggregate information listed in the survey reports; however, if the board would like a more nuanced analysis, the staffer could explore the underlying data to, for example, break down satisfaction by gender or age cohorts. The staffer may rely on the survey instrument to explain how customer satisfaction questions varied among systems and the means that he or she used to align them for the analysis.

EQUITY ANALYSIS

A community group feels that the regional transportation plan under development by the metropolitan planning organization does not sufficiently consider the transit use of low-income residents. The community group queries CATPAD to identify all transit surveys conducted on the three local transit providers. The community group downloads the survey data and conducts its own analysis to show that over the last fifteen years and three rounds of surveys, the average in-vehicle times for low-income riders have increased, while those same times have decreased for higher-income riders. Faced with this evidence, the metropolitan planning organization decides to reconsider its 20-year plan.

METHODOLOGICAL COMPARISON

A major transit agency is facing an active internal debate about using personal interviewers with handheld tablet computers for their next transit survey. To explore the question in a rigorous fashion, they query CATPAD for all surveys using that delivery mode. Then using the NTD IDs associated with each survey, they join their agency list to the NTD to filter out operators that do not meet a specific threshold of annual boardings. They then explore the results of the remaining surveys to address their concerns about this mode. One dissenting voice feels this approach is limited. She argues that the tablet surveys should be compared to paper surveys on similar systems. She also goes to CATPAD to find appropriate paper surveys for the comparison.

TRAVEL BEHAVIOR RESEARCH

A professor at a land-grant university in a rural state wants to explore the changing demographics and use of rural transit services in his state. He queries CATPAD for all bus operators in the state and excludes the operators in the state's two urban areas. His

research finds that rural riders are increasingly elderly and using transit for medical visits and grocery. His research convinces the state legislature to increase funding for rural transit as a lifeline to rural areas.

TRAVEL-DEMAND MODELING

A consultant is hired by a metropolitan planning organization to update their travel-demand model. The region has limited moneys to collect additional data. The consultant uses the CATPAD data from the region's operators to estimate key coefficients to inform the model without requiring additional surveying.

VI. CURRENT STATUS AND FUTURE STEPS

This project has succeeded in establishing CATPAD, the Central Archive of Transit Passenger Data. This development responded to a need from many stakeholders for a resource to store and share transit passenger data. The CATPAD development was guided with reference to other data repositories serving the transportation community. Consequently, in addition to identifying features from other pre-existing archives, CATPAD is coded to be easily joined to both the NTD and the NTM – both resources maintained by the US DOT.

The main portion of the establishment of CATPAD was split into two activities. The first was working with selected transit agencies, departments of transportation, and metropolitan planning organizations to identify the concerns regarding data sharing. The second was developing the database structure itself to effectively meet the needs of anticipated users. These activities yielded a process for gathering information as well as a robust repository for storing and sharing that information.

These activities were both successful, and CATPAD exists as a fully functional searchable database with a clear process for acquiring data from surveying agencies and entering that information into the database. The resource is hosted by the Mineta Transportation Institute as a joint venture with Kansas State University. The Central Archive of Transit Passenger Data has a small set of surveys, primarily those used to design the resource, and is growing slowly as more are added. While CATPAD is fully functional, it has not been publicly launched, because the next steps are to richly populate the archive and improve its user friendliness, as noted below.

While this project only focuses on the establishment of CATPAD, discussing future steps for the archive is warranted. The next stage of development incorporates two tracks. The first is to richly populate CATPAD with transit passenger data sets. The second is to refine the user interface to facilitate easy engagement. Of these two tracks, the former will require far more time and resources.

This project revealed that while transit agencies are wholly supportive of CATPAD, they struggle to make the time to gather the data files necessary to populate CATPAD. While much of this resistance is likely to be overcome by the public availability of CATPAD, some of it is structural. Transit agencies have relatively limited time to spend on gathering historical data when the pressing needs of operating the transit system are at hand. The establishment stage of CATPAD covered in this report only worked with select agencies that had a personal connection to the authors. Even then, it took many emails, phone calls, and reminders to gather the data sets. Furthermore, it took considerable time to process each data set, both to identify the information for indexing and then to enter that information. All things considered, it is clear that a full ramp-up of CATPAD will require substantial person-hours. The FTA has discussed possibly making contribution of data to CATPAD mandatory in the future for agencies receiving federal grants. That requirement would speed the process, but will come only once CATPAD is well-established.

Refining the CATPAD user interface will be much less time-intensive, but will require a series of meetings and webinars with key users to get feedback. In addition, it is intended that a page of links be added to transit survey resources to fully make CATPAD a one-stop shop for everything related to collecting disaggregated transit passenger data. Finally, a longer-term goal is to add a third component that will provide links to research using CATPAD.

There are several medium-term possibilities for fostering the use of CATPAD. In addition to more traditional presentations of the resource to key stakeholders via live presentations and webinars, it would be reasonable to host a conference sponsored by the TRB, possibly via the Transit Subcommittee of the Travel Survey Committee. That conference could focus on developments in the collection and use of transit passenger data. Finally, it would be very reasonable to host a hackathon on using the data for policy analysis. One strong possibility would be to focus on equity concerns and to be co-hosted by an advocacy organization with strong transit commitment, such as PolicyLink.

The ultimate goal of CATPAD is to enhance the analysis of transit to improve the nation's public transportation systems. This establishment of CATPAD represents a first and critical step in that effort.

ACRONYMS AND ABBREVIATIONS

ANSI	American National Standard Institute
CATI	Computer-Assisted Telephone Interviewing
CATPAD	Central Archive for Transit Passenger Data
DOT	Department of Transportation
FOIA	Freedom of Information Act
FTA	Federal Transit Administration
GIS	Geographic Information System
GTFS	General Transit Feed Specification
ID	Identifier
MPO	Metropolitan Planning Organization
MTSA	Metropolitan Travel Survey Archive
NREL	National Renewable Energy Laboratory
NTM	National Transit Map
NTD	National Transit Database
PII	Personally Identifiable Information
RTA	Regional Transportation Authority
TRB	Transportation Research Board
TSDC	Transportation Secure Data Center
UMN	University of Minnesota

ABOUT THE AUTHORS

GREGORY NEWMARK, PhD

Dr. Newmark is an assistant professor of Landscape Architecture and Regional & Community Planning at Kansas State University. His research focuses on the linkages between land use, travel behavior, and the environment, with a particular interest in public transportation policy, use, and finance. He holds a PhD in City and Regional Planning from the University of California, Berkeley, a MSc in Urban and Regional Planning from the Technion - Israel Institute of Technology, and a BA in History and International Studies from Yale University.

HILARY NIXON, PhD

Dr. Nixon is a Professor of Urban and Regional Planning at San José State University. Her research and teaching interests in environmental planning and policy focus on the relationship between environmental attitudes and behavior, particularly with respect to waste management and linkages between transportation and the environment. She holds a BA from the University of Rochester in Environmental Management and a PhD in Planning, Policy, and Design from the University of California, Irvine.

PEER REVIEW

San José State University, of the California State University system, and the MTI Board of Trustees have agreed upon a peer review process required for all research published by MTI. The purpose of the review process is to ensure that the results presented are based upon a professionally acceptable research protocol.

Research projects begin with the approval of a scope of work by the sponsoring entities, with in-process reviews by the MTI Research Director and the Research Associated Policy Oversight Committee (RAPOC). Review of the draft research product is conducted by the Research Committee of the Board of Trustees and may include invited critiques from other professionals in the subject field. The review is based on the professional propriety of the research methodology.

MTI FOUNDER

Hon. Norman Y. Mineta

MTI BOARD OF TRUSTEES

Directors

Research Associates Policy Oversight Committee

MTI MINETA TRANSPORTATION INSTITUTE