Dynamic Data Maintenance for Quality Data, Quality Research

Dilruba Ozmen-Ertekin, Ph.D., PE (Corresponding Author)
Assistant Professor
Engineering Department
Hofstra University
133 Hofstra University
Hempstead, NY 11549
Phone: 516-463 6155
Fax: 516-463 4939
email: eggdzo@hofstra.edu

Kaan Ozbay, Ph.D.
Associate Professor
Civil and Environmental Engineering Department
Rutgers University
623 Bowser Rd.
Piscataway, NJ 08854
Phone: 732-445 2792
Fax: 732-445 0577
email: kaan@rci.rutgers.edu

Paper Submitted for Presentation at the Transportation Research Board’s 90th Annual Meeting, 2011, Washington, D.C.

Statewide Transportation Data and Information Systems Committee (ABJ20)

Re-Submission Date: November 15, 2010
Word Count: 7602 (5102 words + 4 Tables + 6 Figures)
ABSTRACT
Just like any other scientific research field, the value of data quality is undisputed in the field of transportation. From policy planning to performance evaluation, from model development to impact studies, good quality data is essential to generate ideas and clear-cut solutions to be implemented by transportation professionals and decision makers.

In order to improve scientific data quality and function within a continuous quality assessment and management framework, research and development organizations and agencies constantly look for the latest methodologies and technological tools of data management. The New York Metropolitan Transportation Council (NYMTC), for example, has recently awarded a research project titled “Improvements on NYMTC Data Products” to the authors of this paper in an effort to modernize the existing data products (i.e., reports and brochures, both printed and online) and improve the communication between the agency and the public. The main goal of the research project was to perform a through review and examination of NYMTC data products to identify specific issues about the existing data products, website and the current data maintenance process at NYMTC, and then suggest appropriate solutions, both data and website, and process oriented.

This paper reports the results from this research project by giving special emphasis to the issues and solutions related to the data maintenance process.

Keywords: Data Quality, Data Life Cycle, Dynamic Data Maintenance
1. INTRODUCTION

Recognizing the need for quality data for quality scientific research, many organizations and agencies, including those in the field of transportation, constantly look for the latest methodologies and technological tools of data quality and management. The New York Metropolitan Transportation Council (NYMTC), for example, has recently awarded a research project titled “Improvements on NYMTC Data Products” to the authors of this paper in an effort to modernize the existing data products (i.e., reports and brochures, both printed and online) and improve the communication between the agency and the public. Another example regarding transportation agencies’ efforts towards quality data is the research project conducted for the Federal Highway Administration focusing specifically “on quality control procedures, review of archived data that have already been collected and saved by traffic operations systems” (12). As another example, (13) provides guidelines “to improve the quality of the traffic information that supports decisions at all levels of the transportation profession.

This paper reports the results from the aforementioned NYMTC research project. NYMTC is an association of governments, transportation providers and environmental agencies that is the metropolitan planning organization for New York City, Long Island and the lower Hudson Valley. NYMTC’s mission is 1) to serve as the collaborative forum to address transportation-related issues from a regional perspective; 2) to facilitate informed decision-making within the Council by providing sound technical analyses; 3) to ensure the region is positioned to capture the maximum federal funds available to achieve the goals of the Unified Planning Work Program, Regional Transportation Plan and Transportation Improvement Program; and 4) to focus the collective planning activities of all Council members to achieve a shared regional vision (1).

In order to fulfill its mission, NYMTC produces an array of information in support of the planning and policy decisions that shape the region. This information is produced in the form of various reports, brochures and databases, and is used by many transportation specialists and academics both in the NYMTC region and outside the region as well as people from other countries. Given this, NYMTC has recently embarked on a research project, performed by the authors of this paper, in order to modernize the existing data products and ensure that an overall quality of the data products is achieved and maintained. It should be noted that this project is already finalized and the final report was published in November 2009 (10).

The quality of data has various dimensions defining the characteristics of data in measurable forms. A data quality dimension is defined as “a set of data quality attributes that most data consumers react to in a fairly consistent way” (2). Table 1 below provides different categories and dimensions of data quality (3, 4). These dimensions can be used as a starting point in determining the attributes related to data quality that are most suitable and relevant to specific industries.

Just as important as the data quality, is to have a well-established dynamic data maintenance process in place and to adopt an organizational approach in which all the agency’s technical staff, agencies providing data, and other stakeholders make improving the quality of data products a priority.
TABLE 1 Dimensions of Data Quality (3,4)

<table>
<thead>
<tr>
<th>Category</th>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic</td>
<td>Accuracy</td>
<td>The level to which stored data agree with accepted sources of “correct” information</td>
</tr>
<tr>
<td></td>
<td>Objectivity</td>
<td>Whether the information was objectively collected</td>
</tr>
<tr>
<td></td>
<td>Believability</td>
<td>The degree of credibility or trustworthiness of the information</td>
</tr>
<tr>
<td></td>
<td>Reputation</td>
<td>Whether the information has a good reputation for quality</td>
</tr>
<tr>
<td>Accessibility</td>
<td>Accessibility</td>
<td>The ease and breadth of access to information</td>
</tr>
<tr>
<td></td>
<td>Access Security</td>
<td>Whether the information is protected against unauthorized access</td>
</tr>
<tr>
<td>Contextual</td>
<td>Relevancy</td>
<td>Whether the information is useful / relevant / appropriate / applicable to intended purpose</td>
</tr>
<tr>
<td></td>
<td>Value-Added</td>
<td>Overall quality so as to add value</td>
</tr>
<tr>
<td></td>
<td>Timeliness</td>
<td>Freshness and up-to-date state of information</td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
<td>The expectation that certain attributes are expected to have assigned values in a data set</td>
</tr>
<tr>
<td></td>
<td>Amount of Data</td>
<td>Sufficiency/Insufficiency of information</td>
</tr>
<tr>
<td>Representational</td>
<td>Interpretability</td>
<td>Enables the users to make accurate inferences</td>
</tr>
<tr>
<td></td>
<td>Ease of Understanding</td>
<td>Format and presentation that is easy to understand</td>
</tr>
<tr>
<td></td>
<td>Concise Representation</td>
<td>Whether the information is formatted compactly and presented concisely</td>
</tr>
<tr>
<td></td>
<td>Consistent Representation</td>
<td>Whether instances of data are represented in a format that is consistent with the domains of values as well as consistent with other similar attribute values</td>
</tr>
</tbody>
</table>

Loshin (5) argues that “information follows a ‘lifecycle’ consisting of creation, distribution, access, updating, and retirement stages, and that it is necessary that a data quality framework provide protocols for measuring the quality of information at the various stages of that life cycle”. That is why one of the objectives of the NYMTC research project was to identify specific issues about the current data maintenance process at NYMTC and propose an agency-wide, systematic, easily understandable and dynamic data maintenance procedure to guide the NYMTC staff in moving forward with effective data collection and dissemination efforts.

Another objective of the NYMTC research project was to identify specific issues about the existing data products and website, and suggest appropriate solutions.

In order to achieve these two objectives, a thorough review and examination of NYMTC data products from data quality point of view was performed while taking into consideration the latest technological tools, results of the NYMTC’s stakeholder survey, the best practices in various fields in the USA and around the world, and NYMTC’s current data maintenance practices.

This paper reports the results from the NYMTC research project by giving special emphasis to the issues and proposed solutions related to the dynamic data maintenance process.

The organization of this paper is as follows. The next section provides a summary of the data, website, and data maintenance process oriented issues. Section 3 presents specific solutions to the issues identified. Possible ways for the evaluation of the potential impacts from the research described in this paper is given in Section 4. Concluding remarks are presented in the last section.
2. DATA, WEBSITE, AND PROCESS ORIENTED ISSUES

In identifying the issues with NYMTC’s data products, website, and current data maintenance practices, three main factors were taken into consideration.

1. A comparison between NYMTC and best practices was made,
2. Information on the ‘expectations of the NYMTC stakeholders’ was obtained from investigating the responses to the stakeholder survey,
3. The current data maintenance process at NYMTC was reviewed.

In addition to these, understanding the ‘NYMTC’s point of view as the Metropolitan Planning Organization (MPO) serving the New York City, Long Island and the lower Hudson Valley region gave the authors a clear direction to go to in identifying the problem areas that require further improvements. As the largest MPO in the nation, NYMTC provides a collaborative planning forum to address transportation-related issues from a regional perspective and plans and makes decisions on the use of federal transportation funds (1). Thus, the availability, utility, and quality of data is very important for the MPO to properly address a wide range of issues, whether through policy planning, performance evaluation, model development or impact studies, to generate ideas and clear-cut solutions for informed decision making. In fact, one of NYMTC’s new endeavors called iNYMTC, taken as part of the Strategic Data Coordination in the NYMTC Region, aims to offer integrated, interactive, easily accessible data to its wide range of users that comes with sufficient metadata, is GIS-based and user friendly.

2.1. Issues Identified Using Comparison between NYMTC and Best Practices

NYMTC’s data products and website were reviewed to reveal their strengths and weaknesses and to compare them to the best practices that were reviewed by the authors. After inspecting the websites for every department in the U.S. government to identify industries that have various and large amounts of data for the public, and then, for each identified industry, selecting the institutions and agencies with the highest number of citations as data providers, an initial list of sixty plus websites was generated. Then, the final list of best practice websites was prepared using this initial list while keeping in mind three main criterion: total number of databases available, data coverage, and accessibility. For transportation related websites, other state DOTs’ and MPOs’ websites were also considered, as well as other English-speaking countries’ transportation websites. As a result, a final list of 16 best practices, covering different industries including transportation, weather, real estate, finance, health and forestry, both from public agencies and private firms located in the U.S. or overseas, was obtained.

Inefficient data organization and limited accessibility due to occasional lack of tighter integration throughout the NYMTC website, partly limited data archives, long lag times in releasing data products, insufficient and in some cases nonexistent metadata and variable dictionaries, too few format choices, lack of custom data generation and extraction features and insufficient user support and feedback mechanisms were among the possible areas of improvement for NYMTC’s data products when compared to the best practices reviewed.

As far as the advanced services are concerned (e.g. the availability of raw data and statistical analysis feature), there was an equality between NYMTC and the most (75%) of the best practices reviewed; neither of them allow the user to do statistical analysis with the chosen data. For the 25% of the sites that do allow analysis, count, sort, cross tab, graph generation on
map, scattergram, histogram, time-series trend analysis in data, and spatial hotspot analysis on map are allowed in the best practices reviewed.

2.2. Issues Identified Using Responses to Stakeholder Survey
NYMTC’s Stakeholder Survey was examined thoroughly to understand stakeholders’ specific concerns and recommendations about the NYMTC data products and website. The web-based survey carried out by NYMTC in April 2007 had about a 20% response rate. A careful investigation and analysis of the responses to the survey revealed important information about the stakeholders’ data usage, familiarity with the data products, preferences regarding formats, general concerns and suggestions for further improvements. As shown in Figure 1, a one-page survey fact sheet was prepared by the authors of this paper using the responses to the survey so that the overall results could be presented in a compact fashion.

According to Figure 1, the results indicated that Travel Patterns in the New York Metropolitan Area Brochure and Regional Transportation Statistical Report are the top two most popular and mostly used data products. Freight Facilities and System Inventory turned out to be a data product that the users are least familiar with. The survey also revealed that the users’ most preferred data formats are Microsoft Excel spreadsheets and full-color downloadable Portable Document Formats (PDF). Data insufficiency and timeliness issues were among the biggest concerns of the users. Accuracy, consistency, appropriateness of format and ease of understanding, and user feedback related issues were also sources of some concern for the stakeholders. These invaluable pieces of information obtained from the survey were used as inputs in identifying and recommending solutions for the data, website and data maintenance process related issues.
NYMTC STAKEHOLDER SURVEY - FACT SHEET

IMPROVEMENTS ON NYMTC DATA PRODUCTS: WEB-BASED USER SURVEY

**Overview**
NYMTC carried out a web-based survey in April 2007 consisting of 7 questions aiming to find out the most popular data products and formats, and to hear people’s suggestions about various ways to improve data formats and contents. Surveys were sent to 162 participants of public and private transportation agencies in New York Metropolitan area and suburbs, including NYMTC member agencies, via e-mail which included a link to the survey hosted by SurveyMonkey.com. In the end, there were 32 responses (about 20% response rate).

**Profile of Survey Respondents**

- Private Agency, 19%
- University, 3%
- Public Agency, 77%

The following are some notable observations from the survey results:

**Usage and Familiarity**

**Top 5 Mostly Used Data Products**
- Travel Patterns in the New York Metropolitan Area Brochure-Quarterly, and Regional Transportation Statistical Report-Annually (69% each),
- Regional Transportation at a Glance Brochure-Annually (62%),
- Hub-Bound Travel Report-Annually (59%),
- Regional Demographics at a Glance Brochure-Annually, Truck Toll Volume Report-Annually, and County Profiles (55% each),
- Hub-Bound Travel at a Glance Brochure-Annually (52%).

**Top 5 Data Products Users are Unfamiliar With**

- Freight Facilities & System Inventory 50%
- Transp. Safety at a Glance 40%
- Truck Terminal & Warehouse Survey 43%
- Review of Technologies Used in Freight Transp. 43%

**Preferred Formats**
- 97% preferred MS Excel spreadsheet format of data tabulations,
- 92% preferred full-color downloadable PDF,
- 88% would like web-based queries to be able to generate interactive tables, charts, maps.

**Format Preference**

![Format Preference Chart]

**Biggest Concerns/Suggestions for Improvements**
- **Accuracy (5%)**
  - Cross-check with other agencies.
- **Consistency (11%)**
  - Consistency check with other agencies (5%) (specifically, Transit Bureau was mentioned),
  - Concurrence check with other agencies (5%).
- **Amount of Data (42%)**
  - Expand existing data (21%) (e.g., with more details on time/day of travel, vehicle class, County Profiles for the entire region),
  - Provide new data (21%) (e.g., longitude/latitude to facilitate map visualization by user, provide data at disaggregate geographic levels such as town, multiple zip code, crash data, airline boardings, Amtrak boardings).
- **Timeliness (32%)**
  - Keep at most 1-year lag between data and release times (specifically hub-bound travel and regional transportation statistical reports were mentioned).
- **Appropriateness of Format and Ease of Understanding (11%)**
  - Provide better metadata (5%) (specifically hub-bound reports were mentioned),
  - Provide better information on data sources (5%) (specifically hub-bound reports were mentioned),
  - Provide data in single MS Excel file formats for easier use, rather than PDF only (5%),
  - Enhance the data website by including more graphics (5%).
- **User Feedback (5%)**
  - Seek and implement suggestions for improvements from agencies (Mid-Hudson South Transportation Coordinating Committee was mentioned).

**FIGURE 1 NYMTC stakeholder survey results.**
2.3. Issues Identified Considering NYMTC’s Current Data Maintenance Process

The current practice of data maintenance and quality control within NYMTC is summarized in this section based on the *Data Group Methodology Report* received from NYMTC, as well as the information gathered from the NYMTC technical staff during various meetings.

NYMTC’s current data maintenance process is shown in the form of a flowchart in Figure 2 prepared by the authors of this paper for ease of understanding and also to facilitate comparison with the dynamic process that will be proposed later in Section 3 of this paper. In Figure 2, the components that are not mentioned in the Data Group Methodology report and thus that are not part of the officially acknowledged current practice are shown with dashed lines.

NYMTC’s current data maintenance and quality control process starts with ‘Creating the Data Product’. For that purpose, emails are sent to all parties that are required to provide NYMTC with variety of data. Contact information and website addresses for data provider agencies for different data subjects are specified in detail in the “Data Group Methodology Report” (11).

The next step is ‘Conducting the Reviews’. All the reports and brochures that are generated by NYMTC using the raw data provided by the source agencies go through an across-the-board review process. First, the NYMTC data group manager performs the initial internal review and his/her comments are applied. Then, a person from the Public Relations (PR) Department at NYMTC checks the data product for quality and political sensitivity. Next, internal review is carried out by the NYMTC staff and any comments that may arise are then applied appropriately. The next step is the external review of the data product by the corresponding external data source and their comments are also applied. Finally, the data product is reviewed one more time by the data group manager, as well as the PR person. Each review takes about 2 weeks, and the total review process takes about 2 months.

The next step is ‘Publishing the Data Product’. The corrected and completed data product is published on the NYMTC website and an article is written for the *NYMTC-Notes* newsletter to announce the availability of the new data product highlighting any interesting findings or changes.

The logical next step in the data maintenance process, based on the data life cycle concept suggested by Loshin (5), would be to monitor the effectiveness and popularity of the data product from the user’s point of view, and then to improve the overall process by taking into consideration user feedback. Even though these ‘Monitoring Effectiveness’ and ‘Improving the Process’ components are not mentioned in the Data Group Methodology Report, the fact that NYMTC conducted a website hosting evaluation in 2007, conducted the NYMTC Stakeholder Survey in 2007, and initiated the research project in 2008 (which this paper is based on) are clear indications of NYMTC's continuous efforts to monitor, modify and further improve the existing data products, thereby closing the loop in the data maintenance process in a way.
FIGURE 2 NYMTC’s current data maintenance procedure (Put in flowchart form by the authors to facilitate comparison with Figure 5. Dashed lines are not implemented parts of the current process).
3. PROPOSED SOLUTIONS

Specific solutions recommended to all of the abovementioned data, website, and data maintenance process oriented issues are summarized in this section. Data and website oriented solutions are mentioned only briefly here, to the extent necessary for completeness of this paper. Full details on data and website related solutions are presented in great detail in the final report of the project (10) which this paper is based on.

Data maintenance process oriented solutions, however, are presented in detail here, since they are the primary focus of this paper.

All of the recommendations aim to offer to the wide range of NYMTC website visitors and users better integrated, interactive, and easily accessible data that comes with sufficient metadata, is GIS-based and user friendly.

3.1. Data and Website Oriented Solutions

These solutions target the data products in terms of various data quality dimensions shown in Table 1, as well as target the website through which the data products are made available to the users and all the stakeholders. The recommended solutions make sure the data products are accurate, provide sufficient data coverage, easily accessible through an interactive and user-friendly website, up to date, and easy to understand with multiple format choices, metadata, and user support and feedback tools.

Figure 3 suggests a navigation path for data access through NYMTC home page, and Figures 4a and 4b show the current and suggested looks for a typical data product page (the ‘County Profiles’ data page is used as an example).

As indicated in Figure 3, it is recommended that the user gets directed to the webpage for a data product (similar to the one shown in Figure 4b), after selecting the ‘online’ option from the available formats in the Metadata. As can be seen in these figures, with the improved navigation path, it is intended that the user is able to access the data in an interactive manner by means of various filtering options at a desired level of detail and generate customized data in various output formats. Most of the users only need some specific data and they need to have an opportunity for processing the raw data to acquire only what they need.

As can be seen in Figure 4b, with the suggested new format for individual data products, before proceeding with viewing/customizing/downloading the data itself, the user can view the Glossary of terms and acronyms, and Metadata through corresponding links. Metadata, as shown in the lower part of Figure 3, should provide information related to the availability period, geographic coverage, data source along with a link to the website of original owner/collector of this data if NYMTC is not the collector/creator of this data, data collection/creation methodology, clear statement of the purpose for collecting this data, contact person’s email address, phone and fax numbers, and available formats (electronic formats as well as hardcopy. For a product that is available as a hardcopy, shelf location at NYMTC library along with a link to the library webpage for hours etc. should be given. For an electronic data product, a link to the data page should be given).

As it is shown in Figure 4b, the user can either view the standard PDF version of the data product, or can generate customized data tables with chart/graph options with a simple query utilizing drop-down menus as suggested in Figure 4b or with radio buttons for selecting different data options. It is also a good practice to denote commonly selected fields in the query. If the user wishes to add graphs/charts options to his/her data customization process by checking the box that says “check here to create associated graphs/charts”, then s/he is provided with a...
separate interactive window for selecting chart type, data series etc. For the selected data set, users should be allowed to generate basic statistics like count, sort or cross-tab to name a few. Data can be output in various formats such as HTML display with Print option, or XLS, TXT, CSV, PDF etc. exports, as well as DBF export whenever possible (it allows data mining and is preferable by most researchers).

![Diagram of data access navigation path]

**FIGURE 3 Proposed navigation path for data access.**
FIGURE 4 Current (a) and Proposed (b) looks for a data product page (County Profiles in this case) at NYMTC website.
3.2. Process Oriented Solutions
Given Loshin’s (5) argument regarding data lifecycle, a dynamic data maintenance procedure is proposed in this section, which might potentially be applied on a regular basis (bi-annually or whenever a new data is produced and required to be disseminated).

This proposed procedure is built around three main functional components (review, monitor/assess, and modify/improve). The suggested procedure also incorporates the elements from the lifecycle of data (create, distribute, access, update) as shown in Figure 5.

3.2.1. Create
The suggested process starts with creating the data product. A very important input and feature to this step is a series of one-day workshops recommended as part of this dynamic data maintenance framework for the purposes of letting data providing agencies get a better idea about NYMTC’s data needs and data protocol.

Many agencies contribute various data needed by NYMTC. The availability, quality, and timeliness of this data generally depend on establishing a long-term relationship with each agency. However, there is not one institutional approach that is applicable to every external agency. Since the overall quality of data is crucial to the success of NYMTC’s goals, it is important to establish long-term institutional procedures with all the contributing agencies. Moreover, it is also important to better explain and sometimes provide incentives to ensure that the data flow is consistent and satisfactory. Based on various meetings with NYMTC, the major time delay related to having latest data is caused by the delays in obtaining this data from individual agencies. There is little room in reducing time required to publish this data since NYMTC’s internal process is already quite efficient given the rigorous internal and external review requirements of the data to be published.

Thus, to determine novel ways to improve the efficiency of the data acquisition process from participating agencies, it is crucial to conduct a series of workshops (at least two) where all the contributing agencies are brought together to set up a more formal and reliable mechanism for NYMTC to work with these external agencies to obtain necessary data.

3.2.2. Review
The next step, namely the data product reviews, uses the findings from the data and website oriented solutions section when conducting the reviews and looking for any issues in terms of various dimensions of data quality such as coverage, appropriateness of format, availability and content of metadata, accuracy, integration, accessibility and so on as discussed in detail previously in Section 3.1 of this paper. If the data product presents conformity in terms of these various aspects of data quality, it will be distributed to the end-users. If not, it shall undergo necessary modifications. This part of the procedure should be repeated until desired conformity to the desired level of data quality is achieved.
FIGURE 5 Suggested dynamic data maintenance process.
3.2.3. Distribute, Monitor and Assess
Once the data product is distributed by publishing it on the NYMTC website and the users can access it, another functional component of the maintenance process, monitoring and assessment, will come into play. It will basically involve utilization of the following tools and methods in order to determine if any modifications should be applied to the existing data quality dimensions.

1. Online user feedback forms
2. Website traffic monitoring tools such as hit counters to keep track of successful data retrievals and returning users, the popularity of pages and data products, and common entry points
3. Conducting stakeholder surveys
4. Reviewing new studies on the latest technological trends and best practices
5. Use of online multi-user collaboration tools such as wiki (a World Wide Web application designed to allow multiple users to add, remove, and edit content), to create a common community environment for data providers as well as users
6. Implementing emerging web-based technologies that facilitate exchange of information while creating online communities beyond the tools described above. For example, elgg (6), an open source social networking tool is used to create professional and scientific communities such as SciSpace (7). These sites are easy to maintain and thrive on the contribution of its members. Elgg provides fine grained access control that can be easily customized to allow different contributors to have different access privileges. For example, some content that is published on the webpage can be accessible to everybody subscribing to this network while some other content such as agency specific data repositories can be made accessible to the corresponding agencies only.

3.2.4. Modify and Improve
If it is agreed by NYMTC that a number of modifications to the current data maintenance and quality control process are necessary, then, the third functional component, modification and improvement, comes into play and acts as a feedback mechanism and allows the incorporation of the results from the monitoring and assessment component back into the initial step of the process to guarantee the continuing success of the agency-wide data maintenance procedure.

4. EVALUATION OF IMPACTS
In an effort to make an overall evaluation of the potential impacts of the various solutions recommended in this paper, this section aims to answer the following two questions:

- What will be the overall value-added by implementing the recommended solutions? (pros)
- Is it worth doing? (cons)

Table 2 summarizes the anticipated pros and cons of the recommendations made for both data and website related issues, and data maintenance related issues.
# TABLE 2 Potential Pros and Cons of Recommendations

<table>
<thead>
<tr>
<th>Type of Recommendation</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
</table>
| Data and Website Oriented | +The time, money and effort spent cleaning poor quality data may be reduced significantly, thus increasing overall productivity  
+Data users get high quality, accurate, highly accessible, better integrated and customizable data in a timely manner  
+Increased stakeholder and data user satisfaction  
+Help NYMTC better project an image of quality to other agencies, stakeholders and data users  
+Promote website and other types of data user attraction and retention | -Long-term commitment of the time and effort of NYMTC staff and the consultant implementing the project  
-Implementing and maintaining some technology bound solutions requires expertise and money  
-Technology changes constantly |
| Data Maintenance Process Oriented | +Interagency workshops will reduce delays for NYMTC in getting data from external agencies and will provide good input  
+Improvement regarding the notion of quality data and its maintenance over time so that NYMTC staff, agencies providing data, all the stakeholders and data users view data quality as key to success and making valid decisions  
+Provide broader networking and connectivity and create an atmosphere of cooperation among NYMTC, agencies providing data, stakeholders and data users | -Workshops are costly and difficult to establish  
-Expenditures for the stakeholder surveys, and other monitoring and assessment tools |

The overall success and impact of the recommendations made can be assessed over time by calculating several quantifiable measures of performance such as, “Combined Customer Satisfaction Index (CCSI)” and “Importance Index (ImpI)” as given in the illustrative example below, by regularly conducting stakeholder surveys using the same sample of stakeholders. This way, a simple but informative chart as shown in Figure 6 can be obtained depicting the potential effectiveness and success of the suggestions made.
4.1. Illustrative Example—Evaluating the Impacts of Suggested Improvements

Using the results of the NYMTC stakeholder survey, some simple but informative data evaluations were made which involved the calculation of the following:

- Combined Customer Satisfaction Index (CCSI)
- Importance Index (ImpI)

Customer satisfaction indices are widely used in measuring customer satisfaction in a variety of fields both from public and private sector. For example, the American Customer Satisfaction Index (ACSI) reports scores on a 0-100 scale at the national level and produces indexes for numerous economic sectors, industries, companies, and federal or local government agencies using a set of causal equations that link customer expectations, perceived quality, and perceived value to customer satisfaction (8).

The Minnesota Department of Employment and Economic Development (DEED) primarily uses two indices to express customer satisfaction with its services; the Minnesota Customer Satisfaction Index (MnCSI) and ACSI. The MnCSI uses responses to three questions about satisfaction, answered on a scale of 1 to 10, where “1” is the lowest (least satisfied) and “10” is the highest (most satisfied). Then, a single number, which varies from 0 to 100 is generated using Equation 1 below (9).

$$MnCSI = \left( \frac{(Question1-1)}{9} \times 33.3 \right) + \left( \frac{(Question2-1)}{9} \times 33.3 \right) + \left( \frac{(Question3-1)}{9} \times 33.3 \right)$$

(1)

where

- Question1 is the user’s rating on a scale of 1 to 10 to Question 1
- Question2 is the user’s rating on a scale of 1 to 10 to Question 2
- Question3 is the user’s rating on a scale of 1 to 10 to Question 3

A simple generalized formula shown in Equation 2 can be written based on the idea behind Equation 1 in order to obtain a formula that can be used for calculating the Combined CSI (CCSI) that varies between 0 and 100 (0 representing lowest quality, 100 representing...
highest quality) for each of the NYMTC’s data products using the responses to Question 4 in the stakeholder survey, which asked:

“Do you have any suggestions to improve these products?”

After reviewing the survey responses to this open-ended question, the research team calculated the percentages of stakeholders who had concerns regarding the six criteria used (specifically, Accuracy, Consistency, Amount of Data, Timeliness, Appropriateness of Format and Ease of Understanding, and User Feedback) and then scaled those percentages on a 1-10 scale. After that, CCSI was computed simply using Equation 2.

\[
CCSI = \sum_{i=1}^{n} \left( \frac{(C_i - 1) \times 100}{9} \right) \frac{1}{n} \]

where \( C_i \): User’s rating to Criteria ‘i’ on a scale of 1 to 10
\( n \): Total number of criteria (6 in this case)

Table 3 below shows the results from the CCSI calculations. As can be seen, the customer satisfaction index for NYMTC’s data products was computed as about 80%, which indicates a relatively high level of overall quality. However, it should be noted that, considering that the survey had only 32 responses in total, the accuracy of this result should remain rather questionable. A better and far more accurate evaluation would be possible with a larger sample size.

**TABLE 3 Combined Customer Satisfaction Index (CCSI)**

<table>
<thead>
<tr>
<th>No</th>
<th>Criteria</th>
<th>User’s Rating*</th>
<th>Combined Customer Satisfaction Index (CCSI)**</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Accuracy</td>
<td>9.5</td>
<td>15.741</td>
</tr>
<tr>
<td>2</td>
<td>Consistency</td>
<td>8.9</td>
<td>14.630</td>
</tr>
<tr>
<td>3</td>
<td>Amount of Data</td>
<td>5.8</td>
<td>8.889</td>
</tr>
<tr>
<td>4</td>
<td>Timeliness</td>
<td>6.8</td>
<td>10.741</td>
</tr>
<tr>
<td>5</td>
<td>Appropriateness of Format and Ease of Understanding</td>
<td>8.9</td>
<td>14.630</td>
</tr>
<tr>
<td>6</td>
<td>User Feedback</td>
<td>9.5</td>
<td>15.741</td>
</tr>
</tbody>
</table>

CCSI= 80.370

*On a scale of 1 to 10 where 1 represents least satisfied and 10 represents most satisfied
**On a scale of 0 to 100 where 100 represents least satisfied and 100 represents most satisfied

Next, using the responses to Question 3 in the stakeholder survey, a simple Importance Index (ImpI) that varies between 0 and 1 (0 representing low importance, 1 representing high importance) for each of the NYMTC’s data products was calculated by weighing the responses as shown in Equation 3. Question 3 asked:

“Please check one of the three options for each product whether you have used the product or not”

- Important
- Not important
Not familiar with

\[
\text{ImpI} = \frac{0 \times (\% \text{ responding "not familiar with"}) + 1 \times (\% \text{ responding "not important"}) + 2 \times (\% \text{ responding "important"})}{200}
\]

where ‘0’, ‘1’, and ‘2’ are the numerical values assigned to the three degrees of importance and ‘200’ is the value that normalizes the ImpI value to range between 0-1.

The results for ImpI calculations for each of the NYMTC data products are presented in Table 4. As can be seen, survey respondents rated the Regional Transportation at a Glance brochure, Travel Patterns report, and the Regional Transportation Statistical report as the three most important data products. Average ImpI for all eleven of the data products is 0.685, indicating that the stakeholders value the data products with a relatively high level of importance overall.

Again, it should be noted that, considering that the survey had only 32 responses in total, the accuracy of this result should remain rather questionable. A better and far more accurate evaluation would be possible with a larger sample size.

### Table 4 Importance Index (ImpI)

<table>
<thead>
<tr>
<th>Data Product</th>
<th># Respondents</th>
<th>% Respondents</th>
<th>Importance Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Saying Important</td>
<td>Saying Not Important</td>
<td>Who are Unfamiliar</td>
</tr>
<tr>
<td>Reg. Transp. at a Glance</td>
<td>22</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Travel Patterns</td>
<td>22</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Reg. Transp. Stat.</td>
<td>21</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Truck Toll Volumes</td>
<td>18</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Hub-Bound Travel</td>
<td>18</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Hub-Bound Travel at a Glance</td>
<td>17</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Truck Toll Volume Trends at a Glance</td>
<td>16</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>County Profiles</td>
<td>16</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Reg. Demographics</td>
<td>17</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Transp. Safety Stat.</td>
<td>16</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Transp. Safety at a Glance</td>
<td>11</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Avg.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 5. SUMMARY

The New York Metropolitan Transportation Council (NYMTC) has recently awarded a research project titled “Improvements on NYMTC Data Products” to the authors of this paper in an effort to modernize the existing data products and improve the communication between the agency and the public. The main goal of the research project was to perform a through review and examination of NYMTC data products to identify specific issues about the existing data products, website and the current data maintenance process at NYMTC, and then suggest appropriate solutions, both data and website, and data maintenance process oriented.
This paper reported the results from this research project by giving special emphasis to the issues and solutions related to the dynamic data maintenance process. This project was finalized in November 2009 and the final report is available online (10).

A summary of the data, website, and data maintenance process oriented issues, as well as specific solutions to these issues were presented in this paper. Various dimensions of data quality and data life cycle were taken into consideration. Suggestions about possible ways to evaluate the potential impacts from the research described in this paper were also given.

It should be noted that, while the majority of the recommendations presented in this paper are specific to NYMTC’s needs, they can be adopted and implemented by other agencies to address similar issues regarding data quality and maintenance procedures.

ACKNOWLEDGEMENTS
The authors would like to acknowledge the significant contributions made by the NYMTC staff to the research project summarized in this paper. In every task of the project, NYMTC research project manager Mr. Salvatore Mamone, director of technical services Mr. Kuo-Ann Chiao, and from the NYMTC’s technical group Mr. Munneesh Patel and Ms. Leokadia Glogowski provided the authors with timely advice, and sharp, insightful and constructive criticism that significantly enhanced the quality of the final product.

The research project summarized in this paper was funded by a joint grant from and administered through NYMTC and the Region 2 University Transportation Research Center. Their support is both acknowledged and appreciated.

REFERENCES
(11)New York Metropolitan Transportation Council (NYMTC). Data Group Cookbook for Reports and Brochures.