DATA MODELING, A MEANS TO QUALITY IN A REGIONAL TRANSIT DATA REPOSITORY

Paula Okunieff and Manny Insignares
Consensus Systems Technologies (*ConSysTec*) Corp.
PO Box 517, 17 Miller Ave.
Shenorock, NY 10587-0517 USA
+1-914-248-8466, paula.okunieff@consystec.com

This paper discusses the approach to data modeling and data quality assessment/control encountered in the New York State Department of Transportation (NYSDOT) *Transit Schedule Data Exchange Architecture* (TSDEA) project. In particular, the project developed a reference data model and mechanism to ensure data quality transparency for appropriate regional data use by downstream applications.

Key Words: data quality, systems engineering, transit

BACKGROUND

The New York Metropolitan Area TSDEA and supporting *Schedule Data Profile* (SDP), managed by the NYSDOT, provides an efficient, standards-based, framework for managing and exchanging schedule data among agencies and effectively communicating schedule information from multiple NY State transit providers to the public. The effort is focused on collaboratively defining a framework, as well as tools for data development, conversion and exchange, to support regional multi-agency initiatives that use schedule data, including the interactive on-line regional multimodal traveler trip planner *TRIPS123*.

Conceptual Data Reference Model

The purpose of a *Conceptual Data Reference Model* (CDRM) is to describe the “real-world domain” using an unambiguously defined set of data concepts, and model their relationship to each other. The SDP’s CDRM is an abstract model that describes the real world domain of transit scheduling. The technical approach or methodology used included: entities, relationships and attributes, which are captured in a model that is independent of technology implementation.

The CDRM was developed to help ensure that the needs of the applications that use schedule data are met by the SDP. It identified key requirements related to the data validity including: identity, uniqueness, references, attributes, and data format and type. These requirements were captured in the SDP Functional Requirements and modeled in the SDP CDRM. As a result, the model helps ensures that data, when exchanged, are consistent and well understood across applications.

The SDP’s CDRM accomplishes the following:

- Provides a reference for the meaning and relationships of the real-world domain concepts when they are implemented in applications and interfaces;
- Defines the identifiers and uniqueness requirements required for downstream applications;

- Provides a description of core attributes that are needed by most downstream applications;
- Incorporates flexibility to constrain or extend the model given various implementation approaches and tools.

SDP CDRM Development – Concept to Design

The SDP Project used a system engineering approach for developing user driven requirements. The initial stage of the process involved soliciting stakeholder input on how key user groups currently use and might in the future use schedule and related information. This Concept of Operations (ConOps) was developed through stakeholder meetings and interviews. A set of downstream application operational descriptions, high level requirements, and detailed data requirements were documented in a series of white papers. The white papers (also called *Use Cases*) included downstream applications such as Integrated Trip Planning, Dynamic Generation and Presentation of Public Timetables, and Generation of Ad Hoc Scheduling.

The High Level requirements from the Use Cases provided the initial input into the scope and requirements for the conceptual reference model. A draft CDRM was developed, and through a series of stakeholder meetings, the model was refined in order to better capture user upstream practices and downstream application needs. The product from this effort was a comprehensive *Functional Requirements* document. The requirements document not only covered the schedule and related data concepts found in the reference model, but it also described a preliminary set of requirements for naming and formatting an exchange standard based on the XML Schema standard, as well as integration issues for a central repository that fuses individual agency data into a regionally related data set.

SDP CDRM - Implementation

The CDRM can guide more than one way to implement the sharing of transit schedule data. The CDRM is meant to be used as a framework to unambiguously describe the SDP data concepts and their relationship to each other. Different technical methods may be used to implement the SDP CDRM; that is, physically represent and store schedule data. The three methods include: logical data model, physical database, and XML Schema.

DATA QUALITY CHECKS

Quality checks are necessary to ensure that the data provided by the TSDEA is correct and complete. Several techniques will be used to test the quality of data. The CDRM and the set of business rules developed in the requirements phase of the project will be used to generate a comprehensive list of checks and tests to ensure the SDP data integrity, since they unambiguously describe the semantics and relationships among the schedule data concepts.

Regional consistency tests will be based on additional rules and requirements to be developed for regional data sets, such as agreement on regional naming conventions. Those decisions will be approved at a later time, and are not part of the initial project demonstration.
For regional sharing of schedule data to be successful and valuable to the transportation agencies in the Downstate New York region, the schedule data needs to be complete and accurate.

The SDP Functional Requirements document specifies that the quality control and integrity checking of SDP files will go through a minimum of three general levels of data quality checking. At each of the three levels, the submitted file will undergo a set of quality checks and tests. The three levels are described as:

- Level 1: Registration – Ensures that the file contains a well formed and complete SDP XML document.
- Level 2: Authorization – The file content has passed quality checks that are based on business rules and requirements. The file content is deemed logically consistent (semantically and logically accurate).
- Level 3: Regionally Consistent – File content has passed tests to ensure consistency with regional naming conventions and representations.

CONCLUSION

The TSDEA developers used a systems engineering process to develop use cases and requirements for a Schedule Data Profile, a XML schema that serves as a transfer format to share schedule and related data among operators in the NY metro region with a regional transit data repository.

A conceptual data reference model was developed to qualify the structure and verify data relationships and definitions (e.g., data values). Based on the data model and requirements a set of data integrity checks was developed to validate and verify the quality of data to be used in regional transit applications.