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

Presented at:

15th World Congress, New York, NY

TS18: Data Collection Archiving and Processing

Monday, 3:30pm - 5:00pm

by

Paula Okunieff, Manny Insignares

ConSysTec

Case Study on Data Quality

- Example from *NYSDOT Transit Schedule Data Exchange Architecture* (TSDEA) Project

- Developed ***Schedule Data Profile*** (SDP) to ensure consistent transit data quality for use in regional transportation systems.

ConSysTec

What is Data Quality?

- "*Fitness-for-use* of a particular piece of data for your application"
  - From Transit Location Referencing Guidebook

- What if the data will be used for multiple downstream applications and services?

ConSysTec

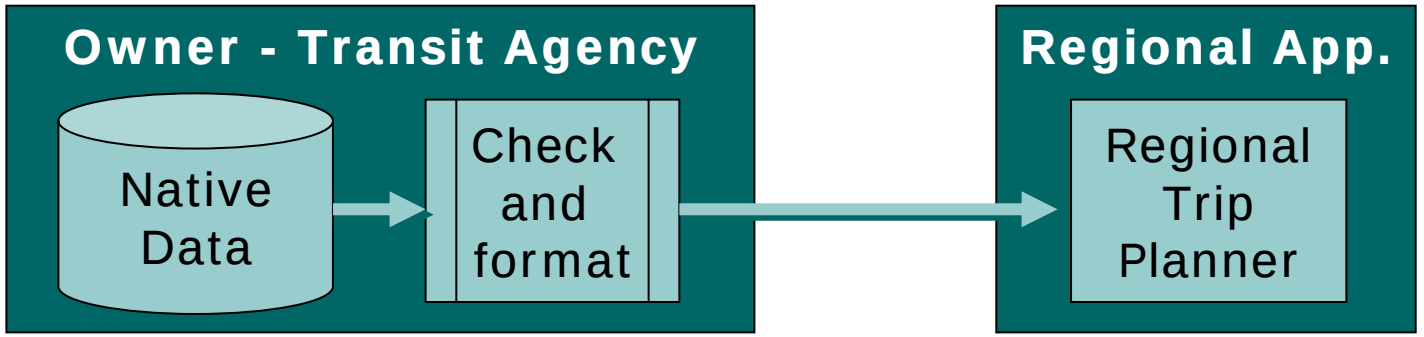
How is Quality Achieved?

- Quality Checking
- Role of Custodian and Users
- Requirement description of downstream uses

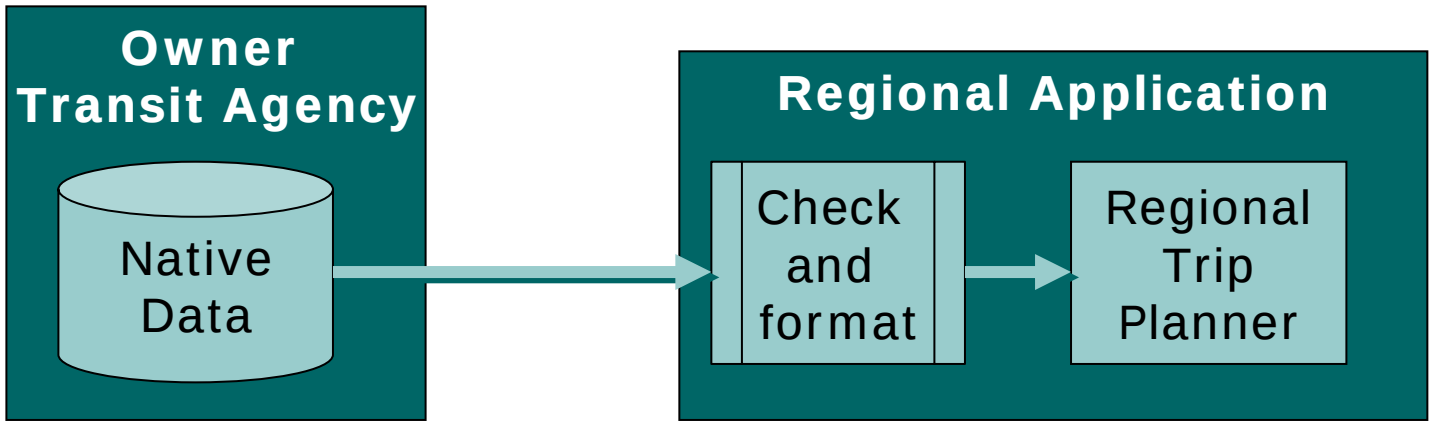
ConSysTec

Data Quality “Owner”

Option 1



Option 2

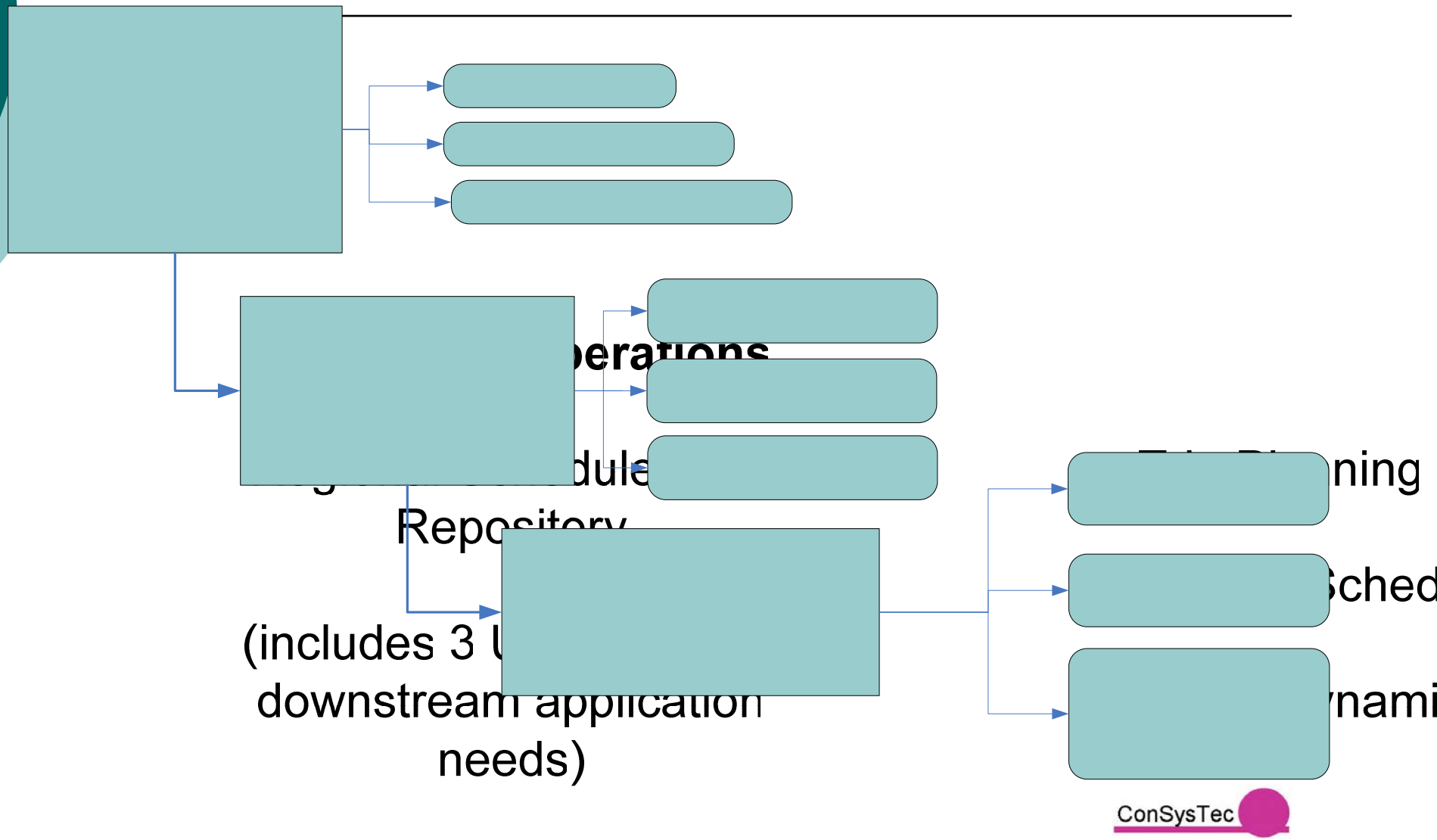


ConSysTec

Quality Elements

- Semantics
  - Relationship among concepts
  - Business Rules
- Syntax/format
- Accuracy, Lineage, Currentness, Logically Consistent, Complete

Focus on Describing, Verifying and Validating Data Requirements



Functional Requirements

Concept to Design

- Semantics: “data concept meaning”
  - Unambiguously define set of data concepts,
  - Model relationship of concepts,
  - Describe business rules for using data.
- Format Requirements:
  - Relational database
  - Exchange while minimizing quality checking
  - Facilitate exchange of large files

ConSysTec

Design to Implementation

- Various Implementation Methods
  - Physical Relational Database
  - XML Schema / Document
  - Comma Delimited (CSV files)
- Quality checks
  - Levels of Quality Checks
  - Syntax
  - Referential Integrity (uniqueness & logically consistent)
  - Dates
  - Other business rules related to data concepts (e.g., day type, pattern, facility, location)

TSDEA Levels of Quality Checking

- 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.

ConSysTec

Project Results

- Demonstration completed
  - Used for NY region's TRIPS123 trip planner
  - Testing use on up-state agency schedule data
- Migrating to Operations
- Plans to develop NY Regional Transit Data Portal (TSIP)
  - using Schedule Data Profile implementation methods (SDP-XML; SDP-csv formats)

ConSysTec