Data Profiling,
Data Governance, and
Data Quality for
Master Data Management

David Loshin
Knowledge Integrity, Inc.
www.knowledge-integrity.com
Agenda

- We will explore how data profiling, data quality, and data governance will support the emergence of the core master data program:
  - Master data management basics
  - Master object identification and modeling
  - Data consolidation and integration
  - Operational master service needs
  - Data governance brings it all under control
  - Review of Data Quality technology for MDM
What is “Master Data”?  

- Core business objects used in the different applications across the organization, along with their associated metadata, attributes, definitions, roles, connections, and taxonomies, e.g.:
  - Customers
  - Suppliers
  - Parts
  - Products
  - Locations
  - Contact mechanisms

“David Loshin purchased seat 15B on US Airways flight 238 from Baltimore (BWI) to San Francisco (SFO) on July 20, 2006.”

<table>
<thead>
<tr>
<th>Master Data Object</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer</td>
<td>David Loshin</td>
</tr>
<tr>
<td>Product</td>
<td>Seat 15B</td>
</tr>
<tr>
<td>Flight</td>
<td>238</td>
</tr>
<tr>
<td>Location</td>
<td>BWI</td>
</tr>
<tr>
<td>Location</td>
<td>SFO</td>
</tr>
</tbody>
</table>
What is Master Data Management?

Master Data Management (MDM) incorporates the business applications, information management methods, and data management tools to implement the policies, procedures, infrastructure that support the capture, integration, and subsequent shared use of accurate, timely, consistent and complete master data.

- Governance
  - Policies
  - Procedures
  - Infrastructure
- Action
  - Capture
  - Integration
  - Sharing
- Quality
  - Accuracy
  - Timeliness
  - Completeness
Challenges to MDM Success

<table>
<thead>
<tr>
<th>Organizational</th>
<th>Operational</th>
<th>Technical</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Uncoordinated Enterprise</td>
<td>• Data Quality</td>
<td>• MDM Architecture</td>
</tr>
<tr>
<td>• Data Ownership</td>
<td>• Data Standards</td>
<td>• Transition Plan</td>
</tr>
<tr>
<td>• Performance Management</td>
<td>• Data Governance</td>
<td>• System Performance</td>
</tr>
</tbody>
</table>

Understanding and addressing these issues provides grounding for the core MDM operational components.
The MDM Component Layer Model

- Business Process Integration
- Business Rules
- MDM Business Component Layer

- Application Integration and Synchronization Service Layer
- MDM Component Service Layer

- Identity Search and Resolution
  - Record Linkage
  - Merging and Consolidation

- Migration Plan
  - Hierarchy Management
  - Identity Management
  - Administration/Configuration

- Data Standards
- Metadata Management
- Data Quality
- Data Stewardship

- MDM Service Layer Architecture
  - Master Data Model
  - MDM System Architecture

- Business Process Management
- Integration
- Identification
- Management
- Governance
- Architecture
Architecture

- Master data model
- MDM system framework
- Service layer architecture
Central Master/Coexistence
Registry
Transaction Hub
MDM Architectural Styles

- Different underlying architectural approaches influence ways of integrating data management services:
  - Data assessment and master object identification
  - Underlying master data model
  - Data profiling
  - Parsing, Standardization, Normalization
  - Cleansing
  - Enhancement

- Two high-level approaches:
  - Repository
  - Registry
Repository – Initial Data Consolidation

1. Data sources are staged and cleansed
2. Cleansed data is brought into a master staging area for consolidation
3. Records are checked for previous registration with the unique identification services
4. Survivorship rules are applied
5. For new records, a unique identifier is created
6. The resulting record is enhanced
7. The enhanced record is stored in the master repository
8. The unique identification service is notified
Registry – Initial Data Consolidation

1. Source data is cleansed and enhanced
2. The unique identification service is consulted with the identifying information to determine if the entity has already been registered, and if not, to generate a unique identifier for registration
3. The records are managed within their respective data locations, and
4. The records are tagged with the returned unique identifier
5. The mapping from the registry to the data locations is registered with the associated identifying information
Master Data Model

- Limited universe of common master objects
  - *Party, customer, product, part, supplier, claim, instrument*
- Universal models may be suitable as starting points
- Challenges:
  - Resolution of metadata in a consistent manner
  - Creating a model that accommodates all applications properly

| CUST | | | | Nightingale-Patterson |
|------|-----------------|-----------------|-----------------|
| First | VARCHAR(15) | | | |
| Middle | VARCHAR(15) | | | |
| Last | VARCHAR(21) | | | |
| Address1 | VARCHAR(45) | | | |
| Address2 | VARCHAR(45) | | | |
| City | VARCHAR(30) | | | |
| State | CHAR(2) | | | |
| ZIP | CHAR(9) | | | |

| CUSTOMER | | | | Nightingale-Patterson |
|----------|-----------------|-----------------|-----------------|
| FirstName | VARCHAR(14) | | | |
| MiddleName | VARCHAR(14) | | | |
| LastName | VARCHAR(20) | | | |
| TelNum | NUMERIC(10) | | | |
Master Object Identification and Modeling

- Document business processes
- Segregate services from dependent data
- Identify target data areas
- Inventory and identify candidate master entities
- Source data analysis
- Assembling the Target Model
Business Process Modeling

- A “business process” is a coordinated set of activities intended to achieve a desired goal or produce a desired output product.
- Models are designed to capture both the high level and detail of the business process.

![Diagram of a business process model with inputs, activity, controls, actors, and output.]
Sharing Enterprise Data Objects

- Interactions between activities depend on shared data:

- Instance values of common data types representing business facts communicate input and control during the business process.
Segregating Master Data Services

- Evaluating the catalog of data elements through which business processes are coordinated yields a candidate list of master data objects.
- Assembling that catalog enables a review of the business application services.

Identify application services at the business level as they access master data objects.
Finding Master Data

- What data elements constitute our “master data”? 
- How do we locate and isolate master data objects that exist within the enterprise? 
- How do we assess the variances between the different representations in order to consolidate instances into a single view?
Centralizing Semantics – Basic Questions

- Master data will be distributed and replicated across the application environments
- The initial goal is to consolidate the replicated copies into a single repository
- Before materializing a single master record for any entity, one must be able to:
  - Discover which data resources may contain entity information
  - Understand which attributes carry identifying information
  - Extract identifying information from the data resource
  - Transform the identifying information into a standardized or canonical form
  - Establish similarity to other standardized records
Structures and Semantics Vary Across the Enterprise

CUSTOMER

- FirstName: VARCHAR(14)
- MiddleName: VARCHAR(14)
- LastName: VARCHAR(30)
- TelNum: NUMERIC(10)

CUST

- First: VARCHAR(15)
- Middle: VARCHAR(15)
- Last: VARCHAR(40)
- Address1: VARCHAR(45)
- Address2: VARCHAR(45)
- City: VARCHAR(30)
- State: CHAR(2)
- ZIP: CHAR(9)
"A customer is an individual who has purchased one of our products"

"A customer is an individual to whom we have delivered one of our products"
Master Object Resolution

- Resolution of candidate master data types requires a complete view of what composes the information architecture
- This entails cataloging data sets, their attributes, data domains, definitions, contexts, and semantics
- This view must facilitate the resolution of:
  - *Format* at the element level,
  - *Structure* at the instance level, and
  - *Semantics* across all levels
- This introduces three challenges:
  - Collecting and analyzing master metadata
  - Resolving similarity in structure
  - Understanding and unifying master data semantics
Collecting and Analyzing Master Metadata

- Metadata sources:
  - Data dictionaries
  - E/R models
  - COBOL copybooks
  - Subject matter experts

- Document all data element characteristics within a metadata repository using a standard representation

- The consolidated metadata repository will enumerate data characteristics in a standardized way

- The standard representation enables statistical analyses such as:
  - Assessing frequency of occurrence of names
  - Comparing the lengths of name fields
  - Discovering dependencies between attribute names and assigned types
Resolving Similarity in Structure

- Different underlying master meta-models are likely to share many commonalities
- Structures will probably reflect similar collections of attributes and relations
  - Example: Many customer data sets will contain customer names along with variant contact mechanisms such as addresses or telephone numbers
- The data analyst can review existing models to identify similar object structures
- Two “hints” in structural similarity:
  - Overlapping structures may reflect data sets carrying identifying information with variant sets of attributes
  - Derived structures potentially reflect an embedding of core master data attributes within a specialized version of the same object
Unifying Master Data Semantics

- What is the qualitative difference between pure syntactic/structural metadata and its underlying meaning?
- Review semantic consistency in data element naming and data types, sizes, and structures
- Next, document business meanings:
  - What are the data element definitions?
  - Are there authoritative sources for these definitions?
  - Do similar objects have different business meanings?
- Resolution of variance will help in standardizing both the representations and meanings for identified master data object types
Proposing the Target Model

- Evaluate the catalog of identified data elements
  - Seek out the frequently created, referenced, modified, retired
- Assess object organizational structure
  - Evaluate conceptual structures as they map to business process use
  - Example: locations are composed of street, city, state, ZIP code
- Identify and resolve anomalies across data element sizes, types, formats
- Propose an object model
- Validate the object model within the information framework
- Validate the object model within the application framework
Data Profiling for Master Data Analysis

- Characteristics of data element metadata is critical for analysis
- Combination of artifact review and empirical analysis
- Data profiling can provide “ground-truth” evidence of consistency with metadata
The Real Data Quality Problem

- There are no objective measures of data quality
- The only industry metrics are based on name deduplication or address standardization
- Data cleansing does not address long term quality improvement
- The answer: Use data profiling to analyze the current state of data and correlate anomalies to data consumer expectations
The Assessment Process

- Data set selection, profiling and profile review
- Impact Analysis
- Rule Definition, Review, and Refinement
- Obtain Objective Measurements
What Can Go Wrong with Data?

- Data entry errors
- Data conversion errors
- Unexpected values
- Mismatched syntax, formats and structures
- Inconsistent values
- Missing values
Reasons for Data Degradation

- Limited tracking of documentation with system
- Exploitation of “shortcuts” for implementation
- Multiple front end interfaces to same back end
- Limited system functionality driving “creativity”
- *Changes in use and perception of data*
Forensics and Archeology

- Using the proper analysis, profiling reveals errors in the data, as well as insight into their origin
- Naïve analysts use profiling to find and correct erroneous data
- Savvy analysts use profiling to find and correct erroneous processes
Hidden Secrets

- Default null values
- Programming errors
- Data element domains
- Format Constraints
- Abstract Data Types
- Augmented Domains
- Overloaded attribute use
- Orphans
- Unused attributes
Default Null Values

- **Social Security Numbers**
  - ‘000000000’, ‘999999999’

- **Names**

- **Phone Numbers**
  - ‘No phone number provided’
  - 000-000-0000
  - 999-999-9999
Programming Errors

- In one application, we had been assured that all records were evaluated prior to insertion into the database to ensure uniqueness.
- Profiling results showed that numerous duplicates existed.
- After bringing this to the attention of the programmers, they acknowledged that a programming error existed and was to be fixed in the next release.
Data Element Domains

- An evaluation of a “Sex” column revealed three distinct values: “M”, “F”, and “U”
Format Constraints

Pattern analysis of a telephone number field showed that there were telephone numbers that contained both numerals and alphabetic letters:

- 1-800-MATTRES
Abstract Data Types

- In legacy systems, pattern analysis can reveal embedded use of abstract data types such as dates, SSNs, telephone numbers, etc.
Augmented Domains

- A set of manufactured values are added to a known domain to account for alternate use of the attribute
  - Additional “local codes” added to FIPS county codes to account for more than one location per county
Overloaded Attribute Use

- In one table a column named FOREIGN_COUNTRY contained telephone numbers and email addresses
Orphans

- In a review of a legacy foreign key relationship for case management, it was determined that a relatively large percentage of records in one table did not have a corresponding record in the other table.
Unused Attributes

- In one data quality assessment, we found one set of columns that were null a significant part of the time, and another that always had the same value.
- These results indicate that the attribute is not actually being used.
More Hidden Secrets

- Metadata
  - Data types
  - Attribute sizes
  - Ranges
  - Value set cardinality

- Structure
  - Embedded tables
  - Relational structure
  - Relationship cardinality
  - Key relationships
  - Master reference data

- Business Rules
  - Domain constraints
  - Dependency constraints
  - Derivation rules
  - Completeness rules
  - Consistency rules
Summary of Data Profiling

- Profiling used to:
  - Provide empirical analysis of data value sets
  - Assess existing metadata against published descriptions
  - Assess quality of each data set
  - Identify outliers for further review
  - Characterize feasibility of consolidation
Identification

- Every object subject to “mastering” is managed using a unique representation within the master repository.
- Any time data intended to refer to that object is seen by an application, its unique representation must be found, verified, and presented back to the application by the MDM platform.

<table>
<thead>
<tr>
<th>Identity Search and Resolution</th>
<th>Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Record Linkage</td>
<td>Merging and Consolidation</td>
</tr>
</tbody>
</table>
Identification

- Unique Identification
- Parsing and Standardization
- Normalization
- Record Linkage
- Survival Strategies
Example – Unique Identification

<table>
<thead>
<tr>
<th>Howard David Loshin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Howard Loshin</td>
</tr>
<tr>
<td>David Loshin</td>
</tr>
<tr>
<td>David Howard Loshin</td>
</tr>
<tr>
<td>H David Loshin</td>
</tr>
<tr>
<td>David H Loshin</td>
</tr>
<tr>
<td>David Loshing</td>
</tr>
<tr>
<td>David Losrin</td>
</tr>
<tr>
<td>David Lotion</td>
</tr>
<tr>
<td>David Loskin</td>
</tr>
<tr>
<td>David Lashin</td>
</tr>
<tr>
<td>David Lasrin</td>
</tr>
<tr>
<td>David Laskin</td>
</tr>
<tr>
<td>Mr. David Loshin</td>
</tr>
<tr>
<td>Loshin, Howard</td>
</tr>
<tr>
<td>Loshin David</td>
</tr>
<tr>
<td>D Loshin</td>
</tr>
<tr>
<td>Jill and David Loshin</td>
</tr>
<tr>
<td>Mr. Loshin</td>
</tr>
<tr>
<td>HD Loshin</td>
</tr>
<tr>
<td>The Loshin Family</td>
</tr>
</tbody>
</table>

Enterprise Knowledge Management: The Data Quality Approach (The Morgan Kaufmann Series in Data Management Systems) (Paperback) by David Loshin


The Geometrical Optics Workbook (Paperback - June 1991) by David S Loshin
Identity Search and Resolution

Objective: Provide the services that will seek the matching record in the master index that represents the “queried” object
Record Linkage – Parsing and Standardization

- **Parsing**
  - Identifying and tagging pieces of each data value within a semantic context

- **Standardization**
  - Correcting terms based on defined rules
  - Assembling components into recognized patterns

- **Normalization**
  - Adjust data elements into a common representation

- **Transformation**
  - Rule-based modifications into target canonical representations
  - Transformation into target format

<table>
<thead>
<tr>
<th>First</th>
<th>Howard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle</td>
<td>D</td>
</tr>
<tr>
<td>Last</td>
<td>Loshin</td>
</tr>
</tbody>
</table>

Loshin, Howard D → Howard D Loshin
Record Linkage – Similarity Scoring

- Records are deemed to be the same when their identifying information matches within a reasonable degree of similarity.
- Specific bits and pieces of identifying information are necessary to enable distinction between unique objects.
- Identifying values are weighted as to their contribution to the similarity score.
- The analysts must perform a continual assessment of similarity criteria to inoculate against false positives.
Merging and Consolidation

- Static merging and consolidation takes existing data sets, seeks to segment records into equivalence classes representing the same entity, and merges data from matched records into a single master record
  - Depends on automation
  - Thresholds for matching set ahead of time
  - Records between thresholds pulled for administrative review

- Inline merging and consolidation matches new instances against existing master records and updates the master records with validated current changes
  - Integrates rules with probabilistic algorithms
Survivorship

- Rules govern the determination of surviving data element values
  - Quality of data source
  - Currency of data source
  - Priority

<table>
<thead>
<tr>
<th>David</th>
<th>Lotion</th>
<th>1163</th>
<th>Kersey Rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Howard</td>
<td>David</td>
<td>Loshin</td>
<td>301-754-6350</td>
</tr>
</tbody>
</table>

| Howard | David | Loshin | 301-754-6350 | 1163 | Kersey Rd |
Identification – Summary

- Values are parsed into recognizable components
- Components are standardized
- Identifying information is collected for search
- Matching is facilitated via similarity scoring
- Matched records are consolidated during integration
Operational Master Data Services

- MDM operations require ongoing support for a number of data services:
  - Inline classification and matching
  - Data cleansing
  - Data enhancement
  - Unique identification
Inline Classification

- Master objects are grouped by their characteristics into hierarchies
- Each presented object is subjected to a classification service to assist in narrowing scope for matching

Diagram:

- Input: $\frac{1}{4}$" metal I thr phil
- Process: Classification Service
- Output: "screw"
Data Standardization/Cleansing

- Knowledge base is used to transform into a standardized format
Inline Matching

- Objects are submitted to the matching service to determine if an instance already exists in the master repository
- Relies on traditional parsing/standardization/linkage
Unique Identification

For unmatched records:
- Identifying information is extracted
- New unique identifier is created by UID service
- Record is stored to master repository
- Master index is updated
- Updates are communicated
Data Enhancement

- Reference data sets are consulted inline for value-added data append
- Example:
  - Customer data is enhanced with geo-demographics
  - Vendor data is enhanced with reliability score
  - Product data is enhanced with tolerance metrics
Governance

- Data Standards
- Data Quality Management
- Data Stewardship
Aligning Data Objectives and Business Strategy

- Clarify and understand the existing “Information Architecture”
- Create an inventory of data assets
  - Applications, data assets, documentation, metadata, usage
  - Inventory of data elements and “owning” application
In any enterprise, we can be confident that information exchanges are understood the same way through the definition and use of data standards.
Standards: Challenges for Critical Data Elements

- Absence of clarity
  - ...makes it difficult to determine semantics
- Ambiguity in definition
  - ...introduces conflict into the process
- Lack of Precision
  - ...leads to inconsistency in representation and reporting
- Variant source systems and frameworks
  - ...encourage “turf-oriented” biases
- Flexibility of data motion mechanisms
  - ...leads to multitude of approaches for data movement
Data Quality Management Goals

- Evaluate **business impact** of poor data quality and develop ROI models for Data Quality activities
- Document the **information architecture** showing data models, metadata, information usage, and information flow throughout enterprise
- Identify, document, and validate **Data Quality expectations**
- **Educate** your staff in ways to integrate Data Quality as an integral component of system development lifecycle
- **Governance framework** for Data Quality event tracking and ongoing Data Quality measurement, monitoring, and reporting of compliance with customer expectations
- Consolidate current and planned **Data Quality guidelines**, policies, and activities
Engineering Data Quality into the System

Analyze/profile data
Assess data quality dimensions
Data quality, Validity, & Transformation rules
Create monitoring system
Recommend data transformations

Flat File
RDBMS
IMS
VSAM
Application

Improved enterprise data quality

Generate data quality reports
Send data quality reports to data owners
Operationalizing Data Governance

- Actualization of data governance activities enables:
  - The identification of explicit and hidden risks associated with data expectations
  - The actualization of the implementation of business policy
  - Oversight of the definition of critical data elements
  - Monitoring and auditing information quality rule compliance
  - Managing enterprise data ownership and stewardship
  - Coordination and oversight of enterprise data quality

In general, data governance provides management oversight for organizational observance of different kinds of information policies
Roles and Responsibilities

- **Executive Sponsorship**
  - Provide senior management support at the C-level, warrants the enterprise adoption of measurably high quality data, and negotiates quality SLAs with external data suppliers.

- **Data Governance Oversight**
  - Strategic committee composed of business clients to oversee the governance program, ensure that governance priorities are set and abided by, delineates data accountability.

- **Data Steering Committee**
  - Tactical team tasked with ensuring that data activities have defined metrics and acceptance thresholds for quality meeting business client expectations, manages governance across lines of business, sets priorities for LOBs and communicates opportunities to the Governance Oversight committee.

- **LOB Data Governance**
  - Data governance structure at the line of business level, defines data quality criteria for LOB applications, delineates stewardship roles, reports activities and issues to Data Coordination Council.
Stewardship: Manual Intervention

- Issues with addressing data quality events:
  - Immediate remediation of flawed data – does this imply data correction?
  - Not all data flaws can be captured via automated processes – this implies manual reviews

- Accuracy may only be measured by comparing values directly

- Carefully integrate manual intervention when necessary in a controlled manner
Summary - Governance

- One of critical success factors for MDM deployment
- Benefits all stakeholders
- Establishes link between business objectives and information quality
MDM Functional Services

- Browsing & Editing
- Rules Manager
- Workflow Manager
- Propagation Synchronization
- Modeling
- Match/Merge
- Identity Management
- Hierarchy Management
- Integration
- Data Quality
- Metadata Management
- Audit
- Security Management
- Authentication
- Access Control
- Create
- Read
- Modify
- Retire
Example: Generic Master Data Services

- "Object Locate"
- "Object Factory"

Master Index
Data Quality Technologies for MDM

- Data parsing and standardization
- Record Linkage/Matching
- Data scrubbing/cleansing
- Data enhancement
- Data profiling
- Data auditing/monitoring
Data Parsing and Standardization

- Defined patterns fed into “rules engine” used for:
  - Distinguishing between valid and invalid strings
  - Triggering actions when invalid strings are recognized
Parsing and Standardization Example

(301) 754-6350

Area Code 301
Exchange 754
Line 6350

(999)999-9999
(999) 999-9999
999-999-9999
999.999.9999
1-(999)999-9999
1 (999) 999-9999
1-999-999-9999
1 999.999.9999
Parsing and Standardization - Approach

- Simplest approach: standard regular expression or context-free parsing with actions attached to success states
- More complex:
  - Standard parsing with variant lookahead
  - Context-sensitive parsing
  - Table-driven string/pattern matching
Data Scrubbing/Cleansing

- Identify and correct flawed data
  - Data imputation
  - Address correction
  - Elimination of extraneous data
  - Duplicate elimination
  - Pattern-based transformations

- Complements (and relies on) parsing and standardization
Cleansing Example
Matching/Record Linkage

- Identity recognition and harmonization
- Approaches used to evaluate “similarity” of records
- Use in:
  - Duplicate analysis and elimination
  - Merge/Purge
  - Householding
  - Data Enhancement
  - Data Cleansing
  - Customer Data Integration
Matching/Record Linkage - Example

Knowledge Integrity, Inc.  301-754-6350

David  Lotion  1163  Kersey Rd

Howard  David  Loshin  301-754-6350

Knowledge Integrity Incorporated  301  754-6350

David Loshin

301-754-6350

Knowledge Integrity
Matching/Record Linkage

- Pairwise comparisons reliant on similarity scoring (inefficient)
- More efficient algorithms block records into candidate sets, then do pairwise comparisons
  - Statistical (Fellegi & Sunter, Jaro, Winkler)
  - Artificial Intelligence (e.g., Borthwick’s MEDD)
- Approximate string matching
  - Phonetic compression (Soundex, NYSIIS)
  - N-grams
  - Winkler (probability-based)
Data Enhancement

- Data improvement process that relies on record linkage
- Value-added improvement from third-party data sets:
  - Address correction
  - Geo-Demographic/Psychographic imports
  - List append
- Typically partnered with data providers
<table>
<thead>
<tr>
<th>Subject</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OCCUPANCY STATUS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total housing units</td>
<td>1,704</td>
<td>100.0</td>
</tr>
<tr>
<td>Occupied housing units</td>
<td>1,681</td>
<td>98.7</td>
</tr>
<tr>
<td>Vacant housing units</td>
<td>23</td>
<td>1.3</td>
</tr>
<tr>
<td><strong>Tenure</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Occupied housing units</td>
<td>1,681</td>
<td>100.0</td>
</tr>
<tr>
<td>Owner-occupied housing units</td>
<td>1,566</td>
<td>93.2</td>
</tr>
<tr>
<td>Renter-occupied housing units</td>
<td>115</td>
<td>6.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SCHOOL ENROLLMENT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population 3 years and over enrolled in school</td>
<td>1,571</td>
<td>100</td>
</tr>
<tr>
<td>Nursery school, preschool</td>
<td>142</td>
<td>9</td>
</tr>
<tr>
<td>Kindergarten</td>
<td>102</td>
<td>6.5</td>
</tr>
<tr>
<td>Elementary school (grades 1-8)</td>
<td>662</td>
<td>42.1</td>
</tr>
<tr>
<td>High school (grades 9-12)</td>
<td>335</td>
<td>21.3</td>
</tr>
<tr>
<td>College or graduate school</td>
<td>330</td>
<td>21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EDUCATIONAL ATTAINMENT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population 25 years and over</td>
<td>3,438</td>
<td>100</td>
</tr>
<tr>
<td>Less than 9th grade</td>
<td>122</td>
<td>3.5</td>
</tr>
<tr>
<td>9th to 12th grade, no diploma</td>
<td>249</td>
<td>7.2</td>
</tr>
<tr>
<td>High school graduate (includes equivalency)</td>
<td>996</td>
<td>29</td>
</tr>
<tr>
<td>Some college, no degree</td>
<td>674</td>
<td>19.6</td>
</tr>
<tr>
<td>Associate degree</td>
<td>180</td>
<td>5.2</td>
</tr>
<tr>
<td>Bachelor's degree</td>
<td>676</td>
<td>19.7</td>
</tr>
<tr>
<td>Graduate or professional degree</td>
<td>541</td>
<td>15.7</td>
</tr>
<tr>
<td>Percent high school graduate or higher</td>
<td>89.2</td>
<td>(X)</td>
</tr>
<tr>
<td>Percent bachelor's degree or higher</td>
<td>35.4</td>
<td>(X)</td>
</tr>
</tbody>
</table>
Data Profiling

- Empirical analysis of “ground truth”
- Couples:
  - Statistical analysis
  - Functional dependency analysis
  - Association rule analysis
- The statistical analysis part is a no-brainer
- The other stuff is hard
Data Profiling – Approach

- Column Profiling/Frequency Analysis (Easy)
  - Done in place using database capability (e.g., SQL), or using hash tables

- Cross-Table/Redundancy (Harder)
  - May be done using queries, or using set intersection algorithms

- Cross-Column/Dependency (Hardest)
  - Uses fast iterative set partitioning algorithms
  - A priori, TANE, FD_MINE
Data Auditing/Monitoring

- Proactive assessment of compliance with defined rules
- Provide reporting or scorecarding
- Useful for
  - “Quality” processes
  - “Data debugging”
  - Business impact assessment
Issues

- Data quality is seen as a *technical* problem requiring a *technical* solution
- Disconnect between information value and achieving business objectives leads to ignoring DQ until it is too late
- Data is often corrected, instead of flawed processes
Integration with Master Data

- Master reference data management
- Semantic metadata integrated with information quality
- Data quality activity tracking
- Service-oriented processing
- Correlating business impacts to information value
- Complex rule frameworks and systems
- Business rule definition and management
Tools Summary

- 1st generation tools focus on “correction”
- 2nd generation tools look at analysis and discovery
- Growing MDM needs:
  - Standardized rules
  - Business impact correlation
  - Performance metrics
  - Semantic metadata
  - Generic metadata-based descriptive capability
  - Historical auditing and tracking
Questions?

- If you have questions, comments, or suggestions, please contact me

  David Loshin
  301-754-6350
  loshin@knowledge-integrity.com