Reference Data Management Concepts and Best Practices

Ivan Milman
IBM
Senior Technical Staff Member
imilman@us.ibm.com
Table of Contents

Business Overview
- What is Reference Data?
- What is Reference Data Management?
- Why Do Organizations Need Reference Data Management?

IBM Solution Overview
- Logical Architecture
- Capabilities & Benefits

Customer Use Cases
- Banking Use Case
- Healthcare Use Case

Q&A
PRE-CONFERENCE TUTORIAL: REFERENCE DATA MANAGEMENT CONCEPTS AND BEST PRACTICES

Ivan Milman Security and Governance Architect, IBM InfoSphere,

Reference Data Management (RDM) is a relatively new offspring of MDM functionality to provide the processes and technologies for recognizing, harmonizing and sharing coded, relatively static data sets for “reference” by multiple constituencies (people, systems, and other data). Such a system provides governance, process, security and audit control around reference data mastering. The RDM system also manages complex mappings between different reference data representations across the enterprise. Clearly, RDM can be expected to become a "ramp up" point of entry for many organizations planning for Customer, Product master and other domains. Moreover, RDM is a major IT initiative being undertaken by a large number of market-leading global 5000 enterprises. Both as an IT discipline and a commercial off-the-shelf software solution, RDM solutions are being brought to market at an increasing pace. This session provides a full spectrum tutorial on RDM concepts for the MDM and Data Governance professional by discussing:

- Building the business case for enterprise Reference Data Management
- Adopting an MDM approach to managing reference data
- Implementing common patterns for Reference Data Management
What is reference data?

- Code tables, lookup tables, abbreviations:
  - State Codes
  - Zip Codes
  - Product Codes
  - Transaction Codes

- Read-only data provided to operational and analytical systems

- Internal or specified externally (ISO, ANSI, etc.)

- Can have relationships – maps and hierarchies (including a taxonomy)

<table>
<thead>
<tr>
<th>Gender Code</th>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not Known</td>
<td>Desconocido</td>
</tr>
<tr>
<td>1</td>
<td>Male</td>
<td>Hombre</td>
</tr>
<tr>
<td>2</td>
<td>Female</td>
<td>Mujer</td>
</tr>
<tr>
<td>9</td>
<td>Not Applicable</td>
<td>No Aplica</td>
</tr>
</tbody>
</table>

Why is it important?

- Reference data is EVERYWHERE

- Codes must be kept in sync across systems for synchronized for interoperability and accuracy
Reference Data – Everyday Example

<table>
<thead>
<tr>
<th>Luxury Airlines</th>
<th>BOARDING PASS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Passenger Name/Title</strong></td>
<td><strong>Flight</strong></td>
</tr>
<tr>
<td>Erik O’Neill, Dr.</td>
<td>HW3285</td>
</tr>
<tr>
<td><strong>From</strong></td>
<td><strong>Date</strong></td>
</tr>
<tr>
<td>New York - LGA</td>
<td>05APR</td>
</tr>
<tr>
<td><strong>To</strong></td>
<td></td>
</tr>
<tr>
<td>Los Angeles - LAX</td>
<td></td>
</tr>
<tr>
<td><strong>Class of Svc</strong></td>
<td><strong>Frequent Flyer #</strong></td>
</tr>
<tr>
<td>Q</td>
<td>DHD7352</td>
</tr>
</tbody>
</table>
Reference data code examples for EU bank

- VAT codes [VAT Codes (Ireland and UK)]
- Company codes
- Currencies [ISO 4217 Currency Codes]
- Countries [ISO 3166 Country Codes]
- Territories
- Organizational Unit types
- NACE codes [NACE Codes (partial)]
- NACE Codes and relation to ISIC
- Sector codes [World Bank Sector Codes]
- Back End System codes
- Language codes [ISO 639.2 Language Codes]
- Postal codes [UK Postal Codes]
- Organizational unit roles
- Price types [Price Types (Market Order Types)]
- Rating [Fitch Ratings]
- Collateral [Collateral Codes]
- Scoring [FICO Score Reason Codes]
- Customer segments
- Facility types [Facility Type Codes (US Healthcare)]
- Country dialing codes
- Internal exchange rates [Exchange Rates Table (US Dollar)]
- Tasks
- Business process
- Sensitivity type
- Public Authority [European Public Authority Codes (Organic Farming Certific...]

IBM Confidential
Reference Data Across The Enterprise

- Reference data is isolated in silos
  - Each system has a different subset, format and internal representation
  - Often no agreed internal standard for a specific reference data
    - Developers will then create their own, making another silo!

- Cross enterprise reference data needs to be reconciled
  - For application integration and business process integration
  - For reporting, compliance, and BI analysis
  - For master data management initiatives

- Need to manage reference data provided by external standard bodies
Typical Problems caused by lack of Reference Data Management

Code tables are typically defined and managed on an application by application basis

- As a result of this, the values and codes for a specific reference data set will differ within each application. When integrating data across applications, it is necessary to reconcile these code tables. Without centralized management, it is almost impossible to accurately and cost effectively manage the mappings and translations between the different representations.
Challenges of Managing Reference Data

- **Manual Changes**
  - Lack of standard process for managing changes
  - No audit trail, no review and approval
  - Difficult to implement security policy
  - Sarbanes-Oxley Act implications

- **No clear ownership**
  - Limited business involvement
  - IT ownership and involvement in application specific code table maintenance
  - Often no agreed internal standard

- **Managed in multiple places**
  - Leads to duplication and introducing errors
  - No way to manage reference data at the enterprise level
  - No process for agreeing on changes to reference data across the business

**Risk**

**Lack of effectiveness**

**Cost related to lack of efficiency**
Justifying Reference Data Management

- **Costs associated with risk factors**
  - Lower quality of data used to make business decisions or supplied to third parties
    - Roll-ups do not work
    - Data is missing but this is not obvious
    - Interpretation of data is incorrect

- **Costs related to lack of efficiency**
  - Replicated functionality
  - Replicated data entry
  - Replicated IT support
  - Cost of manually maintaining mappings and synchronizations

- **Costs related to lack of effectiveness**
  - Correction of errors, and cost of rejected transactions, incomplete reports, and so on
  - Point to Point integration leads to inflexible architecture over time that is costly to maintain and support
  - Brittle integrations make it difficult to respond quickly to business and process change

- **Typical Reference Data Harmonization Project cost**
  - *without a reference data management tool* (IBM Global Services study)
  - One time discovery, cleansing and data reconciliation manually
  - Does not include cost of solution to maintain reference data on ongoing basis

  - For Large company (Fortune 500)
    - Estimate 100,000 hours effort

  - For Medium Size company
    - Estimate 20,000 hours effort
Business Benefits of Managing Reference Data

Reference Data Management

- Improved Business Results
- Improved BI Reporting
- Improved Regulatory Compliance
- Improved Business Agility
- Reduced Operational Risk
- Reduced IT Costs

© 2012 IBM Corporation
Core Domain Objects in Reference Data

- **Sets**
  - Can represent any type of reference data
  - Core properties are *code* and *value*
  - Extendible at a set level and a value level
    - Any type of property
    - Built in properties for lifetimes, lifecycle, etc.
  - Includes translations of values

- **Maps**
  - Links values in one set to another
    - Extensible with properties (like sets)
  - Used for data integration, transcoding and hierarchies

- **Hierarchies**
  - Parent child relationship between values
    - Within a set
    - Between sets of different types
    - Explicit relationship or map link
  - Used in dimensional data and business relationships
Hierarchies Are Core Reference Data

- Hierarchies are used for rollups and relationships
- Multiple types of hierarchies
  - Tree based within a set
    • NAICS
  - Level based
    • City->Province->Country
  - Tree based external to a set
    • Expense codes and categories
  - Combined (Hybrid)
- Consumed by warehouses and applications
Level-based Hierarchy

Level 1
- NA

Level 2
- USA

Level 3
- TX
- CA

Level 4
- AUS
- DAL
- ..
Hierarchy Level Definition
Hierarchy Visualization
Hybrid Hierarchy

Level 1

... Expense codes

Travel

Non-Travel

Level 2

Hotel - 21

Airfare - 24

Phone - 25

Internet - 30
Reference Data Domain Objects

Data Lifecycle / Ownership / Versioning – Governance for Reference Data

Customizable Life Cycle for Ref Data
- Out of the box lifecycles

Versioning
- New versions can inherit values from existing versions
- In V11, can merge between versions

Ownership and Stewardship
- Can assign different owners and approvers for different sets, maps and hierarchies
- Customizable roles for stewardship, administration and data integration
- Attribute level security
Reference Data Management Targets 4 Core User Personas

- **Business User/Data Steward**
  - Easy to create sets, types, maps, hierarchies via Web U/I
  - Easy to bring in and put out reference data in a typical manner that business users do (spreadsheet)
  - Versioning and Workflow support business activities

- **Data Integrators**
  - Help import and export data
  - Define shape of data
  - Leverage maps extensively in data integration hubs

- **Developers**
  - Use web services and REST to get latest reference data and trans-code values between applications

- **IT**
  - Manage and customize RDM environment
  - Make reference data domain infrastructure available as core enterprise asset
Table of Contents

Business Overview
- What is Reference Data?
- What is Reference Data Management?
- Why Do Organizations Need Reference Data Management?

IBM Solution Overview
- Logical Architecture
- Capabilities & Benefits

Customer Use Cases
- Banking Use Case
- Healthcare Use Case

Q&A
IBM MDM RDM Hub - A Master Data Management Approach to managing Reference Data

- Provide governance, process, security, audit control around mastering reference data

- Manage the complex mappings between different reference data representations across the enterprise

- Efficiently manage use of external reference data standards within the enterprise
Why use MDM as the basis for an RDM Solution

- Reference Data and Master Data are
  - Application independent enterprise data resources consumed across the entire enterprise
    - For consistent definition, values, use and understanding
  - Require governance at the business and IT level
- We represent reference data as a specialized domain of master data
- Implement on top of the InfoSphere Custom Domain Hub
  - Provides a **proven, scalable, platform with rich services** for security, business rules, data quality processes, events and notifications, audit, and history.
  - Domain specific reference data management services were created with MDM tooling for creation, maintenance, import, export, transcoding, retrieval and search of reference data
  - Standard MDM mechanism for extensibility (behavior, data, validation, notification, and so on)

- The same concepts and procedures for reference data management (data agreement, integration, governance) are pre-cursors to the concepts and processes for a larger master data management initiative, but with a more focused scope.
Alternative Approaches To Managing Reference Data

- **Dimensional Management in Data Warehouse**
  - Focused solely on the warehouse environment
  - Does not support developers who need reference data standards and maps (for transcoding) in applications
  - Does not support data integrators who need *maps* to manage ETL and transcoding from source into warehouse

- **Domain Specific Reference Data Application**
  - Very common in healthcare (ICD 9 <-> ICD 10)
  - Rich functionality for a specific domain or application, but ...
  - Enterprises want to manage reference data consistently across domains, and not introduce another silo

- **Standalone Reference Data Set Hub**
  - Bespoke functionality for managing reference data
  - Does not inherit or leverage functionality from master data common platform system for extensibility, standardization, services, common infrastructure and tooling
IBM MDM RDM Hub – Logical Architecture

IBM InfoSphere MDM Reference Data Management Hub

- RDM Services
  - Import/Export
  - Reference Data Lifecycle Management

- MDM Services
  - Security Services
  - Business Rules, Events, Notification, Data Quality
  - Audit/History

Reference Data Management User Interface
- View
- Author
- Map
- Approve

Sources → Import

Applications → Consume
RDM Component View
InfoSphere MDM Reference Data Management Hub:

Reference Data Domain Objects
- Hierarchies
- Maps
- Sets

Web Services
- Export

Applications

Data Warehouse
- Reference Data Hierarchy Dimension
- Reference Data Dimension
- Transactional Data

Source Systems

Export

DataStage

Lookup Tables
Standard RDM Paradigm for Import and Export of Reference Data

- Based on customer requirements and feedback, the RDM product moved to a *loosely-coupled* pattern for import and export of reference data between RDM and source/consuming systems targeted at *business users*.

- Usually no direct interaction with managed systems
  - Requires less technical know-how on the part of reference data stewards
  - Uses common data format (CSV) that business users work with
  - But supports XML for *data integrators*, who can use this populate target systems.
Updating target systems with reference data

- Like any other changes in your information supply chain, updates to reference data need to be co-ordinated

- Questions to ask
  - What should trigger a change?
    - Approval of a new version of sets/maps/hierarchies
    - Dates
    - Leverage subscription of managed system to reference data updates
  - What is the process for rolling out changes?
    - Order of systems
    - Do systems need to be stopped and restarted
  - How do you handle “old values”
    - When are they strictly forbidden
    - When are they tolerated and mapped to newer values
    - Use versions to keep a history in place for seeing how reference data has evolved
Customizable Life Cycle for Ref Data

- Out of the box lifecycles

Versioning

- New versions can inherit values from existing versions
- In V11, can merge between versions

Ownership and Stewardship

- Can assign different owners and approvers for different sets, maps and hierarchies
- Customizable roles for stewardship, administration and data integration
- Attribute level security
IBM RDM: Key Features & Benefits

Features
- Reference data defined and managed as centralized asset
- Designed to be configurable, extensible and ready-to-run
- Centralized audit, security, and stewardship
- UI designed for Business users to manage reference data, hierarchies and maps
- Versioning and lifecycle management
- Role-based UIs for review, approval, and publishing of changes

Benefits
- Fast Time to Value
- Consistency
- Auditability
- Security
- Agility
- Limited IT involvement

Impacts
- Fewer errors
- Reduced IT costs
- Improved BI analyses
- Improved compliance
- Faster response to changes
Table of Contents

Business Overview
- What is Reference Data?
- What is Reference Data Management?
- Why Do Organizations Need Reference Data Management?

IBM Solution Overview
- Logical Architecture
- Capabilities & Benefits

Customer Use Cases
- Patterns for Managing Reference Data
- Banking Use Case
- Healthcare Use Case

Q&A
Patterns of Managing Reference Data

- Enterprise reference data standard
- Local and application-specific reference data mapping
- Run-time reference data translation
Typical Banking Scenarios

- Maintaining Industry classification codes (NACE, MiFID) with hierarchies and mappings between versions

- Publishing reference data (changes) to consuming systems

- Management and Consolidation of Reference Data for data warehouse load
RDM Banking Customer -- DNB

**Background:** Norway’s largest bank with offices in 19 countries, and operations including retail banking, mutual funds, life insurance, pensions, heavy presence in shipping industry

**Catalyst:** Government mandate to adopt new classifications for institutional sector classifications in line with revised international standards

**Identified Need:** Business application to manage referential data on a global scale

- Master and source datasets
- Mappings
- Hierarchies
- Import / export within the application
- Governance processes with workflow support within the application

**Benefits:**
- Less time & cost to support new financial regulations
- Greater efficiency in updating, distributing and managing reference data
- Reduces risk in implementing an mdm initiative
RDM Hub provides centralized stewardship of reference data with audit control/history, security, and lifecycle management.
DNB Use Case: Maintain and publish standards

- **External Standards**
  - Eg NACE
  - NAICS
  - ISO 3166 (Country Code)

- **IBM RDM Hub**
  - **Import**
  - **Author**
  - **Approve**
  - **Publish Changes to the Standard**
  - **Subscribe to changes for a Standard**

- **Operational Systems**
Use Case: Maintain and publish both standards and application specific reference sets and their mapping to the standard

IBM RDM Hub

Import

Author

Approve

External Standards

Eg
NACE
NAICS
ISO 3166 (Country Code)

Operational Systems

Manage application specific representation of the code set (eg SAP country code) and mapping to the canonical (eg ISO-3166 code)
Use Case: Managing MDM ref data map files

An MDM Hub needs to map ref data from application specific format to MDM Code table format and vice versa.

Operational Source Systems
Structured/ Unstructured Data

IBM RDM Hub

D/W Map File

MDM Map File

Reference Data

Data

Publish

Exception Process

Author

Approve

Data Warehouse

MDM Hub
Typical Healthcare Scenario

- US Healthcare Industry going through major regulatory changes
  - ICD 9 – ICD 10 standards change

- Healthcare Industry has same issue of managing “simple” reference data as other industries
  - Eg Country codes, State codes

- Healthcare has industry specific code sets and terminologies
  - ICD9, ICD 10, SNOMED, LOINC, etc
  - Need to manage ongoing changes to code sets, mappings, groupings
# Lab Orders Demo/Use Case

<table>
<thead>
<tr>
<th>Test Code</th>
<th>Test Description</th>
<th>Result Code</th>
<th>Result Description</th>
<th>Practice Test Code</th>
<th>Practice Result Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>10015</td>
<td>BASIC METABOLIC PROFILE (CHEM 8)</td>
<td>BUN</td>
<td>BUN</td>
<td>C8</td>
<td>Ay-65121</td>
</tr>
<tr>
<td>10015</td>
<td>BASIC METABOLIC PROFILE (CHEM 8)</td>
<td>CA</td>
<td>CALCIUM</td>
<td>C8</td>
<td>Az-65122</td>
</tr>
<tr>
<td>10015</td>
<td>BASIC METABOLIC PROFILE (CHEM 8)</td>
<td>CL</td>
<td>CHLORIDE</td>
<td>C8</td>
<td>Bd-66100</td>
</tr>
<tr>
<td>10015</td>
<td>BASIC METABOLIC PROFILE (CHEM 8)</td>
<td>CO2</td>
<td>CO2</td>
<td>C8</td>
<td>Bd-6698</td>
</tr>
<tr>
<td>10015</td>
<td>BASIC METABOLIC PROFILE (CHEM 8)</td>
<td>CRE</td>
<td>CREATININE</td>
<td>C8</td>
<td>Br-66114</td>
</tr>
<tr>
<td>10015</td>
<td>BASIC METABOLIC PROFILE (CHEM 8)</td>
<td>DIET</td>
<td>DIETARY STATUS</td>
<td>C8</td>
<td>Fa-7097</td>
</tr>
<tr>
<td>10015</td>
<td>BASIC METABOLIC PROFILE (CHEM 8)</td>
<td>GLU</td>
<td>GLUCOSE</td>
<td>C8</td>
<td>Cj-67106</td>
</tr>
<tr>
<td>10015</td>
<td>BASIC METABOLIC PROFILE (CHEM 8)</td>
<td>K</td>
<td>POTASSIUM</td>
<td>C8</td>
<td>Dv-68119</td>
</tr>
<tr>
<td>10015</td>
<td>BASIC METABOLIC PROFILE (CHEM 8)</td>
<td>NA</td>
<td>SODIUM</td>
<td>C8</td>
<td>Ej-69106</td>
</tr>
<tr>
<td>10015</td>
<td>BASIC METABOLIC PROFILE (CHEM 8)</td>
<td>SPECTYPE</td>
<td>SPECIMEN TYPE</td>
<td>C8</td>
<td>Aj-65106</td>
</tr>
<tr>
<td>10026</td>
<td>ELECTROLYTES</td>
<td>CL</td>
<td>CHLORIDE</td>
<td>LYT</td>
<td>Bd-66100</td>
</tr>
<tr>
<td>10026</td>
<td>ELECTROLYTES</td>
<td>CO2</td>
<td>CO2</td>
<td>LYT</td>
<td>Bb-6698</td>
</tr>
<tr>
<td>10026</td>
<td>ELECTROLYTES</td>
<td>K</td>
<td>POTASSIUM</td>
<td>LYT</td>
<td>Dw-68119</td>
</tr>
<tr>
<td>10026</td>
<td>ELECTROLYTES</td>
<td>NA</td>
<td>SODIUM</td>
<td>LYT</td>
<td>Ej-69106</td>
</tr>
<tr>
<td>10026</td>
<td>ELECTROLYTES</td>
<td>null</td>
<td>SPECIMEN TYPE</td>
<td>LYT</td>
<td>Aj-65106</td>
</tr>
<tr>
<td>10038</td>
<td>COMPREHENSIVE METABOLIC (CHEM 14)</td>
<td>A/GRATIO</td>
<td>A/G RATIO</td>
<td>C14</td>
<td>Ah-65104</td>
</tr>
<tr>
<td>10038</td>
<td>COMPREHENSIVE METABOLIC (CHEM 14)</td>
<td>ALB</td>
<td>ALBUMIN</td>
<td>C14</td>
<td>Ab-6598</td>
</tr>
<tr>
<td>10038</td>
<td>COMPREHENSIVE METABOLIC (CHEM 14)</td>
<td>ALKP</td>
<td>ALKALINE PHOSPHATASE</td>
<td>C14</td>
<td>Ak-65107</td>
</tr>
<tr>
<td>10038</td>
<td>COMPREHENSIVE METABOLIC (CHEM 14)</td>
<td>ALT</td>
<td>ALT (SGPT)</td>
<td>C14</td>
<td>An-65110</td>
</tr>
<tr>
<td>10038</td>
<td>COMPREHENSIVE METABOLIC (CHEM 14)</td>
<td>AST</td>
<td>AST (SGOT)</td>
<td>C14</td>
<td>As-65115</td>
</tr>
<tr>
<td>10038</td>
<td>COMPREHENSIVE METABOLIC (CHEM 14)</td>
<td>BILITOTAL</td>
<td>BILIRUBIN,TOTAL</td>
<td>C14</td>
<td>Avr-65119</td>
</tr>
<tr>
<td>10038</td>
<td>COMPREHENSIVE METABOLIC (CHEM 14)</td>
<td>BUN</td>
<td>BUN</td>
<td>C14</td>
<td>Ay-65121</td>
</tr>
<tr>
<td>10038</td>
<td>COMPREHENSIVE METABOLIC (CHEM 14)</td>
<td>CA</td>
<td>CALCIUM</td>
<td>C14</td>
<td>Az-65122</td>
</tr>
<tr>
<td>10038</td>
<td>COMPREHENSIVE METABOLIC (CHEM 14)</td>
<td>CL</td>
<td>CHLORIDE</td>
<td>C14</td>
<td>Bd-66100</td>
</tr>
<tr>
<td>10038</td>
<td>COMPREHENSIVE METABOLIC (CHEM 14)</td>
<td>CO2</td>
<td>CO2</td>
<td>C14</td>
<td>Bb-6698</td>
</tr>
</tbody>
</table>
Lab Order Data Flow

CHEM8 Order Scenario

Clinic 1

NextGen EMR

Order #C8

Multiple Results

AY  A  Bd  Bb
Br  Cg  Dw  Ej

Hospital A

Cerner EMR

Cerner Lab

Order #10015

Multiple Results

BU  C  A  C  CO
L  2

CR  E
GL  U
K  NA

Hospital B

Allscripts EMR

Sunquest Lab

Order #CHEM8

Multiple Results

BU  C  A  C  CO
L  2

CR  E
GL  U
K  NA
Practice code for Chem-8 test maps to Hospital code
Hospital Chem-8 Lab Test maps to Lab-Test results
Lab Results codes map to Physicians results codes
For More Information

- **Articles**
  - *Intro to IBM InfoSphere RDM Hub v10*
  - *Ontology-guided Reference Data Alignment in Information Integration Projects*
    [http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1038&context=amcis2012](http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1038&context=amcis2012)

- **Books**
  - *A Practical Guide to Managing Reference Data with IBM InfoSphere RDM Hub*
  - *Managing Reference Data in Enterprise Databases by Malcolm Chisholm* (out of print, alas)
  - *Information Quality and Governance for Business Intelligence (forthcoming)*
    - Chapter on reference data [http://www.igi-global.com/publish/call-for-papers/call-details/832](http://www.igi-global.com/publish/call-for-papers/call-details/832)

- **Case Studies**
  - *DNB ASA Centralizes RDM for more effective governance and reduced business risk*

- **Web Sites**
  - *IBMs RDM Home*
Questions?