Data Parsing, Matching and De-duplication

by Kathy Hunter, William McKnight, and Henrik Liliendahl Sørensen

Module 0. About the Course (11 min)

- *About the Authors*
- *Course Objectives*
- *Audience and Pre-requisites*
- *Course Structure*

Module 1. Introduction (17 min)

- *Duplicates within Systems*
- *Matches across Systems*
- *Key Definitions*
- *Setting the Context*
- *Relations with Other Disciplines*
- *The Goals*
- *Target Data*
- *Master Data Domains*
- *Data Formats*
- *Responsibilities of a Data Steward*

Module 2. Implementation Fundamentals (61 min)

- *Parsing and Standardization*
  - Name Elements
  - Postal Addresses
  - Geocoding
  - Date and Time
  - Other Data elements
  - Synonyms and Vocabularies
  - Script Systems
- *Data Matching Techniques*
  - Deterministic vs. Probabilistic
  - Conservative vs. Liberal
  - What is a Match?
  - Data Example
  - Rule-Based Algorithms
  - Probabilistic Matching Algorithms
  - Match & Merge Example
  - Other Common Situations
  - Best Practices
  - Match Codes and Phonetics
  - Fuzzy Logic
  - Evaluating Matching Results
  - Practical Tips
  - Unexpected Side Effects
- *Data Matching Destinations*
  - Reference Data Matching
  - Single-Purpose vs. Enterprise
  - Real-Time Prevention
  - Data Survivorship
  - Identity Resolution

- *Evaluating Data Matching Tools*
  - Matching Tool Packaging
  - Matching Tool Requirements
  - Build or Buy
  - Testing a Matching Tool

Module 3. External Reference Data (45 min)

- *External Data Sources*
  - Address Directories
  - Postal Corrections
  - Business Directories
  - Unique B2B Customer Identification
  - Challenges
  - Customer Identification with DUNS
  - Recommendations
  - Other Party Directories
- *Syndicated Customer Data*
  - Why Syndicated Data?
  - Functions of Customer Profile
  - Customer Lifetime Value
  - What is Syndicated Data?
  - Common Data Elements
  - Reverse Append
  - Matching and Consolidation
  - Syndicators
  - Data Available from Syndicators
  - Syndicated Sourcing
  - Information Architecture
- *Syndicated Product Data*
  - Global Data Synchronization
  - Global Product Identification
  - Electronic Product Code (EPC)
  - EPC Information Flow
  - Operational and Strategic Decisions
  - Other Product Directories
  - Evaluating Directory Services
- *Using the Web*
  - External Business Content “Open Source”
  - Web Syndication

Module 4. Challenges of Global Data (58 min)

- *Introduction to Global Information*
  - Why Go Global?
  - Technology for Business
  - Capitalizing on the Worldwide Marketplace
  - Global Data Challenges
  - Data Management Concerns
  - Implementing a Global Business
  - Global Data Governance

- *Global Data: What You Need to Know*
  - Address Maturity
  - Address Format Differences
  - Japan Address Example
  - Personal Names
  - Personal Name Variations
  - Name Variations by Country
- *Variations by Country and Region*
  - Business Naming Conventions
  - Phone Numbers
  - P. O. Boxes
  - Job Title Issues
  - Specific Regions (France, Spain, Canada, South Korea)
- *Cultural and Legal Impacts*
  - The Importance of Culture
  - Practical Issues Involving Culture
  - Cultural Differences Examples
  - Privacy Laws
  - Privacy Variations
- *Characters and Diacritics*
  - Handling Regional Alphabets
  - Multiple Character Sets
  - Code Pages
  - Unicode

Module 5. Overcoming the Challenges of Global Data (59 min)

- *Data Profiling*
  - Uniqueness, Completeness, Null Values
  - Frequency and Distribution
  - Patterns
  - Address Field Issues
  - Personal and Business Name Issues
  - Job Titles, Email, and Phone
- *Consistent Data Structures*
  - Ensuring Consistency in Data Structures
  - Web and Application Forms
  - Forms – Common Troublesome Areas
  - Example of Good Data Capture Form
  - More Areas to Consider
- *Preparing Global Data for Effective Use*
  - Data Re-engineering
  - Data Parsing
  - Address Validation
  - Data Matching
  - Data De-duplication
  - Consolidating Data Across Sources
  - Matching Households
  - Creating Corporate Hierarchies
  - Address Reconstruction Rules