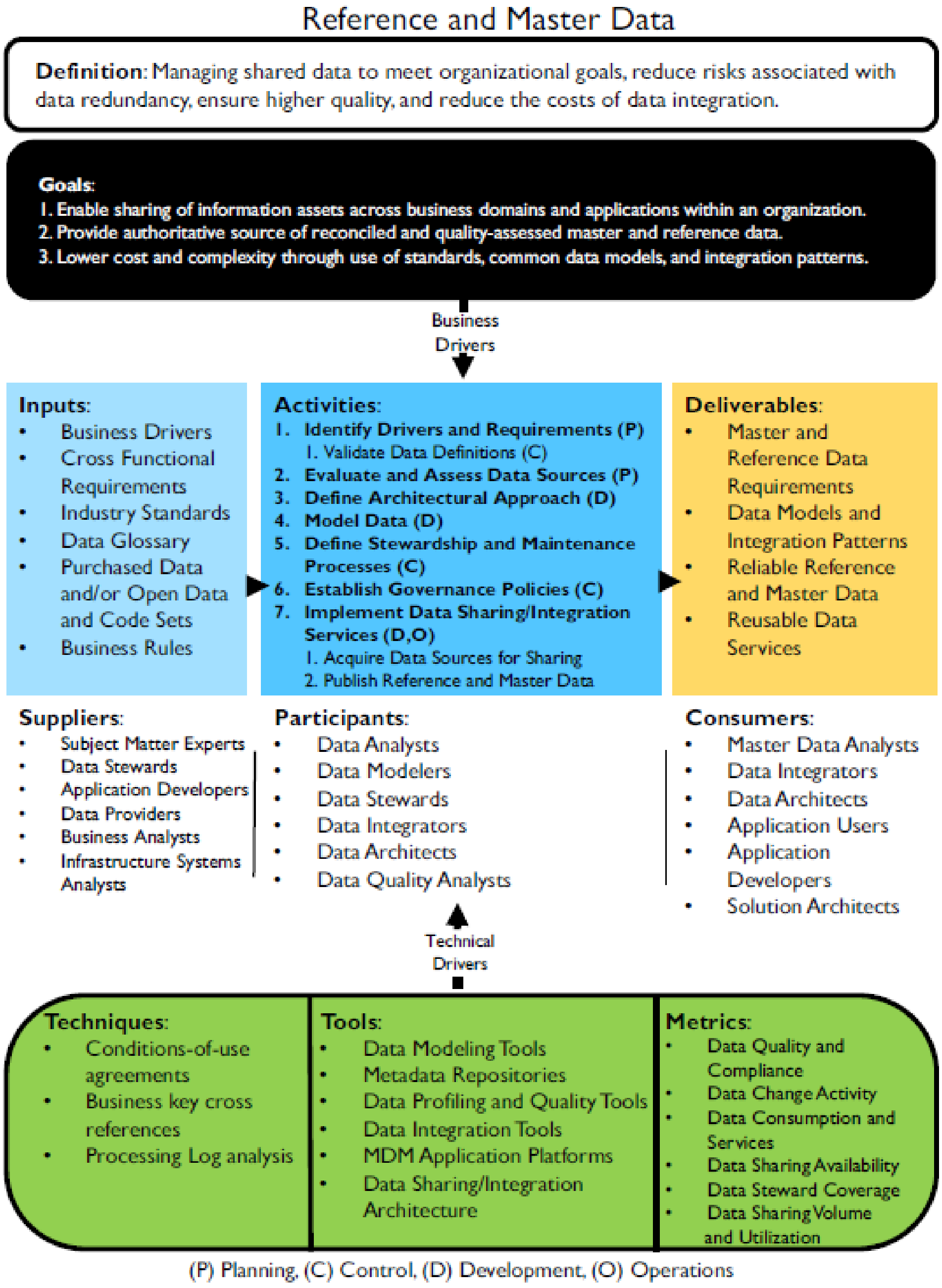
Reference and Master Data

1 Introduction

An organisation and its customers benefit if the data required across business areas, processes and systems is shared, allowing the same customer lists, geographic codes, parts codes etc. to be accessed, to produce a level of consistency.

Systems and data evolve organically resulting in multiple systems executing similar functions isolated from each other, leading to inconsistencies in data structure and values, and increased costs and risks. Both can be reduced through the management of reference and master data.



Reference and Master Data

Definition: Managing shared data to meet organizational goals, reduce risks associated with data redundancy, ensure higher quality, and reduce the costs of data integration.

Goals:
1. Enable sharing of information assets across business domains and applications within an organization.
2. Provide authoritative source of reconciled and quality-assessed master and reference data.
3. Lower cost and complexity through use of standards, common data models, and integration patterns.

Business Drivers

Inputs:
- Business Drivers
- Cross Functional Requirements
- Industry Standards
- Data Glossary
- Purchased Data and/or Open Data and Code Sets
- Business Rules

Activities:
1. Identify Drivers and Requirements (P)
   1. Validate Data Definitions (C)
2. Evaluate and Assess Data Sources (P)
3. Define Architectural Approach (D)
4. Model Data (D)
5. Define Stewardship and Maintenance Processes (C)
6. Establish Governance Policies (C)
7. Implement Data Sharing/Integration Services (D,O)
   1. Acquire Data Sources for Sharing
   2. Publish Reference and Master Data

Deliverables:
- Master and Reference Data Requirements
- Data Models and Integration Patterns
- Reliable Reference and Master Data
- Reusable Data Services

Suppliers:
- Subject Matter Experts
- Data Stewards
- Application Developers
- Data Providers
- Business Analysts
- Infrastructure Systems Analysts

Participants:
- Data Analysts
- Data Modelers
- Data Stewards
- Data Integrators
- Data Architects
- Data Quality Analysts

Consumers:
- Master Data Analysts
- Data Integrators
- Data Architects
- Application Users
- Application Developers
- Solution Architects

Technical Drivers

Techniques:
- Conditions-of-use agreements
- Business key cross references
- Processing Log analysis

Tools:
- Data Modeling Tools
- Metadata Repositories
- Data Profiling and Quality Tools
- Data Integration Tools
- MDM Application Platforms
- Data Sharing/Integration Architecture

Metrics:
- Data Quality and Compliance
- Data Change Activity
- Data Consumption and Services
- Data Sharing Availability
- Data Steward Coverage
- Data Sharing Volume and Utilization

(P) Planning, (C) Control, (D) Development, (O) Operations

1.1 Business Drivers

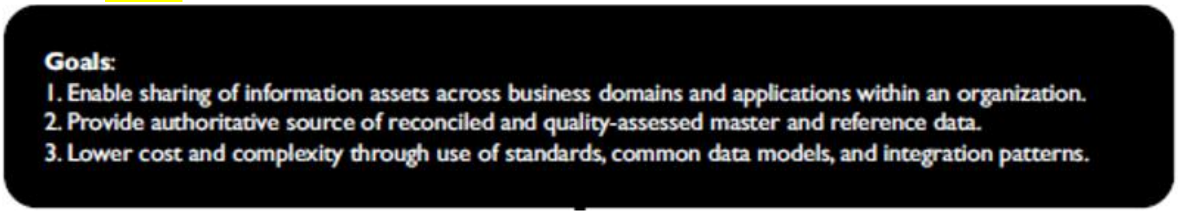
For Master Data management program:

- **Meeting organisational data requirements**: Multiple areas need to access same data sets, which confidence that they are complete and consistent
- **Managing data quality**: MDM enables a consistent representation of critical entities
- **Managing the costs of data integration:**
- **Reducing risk:** Simplifies the data sharing environment

Centrally managed Reference Data enables the organisation to:

- Meet data requirements for multiple initiatives, reduce costs and risks of data integration
- Manage quality of reference data

1.2 Goals and Principles



Goals:

1. Enable sharing of information assets across business domains and applications within an organization.
2. Provide authoritative source of reconciled and quality-assessed master and reference data.
3. Lower cost and complexity through use of standards, common data models, and integration patterns.

R and MDM guiding principles:

- **Shared data:** managed to be sharable across organisation
- **Ownership**: Belong to the organisation. Require high level of stewardship
- **Quality**: Require ongoing monitoring and governance
- **Stewardship**: Business data stewards responsible
- **Controlled Change**:
  - **Master Data**: Represents the best view of currency and accuracy at any point in time. Caution when matching rules that change values. Should be reversible.
  - **Reference Data**: Change follows defined process. Approve and communicate before implementing
- **Authority**: Master data values should be replicated only from the system of record.

1.3 Essential concepts

1.3.1 Differences between Master and Reference Data

Malcolm Chisholm proposed a six-layer taxonomy of data (2008):

- Metadata
- Reference Data
- Enterprise structure data
- Transaction structure data
- Transaction activity data
- transaction audit data

Chisholm's definition of Master Data: An aggregation of Reference Data, enterprise structure data and transaction structure data.

- **Reference data:** code and description tables used to categorise other data in the organisation. Relates data in the database to information outside the organisation
- **Enterprise Structure Data:** Business activity data e.g. chart of accounts

- **Transaction Structure data:** Describes things that must be present for a transaction to occur e.g. products, customers, vendors

DAMA Dictionary Definition (2009): Master Data is the data that provides the context for business activity data in the form of common and abstract concepts that relate to the activity. It includes the details (definition and identifiers) of internal and external objects involved in business transactions, such as customers, products, employees, vendors and controlled domains (code values).

David Loshin describes Master Data objects as core business objects used in different applications across the organisation along with their associated Metadata, attributes, definitions, roles, connections and taxonomies. Master Data objects represent those things that matter most to the organisation, that are logged in transaction, reported on, measured and analysed.

Master Data requires identifying and developing a trusted version of the truth for each instance of conceptual entities, and maintaining the currency of that version. Master Data Management works to resolve the differences in associations between data in different systems and processes to consistently identify individual entity instances in different contexts. This process must be managed over time so that the identifiers for these Master Data entity instances remain consistent.

- Shared purposes of Reference and Master Data:
- Provide context to the creation and use of transactional data
- Reference data provides context for Master Data
- Enable data to be meaningfully understood
- Shared resources managed at the enterprise level
- Reference Data compared to Master Data sets:
- Less volatile
- Fewer columns and rows than transactional or master data sets
- No entity resolution challenges

Different focus of data management:

- **Master Data Management (MDM):** Control over Master Data values and identifiers that enable consistent use of the most accurate and timely data about essential business entities. Ensure availability of accurate current values while reducing risks of ambiguity.
- **Reference Data Management (RDM):** Control over defined domain values and their definitions. Ensure the organisation has access to a complete set of accurate and current values for each concept represented.

RDM is responsible for obtaining data and managing updates, as reference data can originate inside or outside the organisation.

1.3.2 Reference Data

Reference data is any data used to characterise or classify other data, or to relate data to information external to an organisation (Chisholm, 2001). Can be codes and description or more complex hierarchies and mappings.

Common storage techniques:

- Code tables in relational databases linked by foreign keys
- Reference Data Management systems
- Object attribute specific Metadata to specify permissible values for APIs

Reference Data Management entails control and maintenance of defined domain values, definitions and the relationships with and across domain values. The goal is to ensure values are consistent, current and accessible to the organisation

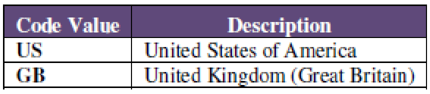
1.3.2.1 Reference data structure:

Depends on the granularity and complexity.

1.3.2.2 Lists

Code value and a description which can be used in a drop down.

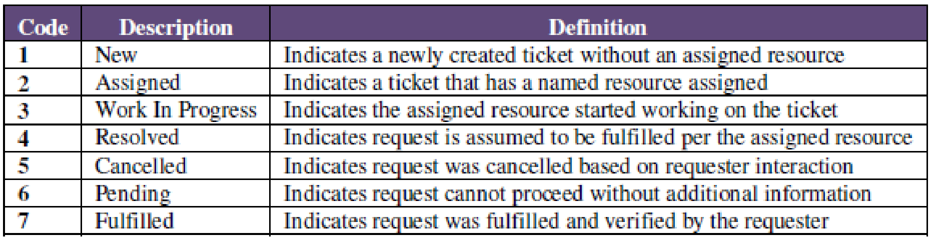
Table 17 Simple Reference List



Code Value	Description
US	United States of America
GB	United Kingdom (Great Britain)

Definitions added for a Help function.

Table 18 Simple Reference List Expanded

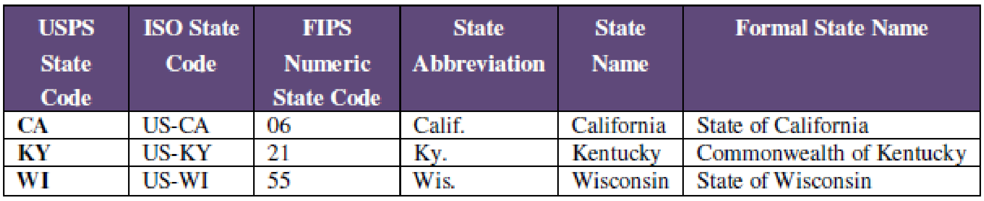


Code	Description	Definition
1	New	Indicates a newly created ticket without an assigned resource
2	Assigned	Indicates a ticket that has a named resource assigned
3	Work In Progress	Indicates the assigned resource started working on the ticket
4	Resolved	Indicates request is assumed to be fulfilled per the assigned resource
5	Cancelled	Indicates request was cancelled based on requester interaction
6	Pending	Indicates request cannot proceed without additional information
7	Fulfilled	Indicates request was fulfilled and verified by the requester

1.3.2.3 Cross-reference lists

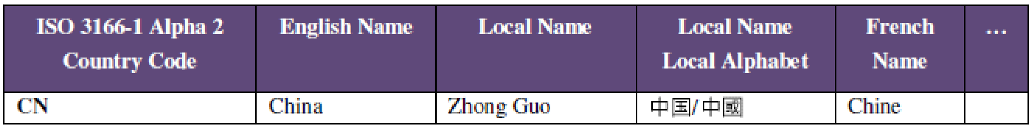
Used to translate code values of the same concept. May be at different granularities or same granularity with different values.

Table 19 Cross-Reference List



USPS State Code	ISO State Code	FIPS Numeric State Code	State Abbreviation	State Name	Formal State Name
CA	US-CA	06	Calif.	California	State of California
KY	US-KY	21	Ky.	Kentucky	Commonwealth of Kentucky
WI	US-WI	55	Wis.	Wisconsin	State of Wisconsin

Table 20 Multi-Language Reference List

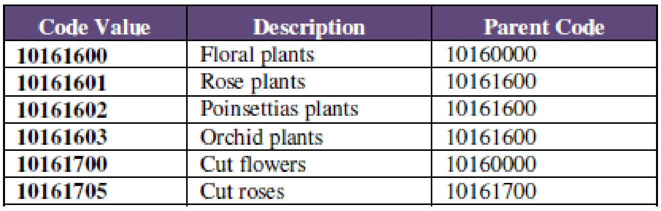


ISO 3166-1 Alpha 2 Country Code	English Name	Local Name	Local Name Local Alphabet	French Name	…
CN	China	Zhong Guo	中国/中國	Chine	

1.3.2.4 Taxonomies

Taxonomic Reference Data structures capture information at different levels of specificity to support multifaceted navigation required by Business Intelligence.

Table 21 UNSPSC (Universal Standard Products and Services Classification)[57]



Code Value	Description	Parent Code
10161600	Floral plants	10160000
10161601	Rose plants	10161600
10161602	Poinsettias plants	10161600
10161603	Orchid plants	10161600
10161700	Cut flowers	10160000
10161705	Cut roses	10161700

1.3.2.5 Ontologies

Ontologies can be part of Reference Data as they are used to characterise other data or relate organisational data to information beyond the boundaries of the organisation.

1.3.2.6 Proprietary or internal reference data

Reference data created within the organisation to support internal systems. RDM consists of managing them and ensuring consistency.

1.3.2.7 Industry reference data

Industry Reference Data describes data sets which are created and maintained by industry associations and government bodies in order to provide a standard for codifying important concepts.

1.3.2.8 Geographic or geo-statistical data

Enables classification or analysis based on geography.

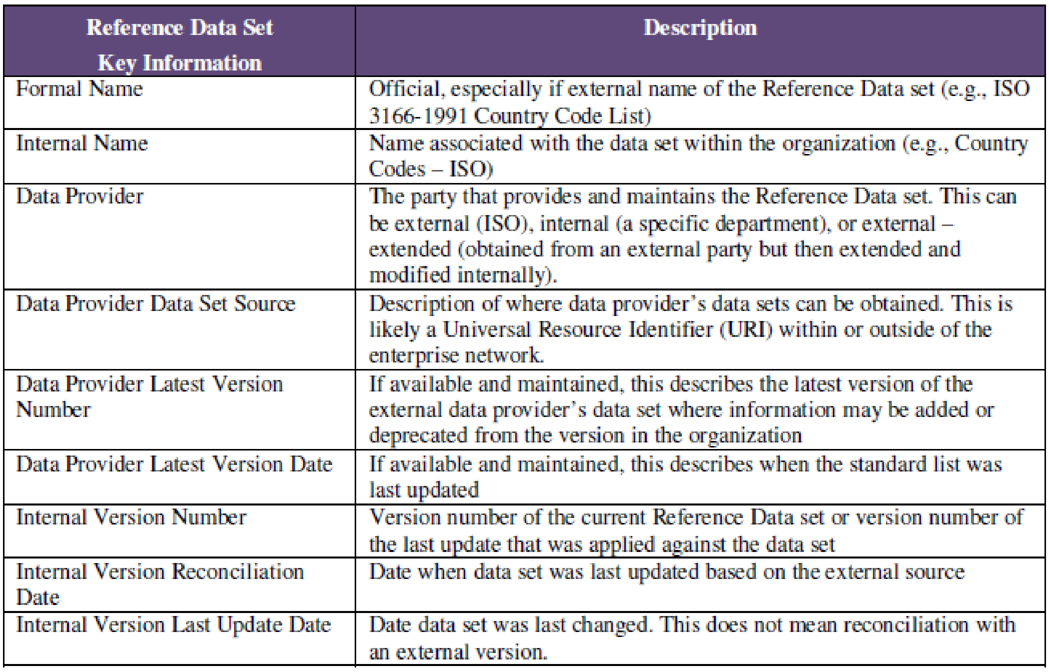
1.3.2.9 Computational reference data

Used for common, consistent calculations

1.3.2.10 Standard reference data set metadata

Maintain key Metadata about Reference Data sets to ensure their lineage and currency are understood and maintained.

Table 23 Critical Reference Data Metadata Attributes



Reference Data Set Key Information	Description
Formal Name	Official, especially if external name of the Reference Data set (e.g., ISO 3166-1991 Country Code List)
Internal Name	Name associated with the data set within the organization (e.g., Country Codes – ISO)
Data Provider	The party that provides and maintains the Reference Data set. This can be external (ISO), internal (a specific department), or external – extended (obtained from an external party but then extended and modified internally).
Data Provider Data Set Source	Description of where data provider's data sets can be obtained. This is likely a Universal Resource Identifier (URI) within or outside of the enterprise network.
Data Provider Latest Version Number	If available and maintained, this describes the latest version of the external data provider's data set where information may be added or deprecated from the version in the organization
Data Provider Latest Version Date	If available and maintained, this describes when the standard list was last updated
Internal Version Number	Version number of the current Reference Data set or version number of the last update that was applied against the data set
Internal Version Reconciliation Date	Date when data set was last updated based on the external source
Internal Version Last Update Date	Date data set was last changed. This does not mean reconciliation with an external version.

1.3.3 Master Data

Master Data is about the key business entities and should represent the authoritative most accurate data available, which can be trusted and used with confidence.

Business rules dictate the format and allowable values. Common organisational Master Data is data about:

- **Parties:** Individuals, organisations and their roles
- **Products and services:** Internal and external
- **Financial structures:** e.g. contracts, general ledger accounts
- **Locations:** e.g. addresses and GPS coordinates

1.3.3.1 System of Record, System of Reference

Where there are potentially different versions of the truth, we need to know more about the data to distinguish between them:

- **A System of Record** is an authoritative system where data is created/captured and maintained through a defined set of rules and expectations.
- **A System of Reference** is an authoritative system where data consumers can obtain reliable data to support transactions and analysis. Examples are MDM applications, Data Sharing Hubs and Data Warehouses.

1.3.3.2 Trusted Source, Golden Record

A **Trusted Source** is recognised as the "best version of the truth" based on a combination of automated rules and manual stewardship. (any MDM system)

A **Golden Record** represents the most accurate data about an entity, also referred to as a "single version of the truth". Not always possible for multiple systems to have one version of the truth.

1.3.3.3 Master Data Management

Gartner's Definition: A technology-enabled discipline in which business and IT work together to ensure the uniformity, accuracy, stewardship, semantic consistency and accountability of the enterprise's official shared Master Data assets. Master Data is the consistent and uniform set of identifiers and extended attributes that describes the core entities of the enterprise, including customers, prospects, citizens, suppliers, sites, hierarchies and chart of accounts.

MDM is a discipline made up of people, processes and technology.

Criteria to assess MDM requirements:

- Which roles, organisations, places and things are referenced repeatedly?
- What data is used to describe people, places and things
- How the data is defined, structured, and the granularity
- Where data is created/sourced, stored, made available and accessed
- How it changes as it moves through systems
- Who uses the data, and for what purposes?
- Criteria used to understand the quality and reliability of the data and its sources

Planning for Master Data Management within a domain:

- Identify candidate sources that will provide a comprehensive view of Master Data entities
- Develop rules for accurately matching and merging entity instances

- Establish an approach to identify and restore inappropriately matched and merged data
- Establish an approach to distribute trusted data systems across the enterprise

1.3.3.4 Master Data Management Key Processing steps

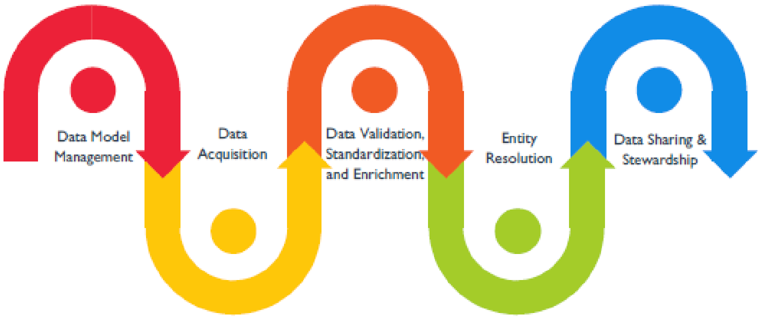
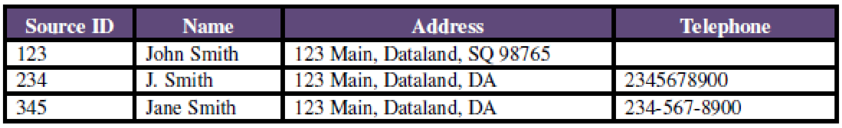

Data Model Management
Data Acquisition
Data Validation, Standardization and Enrichment
Entity Resolution
Data Sharing & Stewardship

Figure 76 Key Processing Steps for MDM

- **Data Model Management:** Clear and consistent definitions make sense to business at the enterprise level
- **Data Acquisition:** Data representing the same entity can look different. Plan for acquiring new data as a reliable repeatable process. High level cleansing tools, and matching rules, then perform data quality on the new data.

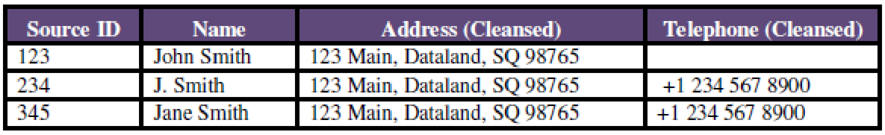
Table 24 Source Data as Received by the MDM System



Source ID	Name	Address	Telephone
123	John Smith	123 Main, Dataland, SQ 98765	
234	J. Smith	123 Main, Dataland, DA	2345678900
345	Jane Smith	123 Main, Dataland, DA	234-567-8900

- **Data Validation, Standardisation and Enrichment:** Reduce variation in format and reconcile values:
  - **Validation:** Identify clearly incorrect or defaulted data
  - **Standardisation:** Data conforms to standard Reference Data values and formats
  - **Enrichment:** Add attributes that improve entity resolution services

Table 25 Standardized and Enriched Input Data



Source ID	Name	Address (Cleansed)	Telephone (Cleansed)
123	John Smith	123 Main, Dataland, SQ 98765	
234	J. Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900
345	Jane Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900

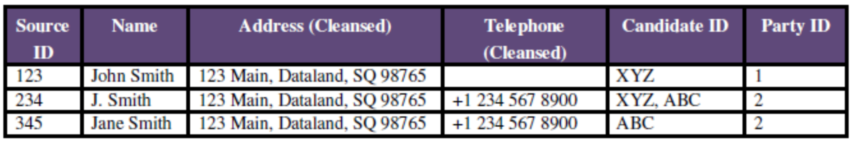
- **Entity Resolution and Identifier Management:**

  Entity resolution is the process of determining whether two references to real world objects refer to the same object or different objects.

  Activities:

  - **Matching:** or candidate identification is the process of identifying how different records relate to a single entity. Use similarity analysis to avoid false positives or negatives.
  - **Identity Resolution:** Keep a history of matches so that those less confident matches due to conflicting values can be undone if found to be incorrect.

Table 26 Candidate Identification and Identity Resolution



Source ID	Name	Address (Cleansed)	Telephone (Cleansed)	Candidate ID	Party ID
123	John Smith	123 Main, Dataland, SQ 98765		XYZ	1
234	J. Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900	XYZ, ABC	2
345	Jane Smith	123 Main, Dataland, SQ 98765	+1 234 567 8900	ABC	2

- **Matching Workflows / Reconciliation Types:** Different scenarios require different workflows:
  - **Duplicate identification match rules:** Specific set of data elements that uniquely identify an entity and identify merge opportunities
  - **Match-link rules:** Identify and cross-reference records that appear to relate to the master record without updating the content of the cross-referenced record
  - **Match-merge rules:** Match records and merge data into a single unified and comprehensive record. Complex.
- **Master Data ID Management:** Two types of identifiers managed in a MDM environment:
  - **Global ID** is the MDM solution assigned unique identifier attached to reconciled records. Should be automatically generated.
  - **X-Ref Management** is management of the relationship between source IDs and the Global ID.
- **Affiliation Management:** Establishing and maintaining relationships between Master Data records of entities that have real-world relationships. Data Architecture design of the MDM system which kind of relationship between entities:
  - **Affiliation relationships** are programmed and are the most flexible
  - **Parent-child relationships** have implied hierarchical navigation structure

- **Data Sharing and Stewardship:** Data Stewards resolve incorrectly matched situations and improve matching algorithms.

1.3.3.5 Party Master Data

Data about individuals, organisations and the role they play in business relationships. Examples from different environments:

- **Commercial:** customers, employees, vendors, partners, competitors
- **Public sector:** citizens
- **Law enforcement:** suspects, witnesses, victims
- **not for profit:** members, donors
- **healthcare:** patients, providers
- **Education:** students, faculty

Customer relationship Management (CRM) systems manage Master Data about customers

Master Data is challenging for parties playing more than one role in an organisation.

1.3.3.6 Financial Master Data

Data about business units, cost centres, profit centres, general ledger accounts, budgets, projections and projects. The central hub of financial Master Data is an Enterprise Resource Planning (ERP) system.

1.3.3.7 Legal Master Data

Data about contracts, regulations and other legal matters.

1.3.3.8 Product Master Data

Can focus on the organisation’s products and services or on industry-wide products and services. Different types of product Master Data solutions:

- **Product Lifecycle Management (PLM):** managing the lifecycle of a product/service from conception to disposal
- **Product Data Management (PDM):** engineering and manufacturing. Enables secure sharing of product information such as design drawings (CAD)
- **Product data in Enterprise Resource Planning (ERP):** SKUs to support order entry to inventory
- **Product data in Manufacturing Execution Systems (MES):** Raw inventory to finished goods
- **Product data in Customer Relationship Management (CRM):** Marketing, sales and support interactions

1.3.3.9 Location Master Data

The ability to track and share geographic information and to create hierarchical relationships based on geographic information.

- **Location Reference Data** is usually geopolitical data handled by external organisations
- **Location Master Data** are address related to parties and businesses

1.3.3.10 Industry Master Data – Reference Dictionaries

Authoritative listings of Master Data entries that can be purchased. They can provide a starting point for matching and linking new records.

1.3.4 Data Sharing Architecture

Each Master Data subject area usually has its own system of record e.g. CRM or ERP systems. Hub-and-spoke model for sharing Maser Data:

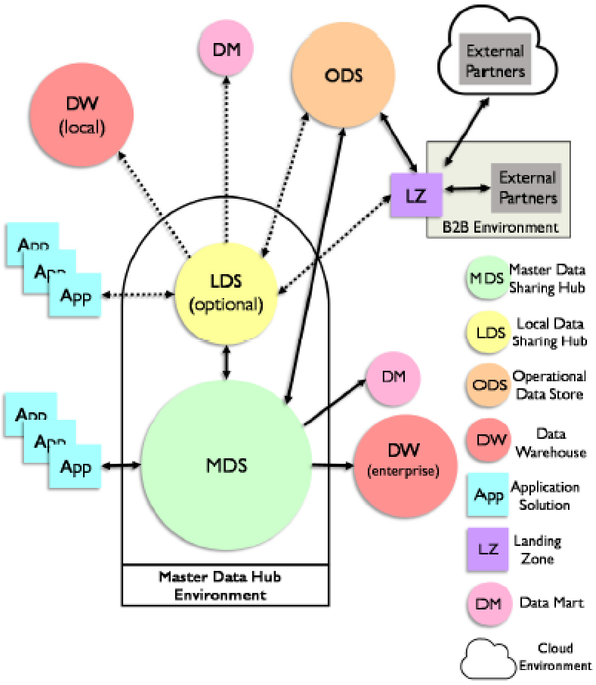


Figure 77 Master Data Sharing Architecture Example

Three approaches to implementing Master Data hub environment:

- **Registry:** An index that points to Master Data in various systems of record which manage Master Data local to their applications. Easy to implement, but complex queries are challenging. Multiple business rules need to be implemented.
- **Transaction Hub:** Applications interface with the hub to access and update Master Data which exists only within the Transaction Hub and not in the applications. The Hub is the system of record for Master Data. Better governance, but business rules reside in the Hub.
- **Consolidated approach**: Systems of record manage Master Data for their applications, and it is also consolidated within a common repository (replication). There is no need to access directly from the systems of record.

2 Activities

2.1 MDM Activities

- **Define MDM Drivers and Requirements:**
  - Easier to define requirements for an application than the whole enterprise
  - Prioritise Master Data efforts on cost benefit of the proposed improvements
  - Start with simplest category to learn from the process
- **Evaluate and Assess Data Sources**
  - Understand the structure of existing data and how it is collected or created
  - Understand the quality of the data, as poor quality data complicates a Master Data project
  - Assess disparity between sources
  - May be able to purchase standardised data such as Reference Directories
- **Define Architectural Approach:** Depends on business strategy, the platforms for existing sources and the lineage and volatility of the data.
- **Model Master Data:** As Master Data is an integration process, model data within subject areas. A logical or canonical model
- **Define Stewardship and Maintenance Processes:** Technical solutions still require the oversight of Data Stewards to address records that fall out of the process and why.
- **Establish Governance Policies to enforce use of Master Data:** The benefits come once people start using the Master Data

2.2 Reference Data Activities

- **Define Drivers and Requirements:**
  - **Drivers:** Operational efficiency and higher data quality
  - **Requirements:** Driven by the most important reference data sets
- **Evaluate and Assess Data Sources:**
  - **External:** vendor that guarantees updates on a schedule and ensures quality data
  - **Internal:** Owners should understand the benefits of central management of their data sets
- **Define Architectural Approach:** Tool should allow for manual updates
- **Model Reference Data sets:** Models help consumers understand the relationships within the reference data sets, and the data quality rules
- **Define Stewardship and Maintenance Processes:** Capture Metadata about reference data sets
- **Establish Reference Data Governance Policies:**

3 Tools and Techniques

MDM requires identity management enabled tools:

- Data integration tools
- Data remediation tools
- Operational data stores (ODS)
- Data sharing hubs (DSH)
- specialised MDM applications

4 Implementation Guidelines

As Master and Reference Data Management are forms of data integration, the same implementation principles that apply to Data Integration and Interoperability.

Implement incrementally through a series of projects defined in an implementation roadmap, prioritised on business needs and guided by an overall architecture.

It is vital to have data governance professionals who understand the challenges of RDM and MDM and can assess the maturity and ability of the organisation to meet them.

4.1 Adhere to Master Data Architecture

The integration process should take into account:

- The organisational structure of the business
- the number of distinct systems of record
- the data governance implementation
- The importance of access and latency of data values
- The number of consuming systems and applications

4.2 Monitor data movement

Monitor data as it flows within the Reference or Master Data sharing environment to:

- Show how data is used across the organisation
- Identify data lineage from/to administrative systems and applications
- Assist root cause analysis of issues
- Show effectiveness of data ingestion and consumption integration techniques
- Denote latency of data values from source systems through consumption
- Determine validity of business rules and transformations executed within integration components

4.3 Manage Reference Data change

Reference data is a shared resource, therefore it should not be locally controlled, but channels to receive and respond to change requests must be provided according to policies and procedures put in place by the Governance Council.

Planned/scheduled changes such as periodic updates to industry codes require less governance than ad hoc changes.

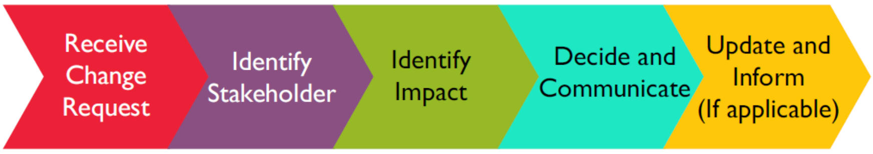


Figure 78 Reference Data Change Request Process

4.4 Data sharing agreements

Data sharing agreements stipulate what data can be shared and under what conditions. Helps when issues regarding quality of data brought in or availability arise. Driven by the Data Governance program and involves Data Architects, Data Providers, Data Stewards, Application Developers, Business Analysts, Compliance/Privacy Officers and Security Officers.

SLAs should be in place so that the quality data can be provided to downstream consumers.

5 Organisation and Cultural Change

It is not easy for people to relinquish control of their data to create shared resources. People may perceive MDM and RDM efforts as adding complications to their processes.

The most challenging cultural change is determining which individuals are accountable for which decisions.

6 Reference and Master Data Governance

Because they are shared resources, Reference and Master Data require governance and stewardship. Governance processes will determine:

- The data sources to be integrated
- Data quality rules to be enforced
- conditions of use rules
- Activities to be monitored and the frequency of monitoring
- Priority and response levels of stewardship efforts
- How information is represented to meet stakeholder needs
- Standard approval gates, expectations in RDM and MDM deployment

Governance processes bring compliance and legal stakeholders together with information consumers to ensure risks are mitigated through definition and incorporation of privacy, security and retention policies.

Data Governance must have the ability to review, receive and consider new requirements and changes. Make principles, rules and guidelines available to all using Reference and Master Data.

6.1 Metrics

- **Data quality and compliance:** DQ dashboards
- **Data change activity:**
    - Metrics denote rate of change of data values
    - Provide insight to systems providing data to sharing environment
    - Used to tune algorithms in MDM process
- **Data ingestion and consumption:** Denote and track data contributing systems and what business areas are subscribing to shared data

- **Service Level Agreements:** Level of adherence to SLAs provides insight to data and technical problems
- **Data Steward coverage:** Used to identify gaps in support
- **Total cost of Ownership:** Must be consistently applied across the organisation to be effective
- **Data sharing volume and usage:** The volume and velocity of data defined, ingested and subscribed to and from the data sharing environment.