Master Data Management

Concepts, Principles, Standards and Guidelines

Authors:	Ben Stevens, Lead Data Architect, Enterprise Data Annie South, Director Data Management CoE, Enterprise Data Mike Stainer, Enterprise Architect, oCTO, ITS Henrique Delamanha Mendonca, Technology Strategist, oCTO, ITS
Version:	1.0
Date:	April 2020
Status:	Final
Issued:	11 November 2020
Location:	Policy Register
Approval Authority:	Master Data Management Steering Committee Information Governance Board

1 Definitions and Key Concepts of Master Data

1.1 What Is Master Data?

Master Data is information that is common to different business areas and is used in multiple business processes. Master Data usually describes things that are participating in a transaction or event. Examples include information about a Course, a Student, or an Employee.

Reference Data is information that is generally used to categorize other data (e.g. status codes), or related to information beyond the boundaries of the organisation (e.g. lists of countries)

By contrast, **Transaction Data** describes an event. The information about the event doesn't change once the event has taken place. However, Master Data is prone to change over time. For example, a staff member may change their surname, or a student may move to a different address.

The re-use of Master Data in different processes, alongside the slowly changing nature of the data itself presents challenges for the management of Master Data.

Master Data Management (MDM) is an approach for ensuring the organisation's Master Data is captured, maintained and referenced in a consistent, uniform and accurate state, and available across all processes and systems where it is needed.

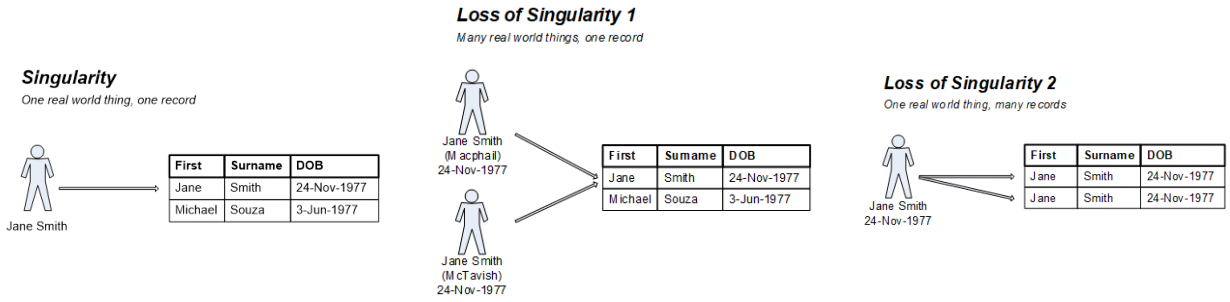
Not all data needs mastering. Mastering efforts focus first on data which is widely shared and tends to originate from outside the organisation rather than concepts *internal* to the university.

Customer Data – Different layers of data have different mastering priorities

1.2 Key Concepts of Master Data Management

Master Data Management relies on the ability identify key areas of data that need to be mastered, and to declare and maintain discipline around the established "master" source of the information for these areas. Typically, these key areas of data are described using data models and business glossaries.

Singularity refers to the state of having a single record for each real world "thing" we wish to master information about. This means keeping data in alignment with the real world.



Singularity is supported by using a **unique identifier**. A **Unique Identifier** is a number which is used to represent a real world thing uniquely in a set of data. For example, a staff number or a student number.

The act of issuing a unique identifier becomes a key governance point for Master Data Management. The process that creates (or issues) a unique identifier is called the **Point of Origin** (or Registration Authority). Importantly, this refers to the *process* of creating the information, not the *record* (or data) that results from that process, nor *IT system(s)* that enable the process(es).

A **Master Record** refers to the designated, authoritative version of mastered information. This master record is created or updated by the designated point of origin process. The **System of Record** is the business system that enables the point of origin processes, issues the unique identifiers, applies processes that increase the likelihood of singularity and holds the master records.

A **Secondary System** is a system that needs to have copy of mastered data in order to operate but is not responsible itself for the mastering process of that data. It's important that secondary systems are not able to over-write or record master information or this may result in value collisions.

A **Value Collision** is where the same piece of information about a mastered entity is captured in different places, and the values recorded are not the same. For example, capturing "Jim Smith" and "James Smith" as someone's full name. From a data perspective, value collisions are difficult to recover from – we can never be sure which of these values is the preferred, accurate or appropriate value to default to.

The **Golden Record** is a record that has been recreated from different sets of information because of mastering challenges. A Golden Record is used in the place of a master record, but it takes more effort to create and is less accurate.

The creation of a golden record requires decisions to be made such as:

- Which value to select for the golden record when value collisions have occurred;
- Whether to merge two or more data records when singularity has been lost; and
- Which attributes to source from different places when data has become fragmented and must be sourced from different places.

All the responsibilities of a System of Record fall on the processes that create the Golden Record.

2 Principles of Master Data Management

Id	Principle	Description
1	RMIT's processes are designed to ensure master data is accurate and consistent.	The use of master data in business processes results in customer or business outcomes that should be based on accurate and consistent data. Processes must ensure that the designated system of record for impacted master data are updated to ensure consistency. Data matching cannot substitute for poor process
2	Master data must have an identified single business owner.	Business may decide at what level ownership applies, e.g. entity or even attribute level. Data owners are accountable for ensuring the data they are responsible for is handled in a way that assures accuracy and consistency.
3	Master data will have an identified system of record.	The data should have an acknowledged 'home'. How and where data is mastered may be by logical entity (e.g. customer, supplier) or by logical role (e.g. lead, student, alumni). Not all attributes related to that entity require to be held in the same System of Record.
4	Master data must only be created or updated by a single, designated System of Record.	Where other systems require local copies of mastered entities, they must source this data from the relevant master, and not from any other intermediary system unless by exception. Updates to core attributes are performed on the master and disseminated. Updates must not be made to the copied master data. If a copy of the data is taken, then the system holding that copy has a responsibility for ensuring the copy held remains synchronised with the master of the data.
5	Copied data must include the Unique Identifier issued by the System of Record, and remain synchronised with the values in the Master Record.	RMIT will ensure unique identifiers exist for all master data. The unique identifier has been mastered to provide a universal key back to a real-world thing. Populating mastered unique identifiers across systems is an important way of ensuring cross-system data integrity. Without it, data integrity is hard to achieve. Not all attributes are required to be retained in the copied data.
6	Master data quality is tracked and managed.	If an entity is worth mastering, it is worth ensuring the data is fit for purpose. This includes: • Monitoring the master for quality issues, and • Monitoring the alignment of the mastered attributes across secondary systems.
7	Mastered entities will be declared using business definitions and business data models.	Business data models and business definitions can be read and understood by business, process, data and technology specialists alike.
8	Analytics environments must not be used as a point of origin for master data.	Analytics environments are subject to data standardisation, consolidation and enrichment activities, and may generate derived attributes. However, these data enrichments are regarded as a different class of data and have their own requirements for governance and management.

Id	Principle	Description
9	It is not necessary for all attributes linked to a critical entity to be mastered, nor mastered in the same system.	See Types of Customer Information. 'Core attributes' are mastered. 'Extended attributes' are candidates for mastering. 'Derived attributes' are not mastered, rather the definition for deriving the attribute should be mastered. 'Related data' may be mastered entities in their own right, hence are not mastered as part of the entity of focus. 'External captured data' is copied from the external systems where it should remain mastered.

3 Standards and Guidelines

3.1 Governance

A governance structure has been implemented to guide the development & operation of RMIT's MDM framework. The governance includes a Steering Committee (SC) and a Working Group (WG).

The Steering Committee sets clear ownership and accountability for master data, establishing a hierarchy of responsibilities that provides and enhances the quality of master data. The Steering Committee also provides overall vision, strategy and guidance.

The Working Group will:

- Define and communicate the principles, processes and controls for master data;
- Review formulated deliverables to achieve master data approach aligned to enterprise data strategy;
- Investigate and recommend master data related solutions to address business pain points and ensure alignment to the enterprise data strategy;
- Investigate and recommend solutions to major master data issues;
- Participate in or provide guidance to projects through the university that impact master data;
- Escalate key topics to IGB for validation / approval; and
- Report on master data metrics.

3.2 Data Standards

The Steering Committee confirms the data entities that are to be mastered.

The Working Group ensures the entities are defined in business terms, including all fields / attributes and acceptable business rules and values for those attributes. These items are then included in the Organisation Logical Data Model.

The Working Group identifies the metrics that can be used to measure the quality of the data. This covers both the data within the System of Record, and across the System of Record and all of the systems that retain a copy of the mastered data.

Reporting is to be implemented to provide a regular view of the data quality of the mastered data entities to the nominated owner.

3.3 Processes

Processes are to be developed in a consistent, standardised model for all mastered data. The processes are aligned to the business processes, but enhanced to recognise the importance of the data, and the mastery of the data in the System or Record. The processes are to be documented, including the process model, descriptions and flow diagrams.

The processes are to cover:

- **Data integration:** The implementation of the System of Record and interfaces to each of the systems requiring an interaction with the mastered data
- **Data maintenance:** The creation, querying, updating, deleting, merging and unmerging of the mastered data, including the impacts to all systems that retain a copy of the data.
- **Data quality assurance and control:** Processes to identify, action and report on data quality issues are required. These processes should leverage on the metrics identified as part of the data standards. The metrics should assist in identifying areas where the processes for data maintenance are failing to meet the expected data quality standards, allowing for the processes to be reviewed and strengthened.
- **Archiving:** A process for removing records from the mastered data set are required. This is likely to require the records be archived for future reference. The archiving should occur not just in the System of Record but be coordinated across all systems that interact with the mastered data (e.g. if a student record is archived, then interaction records with that student should also be removed / archived).

3.4 Organisation

Clear operational roles, tasks and responsibilities are to be defined, aligned to the processes identified above. Resourcing levels are to be committed as part of the move to mastering of the key data.

3.5 Technology

There are 4 common technical approaches for implementing MDM:

- Registry
  A central register is formed with the minimum data required to identify duplicate records in the master data set. Changes to the master data continue to be made in the existing source systems without integration from the register to the source systems. Updates across the sources systems are either handled through manual process or other interface techniques.
- Consolidation
  This approach involves consolidating data from multiple sources in a central hub. The data is matched and analysed for quality, resulting in a 'golden record'. The approach is generally used for reporting and analysis.
- Coexistence
  Similar to the consolidated approach, the coexistence approach takes data feeds to create a 'golden record' which is then stored centrally. Any changes detected from the source systems can be fed back across all source systems.
- Transaction / Centralised
  The mastered data is stored in a central data store. All creations and updates occur through the central data store, with the updates fed back into the systems that require access to that data.

To support any of the approaches above, mapping must be provided between the physical representation of the logical data model in the master data system and the physical representation held within the sources of the data.

3.6 Systems of Record

3.6.1.1 A system of record must provide generic registration functionality.

Generic registration functionality refers to the ability of a system or application suite to be able to register, maintain and distribute the shared information required for a particular business entity.

The principle: "capture information once, at its source" has been widely circulated and accepted. In reality, this ideal has proven difficult to attain in the absence of basic system functionality to support registration processes, and open sharing of the information.

3.6.1.2 A system supporting the process of registration should be fit-for-purpose.

In general, a system supporting the process of data registration should provide data services that:

- state whether a given unique identifier exists in the registration set. For example, answer the question "Does Student Number 12345678 already exist?"
- given the value of a unique identifier, provide its associated attributes, on request. For example, answer the question "What is the First Name, Surname and Date of Birth of Student Number 12345678"?
- stipulate the base set(s) of attributes that can be used as natural keys, ie what minimum set of fields can be accepted as the minimum information required to register a new entity.
- given values for a base set of attributes, verify whether a record currently exists for the real world thing. For example, answer the question "Do you have a John Smith (24 Nov 1972) already registered as a student?"
- given values for any set of attributes, return all records that match the attribute values. For example, answer the question "Tell me all the John Smiths you do have registered?"
- given information comprising a new natural key, create a new record, and issue a new unique identifier for the record. For example, "I know I have a new student called John Smith (24 Nov 1972). Please issue a new Student Number."
- enable remediation of duplicate records for the same real-world entity (merge).
- enable remediation of a single record found to represent two real-world entities (split)

3.6.1.3 A system of record should provide complete and exclusive coverage.

A system supporting the process of data registration should:

- provide **complete** registration coverage of a nominated entity (all possible instances of the entity should be able to be registered in the system). So, for example if the system were registering Student, it would be important that all types of students could be accommodated within the registration process. This might mean overseas students, as well as domestic. Another example might be charities and government instrumentalities, as well as corporations as industry partners.
- provide **exclusive** registration coverage of a nominated entity. That is, there should be no other systems in the enterprise that also register this kind of real-world thing.
  In the context of mergers and acquisitions, this ideal may be difficult to achieve, but in its absence, the registration authority should instead provide the ability to store, hold and publish mappings (translations) between legacy identifiers and the global identifier.

3.6.1.4 A system of record should have appropriate business processes.

A registration authority should have business processes that:

- are highly accurate in their ability to distinguishing between new and existing instances of the real-world thing
- provide for after-the-fact harmonisation to remediate the case where two real-world things have mistakenly been issued with a single data record
- provide for after-the-fact harmonisation to remediate the case where two data records have been created for the same real-world thing

3.6.1.5 Mastering processes should be continuously improved.

Best practice in this area has consistently identified the need for constant re-assessment of registration processes, and information quality analysis. Business processes supporting registration should be sensitive to the business events that cause changes to the underlying master data. Business processes should strive to use every opportunity to capture updated information.

3.6.2 Unique Identifier Requirements

3.6.2.1 The value of unique identifier must remain fixed for the life of the record it is issued to.

To enforce singularity, the unique identifier is essentially the representative of the real-world entity in the data-world – we strive to attain a one-to-one relationship between the real world and the data record. The life of the unique identifier should parallel the life of the real-world entity. If the value of the unique identifier changes, it implies that the corresponding real-world thing has also changed into someone/something else, which is not ideal.

3.6.2.2 The value of a unique identifier must be unique within its set.

To enforce singularity, a unique identifier must be unique to each distinct record (as the name implies). Otherwise ambiguity is introduced, and the identifier is not able to accurately identify which record is being referred to.

3.6.2.3 Once issued to a record, the value of unique identifier must not be re-issued to another record.

Re-issuing unique identifiers in a heterogeneous technology environment leads to sequencing problems, and confusion about the underlying entity purported to be identified.

3.6.2.4 A format chosen for a unique identifier must make enough values available to be issued to support all conceivable instances of the entity in the foreseeable future.

The format of a unique identifier will generally represent a limitation on the number of unique values that can be issued. For example, if a 4 digit numeric value were chosen to represent the Customer Unique Identifier, it would limit the number of customers we could represent to 10,000. The consequences of a poorly chosen format place artificial limits on business operations, and it undermines the ability for the unique identifier to identify instances uniquely and unambiguously. The re-issuing of unique identifiers is an extremely painful process in any organisation to go through, and should be done as infrequently as possible.
The decision regarding "requirements for the foreseeable future" should be an architecture decision, based on the best information available regarding the strategic vision of the organisation.

3.6.2.5 A unique identifier should be free from additional meaning.

The purpose of a unique identifier is to uniquely identify a record, for the life of that record. An example of poor practice would be to encode the first digit of a unique identifier, so that a '2' means 'NSW', a '3' means 'Victoria' etc.
When a unique identifier is overloaded with additional meaning, it becomes less able to do its job as a unique identifier. People begin to make assumptions about the meaning of a unique identifier, which can create confusion about the meaning of the entity itself. If the underlying attributes change (for example, if a person moves from NSW to Victoria), then either the value of the unique identifier needs to change, or the unique identifier no longer accurately represents the attribute.
A unique identifier should not be a name, as these are not linguistically neutral (and hence convey meaning, which is subject to change over the course of time).