**Why Data Governance Matters: Master Data Management (MDM)**

What is Master Data Management?



Photo by [Khamkéo Vilaysing](https://unsplash.com/@mahkeo?utm_source=medium&utm_medium=referral" \t "_blank) on [Unsplash](https://unsplash.com/?utm_source=medium&utm_medium=referral" \t "_blank)

*This is part of a series about*[*Data Governance*](https://towardsdatascience.com/why-data-governance-matters-5385d722c4c6)*and its relevance. Master data management (MDM) is typically considered to be a slice of data governance.*

**Table of Contents**

* What is Master Data?
* What is Master Data Management?
* What problem does MDM solve?
* Why is duplicate data bad?
* How does MDM work?
* Conclusion

**What is Master Data?**

Let’s start with understanding **master data**. It is, after all, what we are aiming to manage.

A definition on [Wikipedia](https://en.wikipedia.org/wiki/Master_data): *Master Data represents the business objects that contain the most valuable, agreed upon information shared across an organization.*[1]

Another one: *Data about the business entities that provide context for business transactions.* [1]

A bit vague right? It’s a tricky term to define. Below are a few examples of data considered “master”:

* Customers
* Products
* Services
* Vendors

If you don’t get it, don’t worry, it will make sense as we move on.

**What is Master Data Management?**

In my own words, I would describe MDM as “trying to make sure master data is distinct and has minimal quality issues”

Wikipedia: *Master data management (MDM) is a technology-enabled discipline in which business and information technology work together to ensure the uniformity, accuracy, stewardship, semantic consistency and accountability of the enterprise’s official shared master data assets. [2]*

I think the only way to truly understand MDM is by first understanding the issue it aims to aims to resolve. But hopefully these definitions at least vaguely set the stage.

**What problem does MDM solve?**

MDM aims to resolve **duplicate master data.**

The main idea here is there are duplicate records of master data across the enterprise. Examples include:

* Multiple instances of the same customer with different addresses and phone numbers, etc.
* Multiple versions of products sold with the same name but different attributes. Or worse, multiple products with slightly different-worded names with the same attributes
* Multiple Vendors with the same name but different attributes

Let’s look at duplicate customer examples at a fictional Auto Repair shop, “Car Repair ABC”.

* Ben has an existing account with Car Repair ABC, but he opens up another membership with a different email accidentally. This leads to two instances of Ben in their system. *This is an example of duplicate records being created due to an internal control limitation.*
* The advertising department of Company ABC has its own database which contains both current customers and prospective customers. This database does not have a 1:1 match with the customer records of sales, accounting, etc. *This is an example of duplicate records existing due to siloed / segmented source system processes within an enterprise.*
* Merger and acquisition: Car Repair ABC purchases a smaller auto shop. If Sally has accounts at both stores pre merger, there will likely be multiple instances of Sally post merger. *This is an example of M&A-related duplicates.*

These are a few basic examples of duplicate data. It’s important to remember this issue isn’t solely for customer data. The same principles can apply to the products, services, employees, or any other [master data](https://en.wikipedia.org/wiki/Master_data).

**Why is duplicate data bad?**

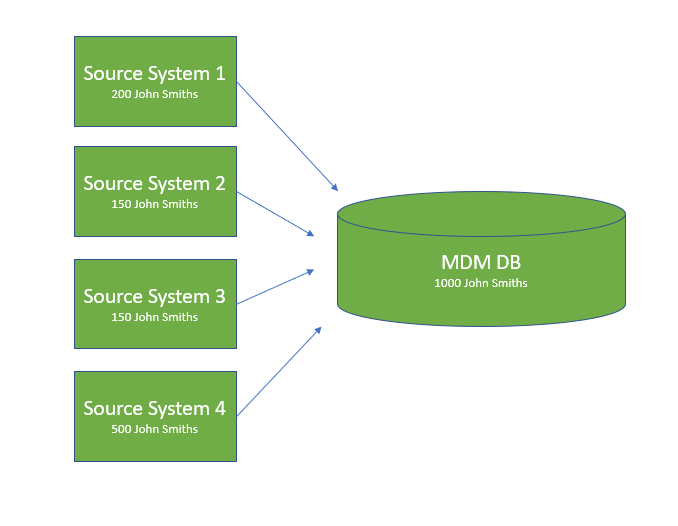
With all this duplicate data, it increasingly makes it difficult to know what is true and what is not. Who are our customers exactly? What are our products?

*“Duplicate master data and missed opportunities for revenue, margin, and efficiency have been linked in case study after case study… In a traditional architecture environment, data is not integrated so that a customer is uniquely identified and consistently described across all applications, or that a product is tracked in uniform throughout the company. An organization that lacks such data cohesion cannot possibly recognize the total value of their business ” — Val Lovicz*[3]

Additionally, without a foundation of governed master data to work with, data scientists are limited in their ability to draw insights or create models. What kind of value does a customer regression model bring if duplicates are making our conclusions inaccurate?

**How does MDM work?**

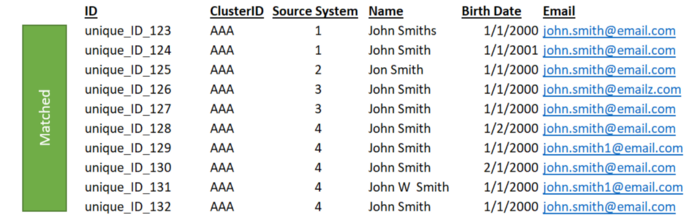
The first step is to create unique master records, a single source of truth, in a new location. Let’s say there are 1000 instances of ‘John Smith’ across all of the various enterprise source systems. How many unique ‘John Smiths’ are there and how many are duplicates? This is what MDM is aiming to solve!



High level visual of what the MDM ETL process looks like — centralizing all instances of a specific type of master data into one place. (Image from author)

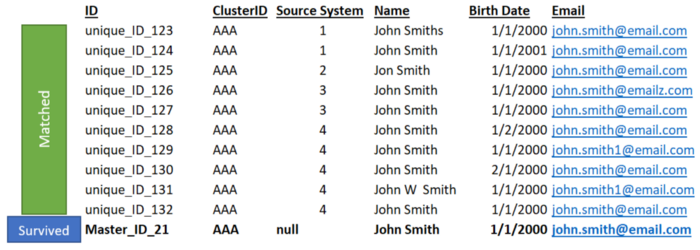
After compiling all the source records into a single database with unique IDs for each instance, there are 2 steps that occur: Matching and Survivorship.

**Matching**: The basic concept here is there’s an algorithm which goes though all of the records (in our case, the 1000 John Smiths). It will then create clusters of John Smiths that it believes are the same person. These clusters are determined through a logical series of matching rules which prioritize certain shared attributes above others. For example, the rules may prioritize a match on birth date over a match on email.



The matching algorithm has determined that these 10 instances of John Smith (from our mdm db of 1000 instances) are all duplicates of the same individual (Image from author)

**Survivorship**: After there are clusters of John Smith’s formed via the matching process, the next step is to select the attributes shared by each cluster that are the strongest and “survive” those into a single master record.



The new “survived” golden record at the bottom has pulled the strongest instances across all records, within the cluster, to form the most accurate, unique instance of this specific ‘John Smith’ (Image from author)

Even though a few of the instances are not consistent (for example, the name ‘Jon Smith’ or the email ‘john.smith@emailz.com’), the strongest attributes are survived into the bolded master record at the bottom. The ‘strongest’ attributes are shared by the most instances within each cluster, or are determined by more nuanced logical rules in the survivorship configurations.

**Conclusion**

MDM is a vital slice of the data governance pie, and it is used in virtually every large enterprise. It’s an essential concept for data analysts/scientists to know as they navigate various data landscapes in their careers.

I would recommend checking out the article listed in the sources below, “Beyond Matching” by Val Lovitz, for additional insight into how matching works with more thorough examples.

Hopefully this article has added value to anyone who made it to the end. Cheers!