CHAPTER 9

Data Quality Management

This chapter covers the discipline and practice of Data Quality Management (DQM) in a multi-domain Master Data Management (MDM) framework. It discusses how to define and apply a DQM model consistently across a multi-domain environment, and how to expand the practice of DQM to an enterprise level. This chapter starts by presenting how DQM fits into a multi-domain MDM environment and how important it is to manage DQM strategies, decisions, and execution. It continues by introducing a DQM model that is critical to supporting and scaling the discipline beyond MDM. Finally, it covers a data-quality improvement life cycle, with the required steps and activities to analyze, plan, execute, and monitor data quality projects effectively and efficiently.

DQM in a Multi-domain MDM Environment

Trusted data delivered in a timely manner is any company's ultimate objective. No company will ever question the need for high-quality information. The major challenge is to establish a model that can both minimize bad data from being introduced and efficiently correct any existing bad data. Data quality requires both strategic and tactical focuses in addition to strong business and information technology (IT) collaboration. Strategically, a company needs to start by establishing a culture of quality. Data quality is everyone's job, but it's up to management to communicate that view, and most important, establish the proper foundation and channels to help people succeed. Tactically, organizations need to identify and assess areas in need of data-quality improvement, and conduct data-quality projects to both fix issues and prevent them from happening in the future.

Let's discuss data quality in the context of a multi-domain MDM implementation. One of the selling points of MDM is that it helps improve the quality of data. But why is MDM so conducive to data-quality improvement? The answer is simple: first, MDM requires data to be consistent across multiple sources so that proper linkage is established; second, MDM fosters the assembly of fragmented information across disparate sources; and third, MDM becomes a central point to maintain, improve, measure, and govern the quality of master data. All those three points have a data-quality focus, which forces companies to dig deep into their data by performing meticulous analysis, and once issues are identified, to perform the proper correction.

Companies will have to address data-quality concerns during an MDM implementation due to the intrinsic problem that MDM addresses. It is important, post-MDM implementation, not to lose that momentum and use it to establish a solid foundation to continue a sound data-quality program even after one or more domains are completely integrated into an MDM hub. Furthermore, it is typical to gradually master different domains in multiple phases. As such, there is a clear advantage in increasing the maturity of DQM to shorten future phases. Even a single domain can be implemented in multiple phases due to a large number of sources to integrate. Any data-quality methodology established and knowledge acquired should be clearly documented for future reference and reuse.

DQM is very broad and deep. Certain aspects of DQM can easily be re-used across many different domains, but others not so much. For example, validating customer data can be quite different from validating product data. Still, many techniques of data profiling and data cleansing can be applied similarly across multiple domains. Solid leadership and expertise in DQM are required to differentiate when to generalize and when to specialize data-quality components, processes, and methodologies for multiple domains.

A Data-Quality Model

There are many factors influencing how a company will address data quality, such as the multi-domain MDM approach being implemented; how multiple lines of business (LOBs) interact with each other; the level of collaboration between business and IT; the maturity level of data governance, data stewardship, and metadata management; the degree of high-level management engagement and sponsorship; technological resources; and personnel skills.

Data-quality efforts cannot be isolated into a corner or managed like a software development team. Data-quality specialists must be engaged into other activities within the company, and be actively involved into data-related decision making subjects. Data-quality experts should be assigned to projects, working side by side with data architects, data designers, system integrators, software developers, data stewards, and so on.

Data quality can be managed within its own Project Management Office (PMO), as a function of the data governance PMO, or as a function of the MDM PMO. Still, these practices need tight strategic alignment to avoid charter and priority conflicts. Furthermore, in spite of the specific model chosen, the data governance PMO should be highly engaged to DQM for proper strategic directions and oversight. Figure 9.1 depicts a model for DQM. Notice how it shows a separate DQ PMO, but it also indicates the need for alignment with the data governance and the multi-domain MDM PMOs. The separate DQ PMO is presented to allow for explanation on specific DQM activities and functions.

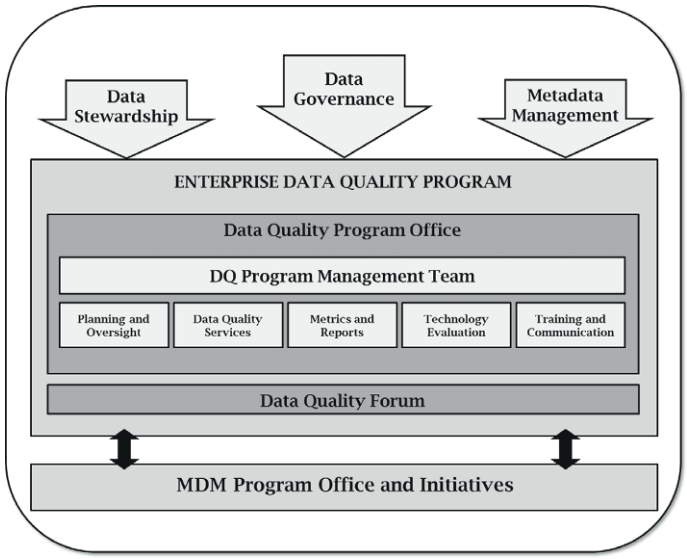

Figure 9.1: A DQM model

Planning and Oversight

A data-quality program is ongoing. There will be many data-quality improvement projects, but the overall goal is to have a strategically sustainable office to direct and foment initiatives. Strong data-quality leadership needs to work closely with a data governance office to align business needs and continuing prioritization of efforts. A data-driven company needs to encourage everyone to detect issues and propose data-quality improvements. Of course, any proposed initiative will have to be analyzed, reviewed, and prioritized. A data quality office can oversee the analysis of data issues, evaluate the impact that it has on the business, propose solutions to cleanse existing data, and either mitigate or prevent future occurrences.

In multi-domain MDM, it is natural to have certain data-quality requirements and activities already presumed since they are intrinsic to consolidating and synchronizing master data, as explained earlier in this chapter. Therefore, the data-quality roadmap is somewhat embedded to the overall MDM roadmap. Although MDM provides the opportunity to start a data-quality practice, DQM needs to be built as a strong practice with well-supported capabilities and leadership. To be sure, they need to be collaborating functions. But if companies do not already have a data-quality program, they need to take this opportunity to start one and expand the role of data quality beyond master data.

Data-Quality Services

Data-quality services are the primary reason for the existence of a data-quality program: the delivery of actual data-quality improvements and a constant drive toward a more mature and

proactive organization. The ideal way to solve a data-quality issue is by avoiding it from occurring. But even mature companies will have ongoing data-quality issues that require reactive processes in place to solve them. Later in this chapter, we will cover data-quality activities in detail as part of a data-quality improvement cycle.

In any event, in addition to clear requirements and goals, a successful data-quality program also needs to set the right expectations through a well-communicated service-level agreement (SLA). Be sure not to overstate your deliveries, as it can lead to a lack of confidence on the entire program. Data-quality improvement is harder to achieve than people imagine, because it is both a technical and a business issue. Affected business teams must be involved at all times to evaluate whether the delivery meets their needs. Very often, rules and regulations may limit data updates, which can hinder data fixes. In some cases, one or more corporate regulatory offices, such as legal, compliance, or privacy, must get involved in approving corrections. These are usually slow-moving processes that will extend delivery time.

Remember that with MDM, the same master data is shared among multiple LOBs. That is a double-edged sword. On the one hand, data-quality correction only needs to occur in one location, which is then propagated to other sources consistently. On the other hand, data changes must be accepted and approved by all affected business units. Data governance can facilitate the approval process, although a data quality forum is also a viable alternative to engage business liaisons along with data-quality experts to expedite analysis, testing, and acceptance of proposed modifications.

Metrics and Reports

Metrics and reports are about measuring and reporting the overall performance of the data-quality program itself, not actual metrics and reports on data-quality issues, which is considered a data-quality activity in this book, and is explained later. A data-quality program must have measurable goals with meaningful key performance indicators (KPIs) and well-established return on investment (ROI). As such, those goals must be measured and reported to demonstrate the value of the data-quality program as a whole.

As stated previously under planning and oversight, a complete data-quality roadmap may or may not be included in the overall MDM roadmap. It is more advantageous for a company to expand the scope of data quality beyond MDM. Programs and projects can have various forms and levels of DQ metrics, either within MDM or not. Still, DQM standards should be approved by data governance to insure proper reporting alignment and use across multiple programs.

Technology Evaluation

Tools for data quality are constantly evolving. In addition, do not expect a vendor-provided MDM solution to have all the technology needed for a comprehensive DQM program.

They will usually provide what is required for MDM-related data quality, which is just one portion of all data-quality activities. There are many facets to data quality, which may require multiple tools and multiple skills. The many activities around data quality are covered in the next section. The data-quality program office should spend a reasonable amount of time evaluating new technologies and staying abreast of new capabilities that can help prevent data issues and improve competencies to profile data and expedite data-cleansing initiatives.

Furthermore, using reference data to either verify or correct other data is typically a good strategy. If a trusted source exists for a particular subject, it should be explored as part of a data-quality effort. Therefore, it is important that reference data management and DQM collaborate as part of an overall data management organization. In the context of multi-domain MDM, certain domains will offer more or less options for reference data depending on how general the domains are. For example, depending on how specific a product is, it can be much more difficult to find reference data related to the *Product* domain than to the *Customer* domain.

Training and Communication

Data quality is everyone's responsibility, and that message must be communicated clearly in a truly data-driven enterprise. Just about everyone in the company will recognize data-quality issues do exist; however, no one person or business group will typically understand the extent of those issues, their cost, how the issues affect their current decisions, and how to fix them. Therefore, simply communicating how important it is to have high-quality information is not sufficient. It is necessary to present tangible information and techniques that can be converted to actions to detect, assess, and improve the overall quality of the data.

Data issues are typically identified on a singular basis. A certain user notices an issue to a certain record while interacting with a customer or partner, or while performing a given business process. The instinctive action is to correct that one record and move on. A truly proactive company should have mechanisms for this user to report the problem, which is analyzed to identify whether it is indeed a single occurrence or it is a systemic problem that must be addressed. This certainly requires proper infrastructure, but awareness and proper training are key elements to make it successful.

Technology alone is not enough to prevent and correct issues. It requires people to be sensitive to issues and how their actions have the potential to negatively affect downstream systems and processes. This is as much of a business issue as it is an IT issue. It is a business issue because data is mostly entered and maintained by business teams. It is an IT issue as well because, as custodians of the data, IT teams are responsible for providing a solid infrastructure that minimizes data anomalies according to how data is used.

In a multi-domain MDM environment, bad data can be magnified. MDM is about eliminating data siloes by bringing disparate sources together, eliminating inconsistencies and

fragmentation, and finally distributing golden records throughout. That is obviously fine, so long as data coming into an MDM hub are either good, or just bad enough that they can be fixed. However, if the data are of such low quality that cannot be fixed at all, or its low quality can't be automatically detected, the MDM hub will distribute the bad information to the enterprise. That is why the functions of stewardship, quality, and governance are even more important. The consequences of bad data entered into what was previously a siloed system are now exacerbated, making the need for proper training and communication even more relevant to avoiding negative consequences.

Continuous training, both formal and informal, is essential to achieve everyone's participation and strengthen a culture of focus on data quality. Actually, several studies have shown that informal learning can be more effective than formal learning. With that in mind, companies need to find creative ways to disseminate DQM information, such as with mentoring and coaching programs, brown-bag sessions, reviews of lessons learned, and so on. Technology should be leveraged to increase collaboration. Social media still has a long way to go in the workplace, but that channel needs to be considered as a mechanism to increase collaboration and promote information sharing. There are multiple categories of social media applications, such as blogs, microblogging, social networking, wikis, webcasts, podcasts, and more. Balancing what resources to use and to what extent is a challenge. Companies in certain industries may have less difficulty in adopting some of those applications. As an example, it is likely that high-tech companies are more prepared than healthcare companies to embrace and spread the use of social media in general. When defining what will be most effective in a company, it is necessary to take into consideration the company culture, computer resources, and human resources and skills.

Data-Quality Improvement Cycle and Activities

There are many aspects of DQM, including the source of bad data, as well as the actual definition of what bad data really is and its associated representation. Let's take a look at a couple of examples to illustrate this further:

- There is no contention about what the two acceptable values are for the gender attribute. However, one system may represent it as M/F, another with 1/0, another with Male/Female/MALE/FEMALE, and yet another without any validation whatsoever—with a multitude of manually entered values that could be missing, correct, or incorrect. Furthermore, it is possible to have a correct gender value improperly assigned to a given person. In the end, this information can be critical to companies selling gender-specific products. Others, however, may not be affected as much as if a direct-mail letter is incorrectly labeled Mr. instead of Mrs.
- Some data elements may not even have an obvious definition, or its definition depends on another element. An expiration date, for example, has to be a valid date in the calendar, as well as a "later than" effective date.

- In another scenario, some customers are eligible for a certain service discount only if they have a gold account.

These examples show just one facet of data quality or lack of it. Data suffers from multiple problems, including fragmentation, duplication, business rule violation, lack of standardization, incompleteness, categorization, cataloging, synchronization, missing lineage, and deficient metadata documentation.

One may wonder why companies get in such a mess. Difficult situations usually are caused by a multitude of factors, some potentially more avoidable than others. Certain companies grow at an incredible and sometimes unpredictable pace. Mergers and acquisitions are very common vehicles to increase market share or to tap into new business endeavors. Every time a new company is acquired by another, its data must be integrated. That can translate to more data-quality issues and defects being injected into the integrated environment. Companies usually do not have time to cleanse the new data coming in as part of the migration process, except when it is absolutely required to make them fit into the existing structure. "We'll cleanse the data later" is a common motto and rarely achieved.

Additionally, software applications have historically been developed to solve a particular aspect of the business problem and rarely consider the impact to other business processes or data consumers. That has led to years of multiple distributed software and business applications with disparate rules. Different systems might contain multiple instances of a customer record with different details and transactions linked to it. Because of these reasons, most companies suffer from fragmented and inconsistent data. The net effect is that companies face unnecessary and increased operational inefficiencies, inconsistent or inaccurate reporting, and ultimately incorrect business decisions.

Even enterprise applications, such as ERP and CRM, have remained silos of information, with data being constantly duplicated and laden with errors.

There is also the business process as part of the equation. It is human nature for people to look for creative ways to solve their problems. That means, when users have technical problems or run into business limitations during data entry, they will find ways to do it, even if it means breaking business rules or overriding well-defined processes. From a data-quality perspective, this is not a good thing, but should the users be blamed? After all, they may be facing a particular customer need that doesn't fit well into an existing business process or lacks sufficient business rules, or there is a system defect that is delaying a high-profit transaction requiring the user to take alternate actions that can lead to data capture issues.

Let's assume a company does have all the correct elements in place, such as data governance, data stewardship, data quality, IT support, and so on. Users are less likely to engage the proper teams if their confidence on the support process is low. They may think: "by the time I get this problem resolved through the proper mechanisms, I'll have a customer satisfaction

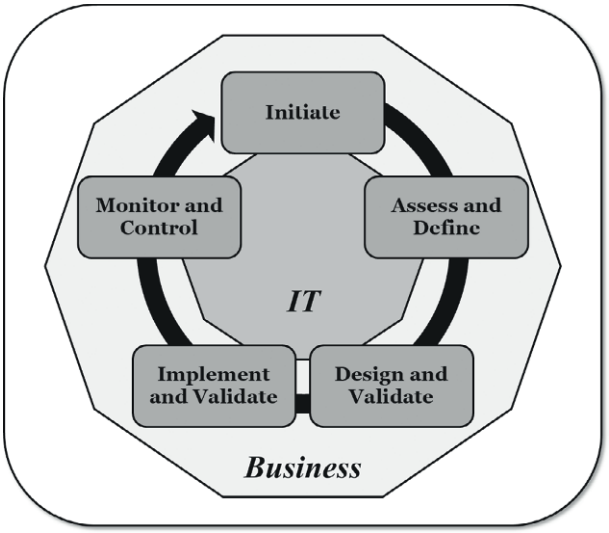

Figure 9.2: Data quality improvement cycle

issue beyond repair." Therefore, for the benefit of the company, they act with imagination and independently solve the immediate problem to their satisfaction and possibly with nonstandard solutions and non approved data entry. Making matters worse, these out-of-spec practices and associated data issues are usually difficult to monitor, detect, and correct.

With that said, the primary goal of a company should be to not only have the proper elements of a well-governed entity, but have them working effectively as well. This comes with maturity and a constant focus on process improvement. Simply improving the data entry process alone is not enough. It is necessary to improve the support process around it. Just about everything is constantly changing: business needs, business landscape, technology, people, and so on. The only hope is to have an efficiently adaptive model that in spite of all these changes can continue to deliver results quickly. The topic of continuous improvement is addressed in Chapter 12.

Figure 9.2 depicts a data quality improvement cycle, which conveys the steps to identify, correct, and control data-quality issues. Notice that the steps intersect both IT and business to emphasize the strong collaboration needed between them. Each of the steps has multiple data-quality activities embedded. Let's describe the steps and their accompanying activities next.

Initiate

A data-quality effort can be initiated in many ways. Drivers are essentially the initiators of data-quality activities and the means by which they bring data-quality issues to the attention of the proper people. A company with a mature data-quality practice should be able to

support a multitude of drivers. Not only that, it should demand that everyone across the company participate in improving the overall quality of the data. After all, data quality is everyone's responsibility.

Essentially, data-quality initiatives fall into two categories: (1) reactive and (2) proactive. In general, proactive initiatives are measures established to avoid problems from happening or worsening, while reactive initiatives are measures adopted after the problem has already occurred and must be corrected. Drivers throughout the company, acting on their particular roles and driven by specific business needs, will either be reacting to data-quality problems or proactively preventing new problems from happening or existing problems from getting worse.

Users following a particular business process for data entry, for example, may detect irregularities with the data due to a system defect, a bad practice, or weak enforcement of business rules. They will not necessarily know the root cause of the problem or the best way to resolve it, and that is to be expected. But they need a mechanism for presenting the problem and requesting a correction.

Most companies will implement a type of trouble-ticket system that will allow users to communicate any problems they see. These trouble tickets are then categorized and routed to a suitable team for proper action. In this scenario, the problem entered by the user on the ticket becomes the requirement or problem statement.

The trouble ticket is just one mechanism by which a company should support requests for data-quality improvements. Special projects and certain activities commonly performed are very likely to have data management effects and should be supported with proper engagement of the data-quality team according to preestablished policies and SLAs. Here are some examples of activities that will require close data-quality participation:

- Migrating data from one system into another due to mergers and acquisitions, or simply to consolidate multiple systems and eliminate redundancy.
- Changes in system functionality, such as adding a new tax calculation engine, may require a more complete, consistent, and accurate postal code representation than before.
- Regulatory compliance, such as new financial reporting rules, the Sarbanes-Oxley Act (SOX), the U.S. PATRIOT Act, or Basel II.
- Security compliance, such as government data requiring particular access control rules.
- Bad or inaccurate data values that cause additional business costs and customer dissatisfaction, such as bad mailing addresses, inaccurate product descriptions, out of date product IDs, and more.

The drivers behind these activities will vary depending on the organizational structure. Even within a company, the same category of change could come from different organizations. For example, an IT-initiated system consolidation task may be the driver of a data migration activity, while a merger or acquisition is a business-initiated activity that could also lead to

data migration. In another example, a regulatory compliance requirement can come either from a financial organization or from an enterprisewide data governance initiative.

Most important to remember in such cases is that the data-quality process supports all data-driven requests, no matter what the driver is. Remember that the goal is to create a culture of focus on data quality. If a company limits who is allowed to report data-quality issues, it will create skepticism regarding its true objectives, which could ultimately lead to a companywide DQM failure.

In the context of multi-domain MDM, the initiation of certain data-quality activities is immediate. As stated in the beginning of this chapter, one of the goals of MDM is to tackle data-quality problems. Therefore, MDM itself is a major driver for data-quality improvements.

Assess and Define

The actions of assessing and defining are what make DQM different from other solution design activities, such as application development. With typical application development, business units have a very clear understanding of the functionality required for them to perform their jobs. New requirements, although vulnerable to misinterpretation, are usually very specific and can be stated without much need for research. Also, software defects sometimes might be difficult to replicate, diagnose, and fix, but again, the functionality to be delivered is mostly clear.

With data, the landscape is a bit different. One of the mantras for data quality is fitness for purpose. That means data must have the quality required to fulfill a business need. The business is clearly the owner of the data, and data's fitness for purpose is the ultimate goal. However, there are some issues. First, the business does not know what it does not know. A lot of the time, company personnel have no idea of the extent of the problem they face when it comes to data quality or lack thereof. Second, fitness for purpose can be difficult to quantify. There are certain situations where you just want to get your data as good as possible to minimize any associated costs, but it is not clear how much they can be improved. Third, the business might not fully understand the ripple effects of a data change.

In a multi-domain MDM implementation, there are certain predefined data-quality expectations, such as the ability to link inconsistent data from disparate sources, survive the best information possible, and provide a single source of truth. Still, the current quality state is mostly unknown, which tremendously increases the challenge of scoping the work and determining the results. The business might have a goal, but achieving it can be rather difficult.

The only way to really understand what can be accomplished is to perform an extensive data-quality assessment. It cannot be stated enough how data profiling is important. To efficiently and effectively correct data problems, it is a must to clearly understand what the issue is, its extent, and its impact. To get there, data must be analyzed fully and methodically. For an extensive discussion of data profiling, consult Chapter 8.

A recommended approach to organize data-quality problems is to categorize them into dimensions. Data-quality dimensions allow complex areas of data quality to be subdivided into groups, each with its own particular way of being measured. Data-quality dimensions are distinguishable characteristics of a data element, and the actual distinction should be business-driven. Since it is possible to characterize a particular data element in many ways, there is no single definitive list of data-quality dimensions. They can be defined in many ways with slight variations or even overlapping context. The following list represents the type of dimensions and definitions generally used:

- *Completeness:* Level of data missing or unusable.
- *Conformity:* Degree of data stored in a nonstandard format.
- *Consistency:* Level of conflicting information.
- *Accuracy:* Degree of agreement with an identified source of correct information.
- *Uniqueness:* Level of nonduplicates.
- *Integrity:* Degree of data corruption.
- *Validity:* Level of data matching a reference.
- *Timeliness:* Degree to which data is current and available for use in the expected time frame.

Data-quality dimensions are also very useful when reporting data-quality metrics as part of the Monitoring and Control phase, as will be discussed later in this section.

At the end of this step, it is possible to have a clearly stated set of requirements on what needs to be accomplished for each specific data-quality activity. Remember that most of the time when the process is initiated, there is not yet enough information to produce a complete set of requirements. This step complements the initiation by adding much-needed analysis.

Design and Validate

Once requirements and data analysis are complete, the next step is to design the solution. Notice there are many types of data-quality initiatives. The most common is a reactive data-quality anomaly that must be corrected. How companies address this common issue is a strong indication of their DQM maturity. The more mature companies will take this opportunity to do the following evaluation by answering these questions:

- How is business affected by this issue?
- What is the quantity and frequency of the problem?
- Is this an application bug or a process problem?
- Can this issue be prevented in the future?
- Does this issue require proper training to be avoided in the future?
- Are there any legal implications if data is corrected or left intact?
- Have any wrong decisions been made due to this issue, and who, if anyone, should be informed about the situation?

- Is data governance aware of the issue?
- What policies and procedures affect or are affected by the problem?
- Is it necessary to monitor this problem moving forward?
- What is the cost to the company if this issue is not corrected?

The reason for most of these questions is obvious. But notice the sixth item related to legal implications. Often, there are legal aspects associated with changing the data even if the data is wrong. For example, it might be illegal to modify a certain piece of information provided by a customer without their legal consent. Similarly, the ninth item evokes the need to contemplate any policies and procedures related to the data or to changing them. With multi-domain MDM, there could be many organizations affected by the change—make sure to consider all of them. A data governance body can facilitate the process of obtaining the proper approval for a change.

Let's cover the many data-quality activities (depicted in Figure 9.3) that should be considered when designing and eventually implementing a data-quality improvement effort.

Error Prevention and Data Validation

The best way to avoid data-quality issues and their costly effects is to prevent them from occurring. Of course, this is easier said than done. Technological limitations, time constraints, business complexities, and ever-changing business rules are some of the factors preventing real-time validation at the data-entry point. In addition, there are many interconnected systems, internal and external, that make it very difficult to constantly ensure the high quality of all data flowing in and out.

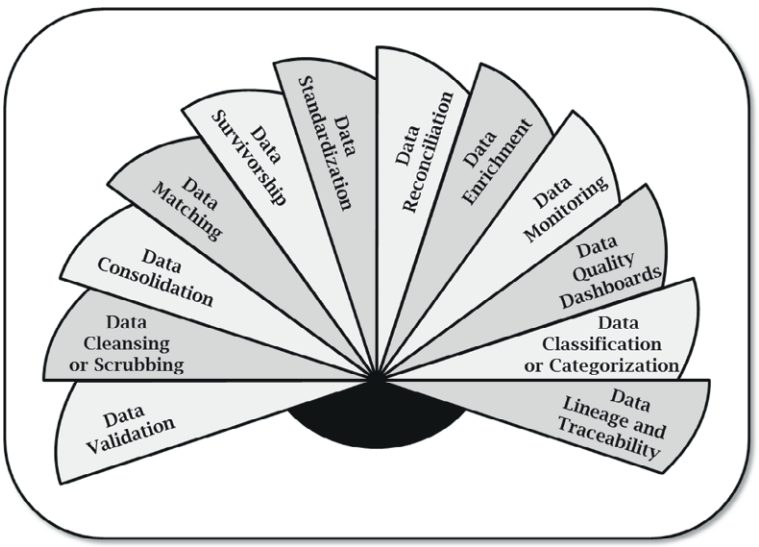

Figure 9.3: Data-quality activities

The technology behind multi-domain MDM varies from vendor to vendor, but it will typically encompass software applications, middleware, batch jobs, and databases. Therefore, data validation can be accomplished by a combination of client- and server-side techniques. Both client- and server-side validations have advantages and disadvantages.

Thin or fat clients will dictate the type of tools available for client-side validation, which will include varying levels of built-in code, along with graphical user interface (GUI) elements such as drop-down lists, check boxes, list boxes, radio buttons, and lookup tables.

The advantages of client-side validation include the following:

- Some GUI elements prevent users from entering incorrect values because they have to select from a predefined list.
- Performance is improved since there is no need to process a request/response message exchange with a remote server for further validation.
- When presented with a predefined list of options for a given attribute, users have the opportunity to choose the best one.

The disadvantages of client-side validation include the following:

- Any operation not done using the front-end interface will miss the client-side validation logic, potentially corrupting the data.
- Most GUI elements used for client-side validation are impractical when dealing with large amounts of data, adversely affecting user experience and productivity.
- Client-side validation can be maliciously bypassed and is subject to security concerns.
- Validation done via application code on the client side can be difficult to update and distribute, depending on the architecture in use.

Server-side validation can also be accomplished in multiple ways, with the most common being database referential integrity rules and software code embedded in the middleware layer or the database layer via stored procedures and batch scripts. Some products will offer database-independent solutions, with most of the server-side validation accomplished in a middleware layer, while others will be heavily database-driven.

The advantages of server-side validation include the following:

- Technology on the server side typically allows higher sophistication of validation rules.
- Validation rules developed directly into the database via referential integrity rules or triggers and stored procedures cannot be bypassed, providing the highest level of corruption prevention.
- Changes to existing rules are typically easier to make because they are centralized.

The disadvantages of server-side validation include the following:

- User response time is affected because of the additional request/response message exchange with a remote server.

- Valid options are not immediately obvious to the user.
- Software validation code embedded in the middleware or database batch scripts can still be bypassed with direct access to the database, risking data corruption.

The best approach is to combine client- and server-side validations to achieve optimal results. It is also important to minimize validation code duplication to prevent discrepancies or redundant work when rules change.

Companies also try to prevent data corruption through business process rules. However, this approach can be very ineffective depending on the size of the company, the quantity and skills of users, the level of training, and the emphasis on quality. For data-quality error prevention to be effective using business process rules, it is necessary to have constant communication and ongoing training and monitoring, coupled with incentives and rewards for high-quality data entry and maintenance.

Business rules, policies, and procedures can be dynamic. As they change, software upgrades may be required to incorporate additional data entry or maintenance validations. Those software changes can take a long time to occur due to IT resources and constraints. The business will typically opt to amend business processes to convey the new rules until the actual software change is complete. In this case, it is also critical to plan for a comprehensive data cleanup of previous data that no longer comply with the new rules. In summary, two additional activities may be required to address a business rule change:

- Making any necessary software changes to support the new rules and associated validations
- Cleaning up data that is no longer compliant or data that is temporarily entered or changed incorrectly until the change goes into effect

These two projects will need to be planned, prioritized, and executed independently according to business needs, also taking into consideration their mutual dependencies.

Unfortunately, it is difficult to totally prevent bad data from entering the system. Many data elements can be very difficult to validate in real-time, either because reference data does not exist at the level needed or because it can take a long time to search when you consider that a customer could be on the phone while data is being entered, and you do not want to make that person impatient. Mistakes happen when users are entering information into free-form fields due to incorrect interpretation of business processes, typos, or miscommunication with customers, partners, and vendors.

Fundamentally, it is important to anticipate data issues and include error prevention in the solution design. That is why a company that has instituted more proactive DQM measures will have fewer data issues affecting business and analytical operations. It requires data-quality participation from the beginning. A typical situation is the following: Data quality starts with sound data model design, making a data-quality specialist a must from the very

beginning. Data quality is not maintained just for the sake of data quality. Data have to have the required level of quality to fulfill a specific business need. However, if data models can be designed to prevent data anomalies to occur in the future, unquestionably it should be implemented right away, even if the business cannot foresee their future benefits.

Let's use marital status as an example. Assume that the business did not specify that it needed strong data validation. It simply stated that it had to ascertain marital status as part of a data entry process. Assume the following implementation scenarios:

Scenario A

Marital status is implemented as a free-form entry field. The team designing the data model does not have a data-quality specialist reviewing its work, and implements the plan without any data normalization.

Scenario B

Marital status has been normalized and implemented as a reference look-up field.

Both implementations fulfill the immediate business need. As data is captured as part of the newly data entry process, let's see what happens in both scenarios.

Scenario A

In this implementation, it is possible to enter many variations of the same marital status, invalid values, or no values at all. In this case, it is possible to violate the following dimensions of data quality: completeness, validity, conformance, and accuracy.

Scenario B

Only values from a predefined list are available for selection, avoiding the violation of completeness, validity, and conformance as occurred in scenario A. Accuracy might still be an issue in case data is provided incorrectly or users mistakenly select the wrong option. However, there is no question that this scenario leads to a much better outcome from a data-quality perspective.

Imagine now that a marketing campaign using demographics is initiated, and marital status is a key element to determine how products are consumed. Of course, the quality of marital status is now critical, and here is what happens in both scenarios.

Scenario A

Data analysis to assess quality issues and the ensuing cleansing effort must be executed in order to correct issues. As indicated previously, many data-quality dimensions will have to be addressed, making it potentially difficult to fix data so that they achieve the required fitness for purpose.

Scenario B

Existing data is likely as good as it gets, no matter what. Depending on data-entry practices, it is possible that accuracy issues are large. Still, there is no question that this scenario presents a much better situation even if some analysis and cleansing are still required. It is certain that the number of issues will be much less than in scenario A.

This example with these scenarios is very basic, but it clearly shows how a simple design approach can avoid a lot of aggravation in the future. It also shows that we shouldn't expect the business to spell out every single requirement, even if that would be helpful when trying to solve a business problem,. A proactive design should be the focus of a high-quality IT organization.

In the end, companies need to be creative with their data-quality efforts and reorganize themselves to make error prevention a top priority. Data-quality specialists need to be engaged in many data-related projects to ensure that good practices are followed at all times. MDM increases the opportunity for preventing data-quality issues for master data. Before MDM, inconsistent and duplicated information was entered all across the company. A sound MDM implementation will minimize those problems, and in addition, it will provide the opportunity to cleanse, standardize, and enrich information, as will be discussed next.

Data Cleansing or Data Scrubbing

Data cleansing (aka *data scrubbing*) refers to a process to correct or improve data. This implies that the data were bad already, which means that error prevention or validation wasn't done, and now any anomalies must be corrected.

But there are many things that could be wrong with the data. In the previous section, data-quality dimensions were covered. In essence, different data-cleansing activities need to be applied depending on the dimension associated with the failed quality rule. For example, if conformance is an issue, then data standardization must be applied to the data. If inconsistency is a problem, then data reconciliation is necessary, and so on.

It is at the core of multi-domain MDM to address many of these data-quality issues. It will typically attack the following topics:

- *Completeness:* Before MDM, data are fragmented across multiple sources. By matching data among many sources, it is possible to complement data from one system with another, hence increasing the level of completeness.
- *Conformity:* Data standardization is usually a prerequisite to data matching. It is very difficult to match data when they do not follow a particular format. Granted, there is fuzzy matching. Still, one form of standardization or another will be needed to support the matching of data across multiple sources as part of MDM data integration. This will improve conformity.

- *Consistency:* The end goal of MDM is to have a single source of truth. As such, inconsistencies are not tolerated and are, expectedly, addressed by MDM.
- *Accuracy:* MDM requires a particular set of data to be the surviving information for a given master record. Usually, a certain system is picked over another due to its status of the system-of-record (SOR) for that particular set of information. Assuming that the selection is right, this will be indeed the most accurate information related to a given entity inside the company. Furthermore, MDM advocates the use of reference data to ensure a more accurate multi-domain MDM hub.
- *Uniqueness:* MDM is about establishing and tracking the various sources of master data, and one major focus is to eliminate duplicates.
- *Integrity:* There are many contexts to integrity of data. For example, it could be integrity in the context of referential integrity in a database or the level of data corruption. A good multi-domain MDM hub should deliver solid referential integrity within and across multiple domains, preventing corruption of information.
- *Validity:* Whenever possible, the usage of reference data specific to each domain is encouraged to increase the legitimacy of master data.
- *Timeliness:* Without MDM, companies would struggle to deliver a complete and accurate list of master data records on a timely manner. With a multi-domain MDM hub, this is readily available to meet the ever-growing need for rapid delivery to gain an edge on competition.

Notice how MDM is intensive about improving the quality of master data. Companies need to leverage the practice of data cleansing throughout. Remember that every cleansing effort should include two additional efforts:

- Preventing this issue in the future: root cause analysis and mechanisms to avoid the problem altogether
- Monitoring the occurrence of this problem moving forward

Data Consolidation, Matching, and Survivorship—Entity Resolution

Three data-quality activities, consolidation, matching, and survivorship, are usually applied collectively. As a matter of fact, one definition of data consolidation implies the actual deduplication of information, which implies matching and survivorship. One other definition of data consolidation, however, is simply grouping data together without any particular intelligence behind it. This latter definition is more of a system consolidation than a data consolidation. Nonetheless, it is important to clarify to what extent data are truly consolidated.

Multi-domain MDM is all about fully consolidating data by matching and deduplicating data from multiple sources. This is at the core of entity resolution, which was covered to a great extent in Chapter 8. Nevertheless, as with many other data-quality activities, companies

should be looking to extend these data-quality practices outside of MDM, or even prior to starting MDM. Single systems can have lots of duplicates already. If data consolidation is started on an individual system basis even before MDM is started for that particular domain, it will tremendously help any later efforts to consolidate multiple systems.

Data Standardization

With multi-domain MDM comes a stronger need for enterprise-level data standardization. Integration of data is intrinsic to MDM, and for it to be effective, it requires the alignment of data definitions, representation, and structure. Therefore, data standardization becomes a core data-quality technique to facilitate the connectivity, consistency, and synchronization of information.

The ideal scenario would be for all matching information to conform to a unique standard in all systems across the enterprise, but that is not always feasible. Companies today have many commercial off-the-shelf (COTS) and in-house-developed applications spread across the many different organizations with different structures and models. In addition, there are many external data items from vendors and partners that reach internal systems through many interfaces. Those external systems are also structured differently with their own models. Here are some examples of how similar information can be represented very differently across systems:

- A person's name may be captured as separate fields for First, Middle, and Last names in one system, but as a single field for Full Name in another.
- An address in one system may have one single address line for street number and name, but two or more in another.
- One system might have a single field for the entire US ZIP+4 information, while another has two separate fields; and dashes may or may not be required.
- One system might store a social security number as a string, while another stores it as a number. Again, there could be multiple fields or not, and dashes or not.
- One system might capture gender as M and F, while another captures it as 0 and 1.
- One system might use a two-digit state code, while another spells out state names.

These are only a few of the many differences that can occur across multiple systems for similar data elements that are semantically the same. It is easy to see how unlikely it can be to modify all systems to conform to a single standard. From data models to data-view layers, IT systems can be very specific and expensive to change just to comply with a particular standard. Does that mean that you should give up on data standardization? Of course, the answer is no. However, the scope of data standardization is wider than usually expected.

Data standardization is not only about creating enterprise rules on what the data should look like, but also about how they need to be transformed to fit specific systems that cannot follow

an enterprise-level definition. A company should strive to have a single standard, but it must be flexible enough to adapt and transform data as needed when moving information across systems with noncompliant constraints.

To understand this concept, imagine a set of enterprise standard data definitions. Not all IT systems will follow all those standards, for the reasons explained previously. However, the intent is for a multi-domain MDM hub to follow those standards because one of the reasons to implement MDM is to improve the overall quality of information. As data flows from the hub to other systems, it can be transformed as needed if downstream systems cannot conform to predefined standards. Likewise, as data flow into the hub but are not already compliant, they need to be transformed to conform to predefined standards. Therefore, MDM implementation provides the perfect opportunity to comprehensively standardize master data across the enterprise.

Let's break the standardization process down into parts to explain it better. For this example, let's assume a hybrid-style multi-domain MDM, with a publish/subscribe method to synchronize data across all sources. Figure 9.4 depicts the high-level architecture.

In the hybrid MDM hub style, master data flow into the hub from multiple sources. As they reach the hub, they typically go through a series of data-quality improvements to cleanse, standardize, consolidate, and enrich data. The focus of this section of the book is on data standardization, which is shown in Figure 9.5.

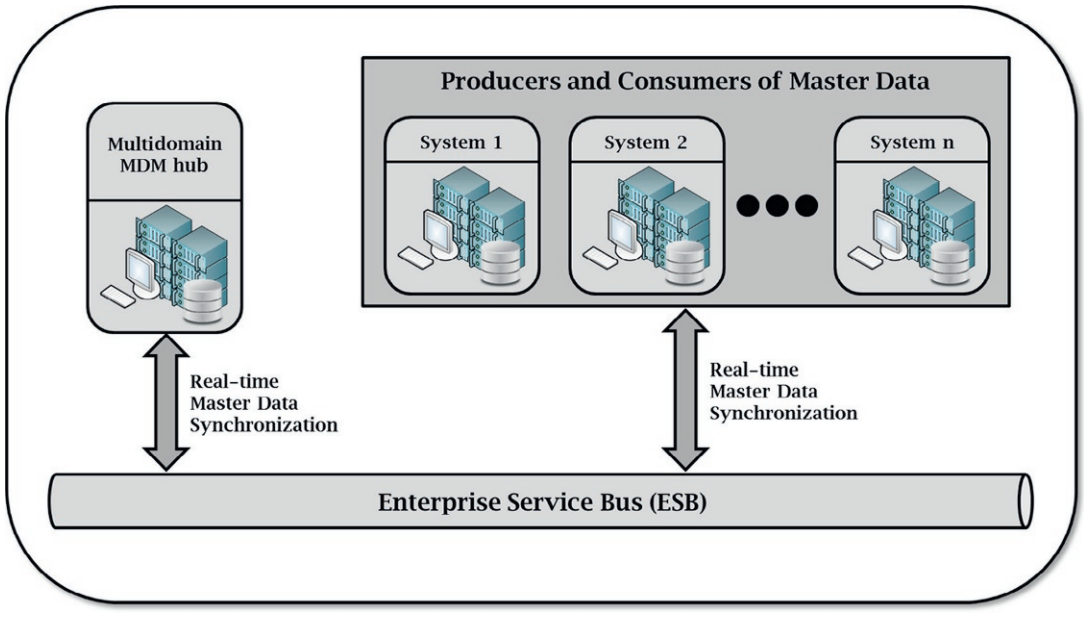

Figure 9.4: Hybrid-style multi-domain MDM with publish/subscribe synchronization

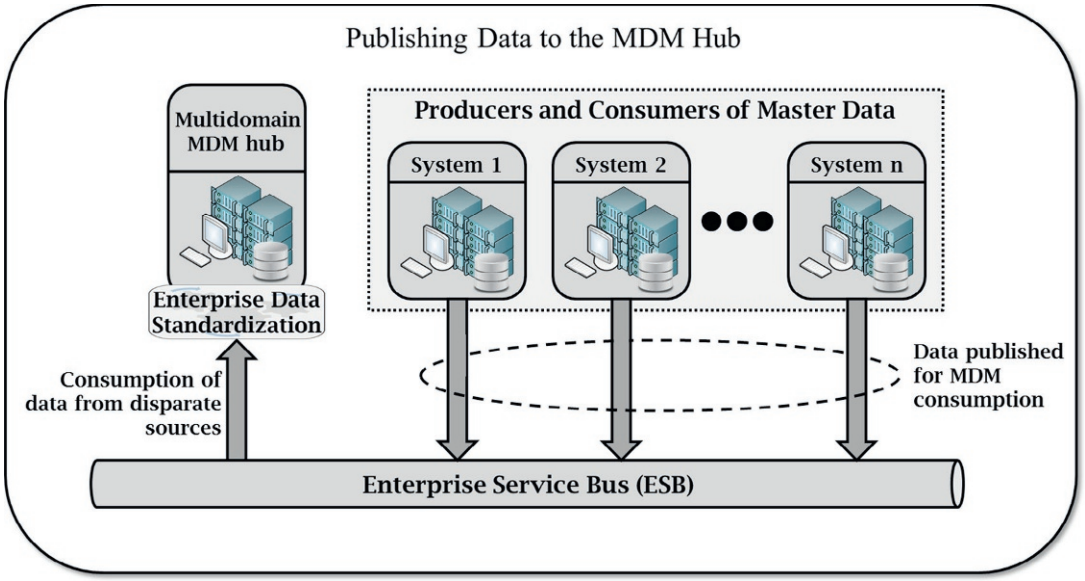

Figure 9.5: MDM consumes and standardizes data from multiple sources

Now that the data are standardized in the MDM hub along with multiple other data-quality improvements, it is obviously beneficial to propagate good data back to their original sources. However, each system will have its own idiosyncrasies, making it practically impossible for all of them to follow the same standards. As such, a series of data transformations must occur to convert the enterprise standard to a local standard when necessary. Needless to say, a strong effort should be made to have all systems comply with a singular standard. Noncompliance should occur only when the cost to implement doesn't justify a tangible business benefit. Figure 9.6 depicts an approach to customize standards when absolutely necessary.

Let's illustrate this concept with a data example. Many systems in a company store addresses. Let's assume three existing applications, as follows:

- System 1 offers three address lines: one for street number and name, a second for apartment/suite number, and a third for other information (building number, floor number, etc.).
- System 2 offers two address lines: one for street number and name and another for any other complements.
- System 3 offers separate fields for street number, street name, street type (St, Rd, Ave, Blvd, etc.), pre-direction (N, S, E, W), and post-direction (N, S, E, W).
- To simplify, let's assume only U.S. addresses and all three systems have the same number of fields for city, state, county, and ZIP.

When implementing a MDM hub, it is important to decide what standard to adopt. Data governance can drive the decision by working with the solution architects and the proper

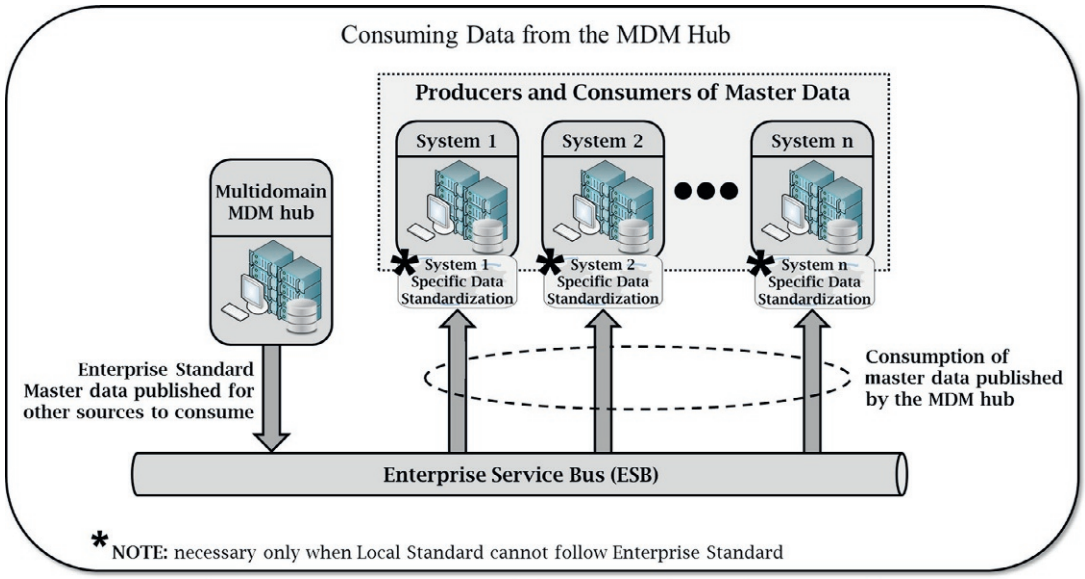

Figure 9.6: When enterprise standards are not supported by specific systems

business units. Let's assume that the enterprise standard adopted matches system 2. Therefore, when data from systems 1 and 3 are integrated into the MDM hub, they must be transformed to the enterprise standard. In addition, it is likely that all addresses coming into the hub will be cleansed, validated, and enriched by a reference source. In the end, there is a clean, valid, and standardized address in the hub. As this address is synchronized now with the original sources, there is a challenge. System 2 can simply pick up any changes and overwrite its contents since the structure is the same. However, systems 1 and 3 cannot simply take the data as they are. Assuming that system 1 has no business function relying on the third address line, it can simply start using only two lines and ignore the third one. Still, some minor transformation needs to happen to blank out address line 3 when receiving data from the hub. On the other hand, system 3 has a lot of work to do. Fitting two address lines into distinct elements for street number, name, type, pre-direction, and post-direction requires some heavy transformation. In addition, modifying system 3 to using only two lines can be very expensive since it may affect databases, user interfaces, business processes, and so on. In this example, systems 1 and 3 will require some type of customization to accommodate data consumed from the MDM hub.

The question becomes: Where should those necessary transformations occur? Enterprise standardization for data coming into the hub is typically part of a series of data-quality transformations already provisioned for as part of the multi-domain MDM hub architecture and design. The other necessary transformations for data flowing from the MDM hub to other sources can be implemented in different ways. Some companies opt to increase the role of

the data integration layer, such as the enterprise service bus (ESB), to transform the data to specific system needs. Other companies opt to add adapters between the integration layer and each system to perform the necessary conversion. Finally, the change can occur directly at the interface layer in each source system itself, which is modified to accept an enterprise standard, but convert internally to accepted format. There is no one-size-fits-all solution. Due diligence and proper evaluation should be performed to assess what is best for your company.

This overall data standardization approach should be extended to external systems as well. External sources will have their own models and definitions, which are completely out of control from an internal data-governing body. Therefore, data might need to be transformed as it comes into the company to conform to internal standards. The same issue exists with outbound data, with vendors and partners requiring specific formats to data that they receive. Adapters or other layers of data integration will need to perform proper transformations, especially because external systems are generally very difficult, if not impossible, to change.

Data Reconciliation

In general, data reconciliation is about measuring and correcting inconsistencies among data sets. Multi-domain MDM as a whole can be seen as a system to reconcile master data from multiple sources. Data reconciliation is very common as an activity to ensure that the process to transform and move data went as expected. For example, data reconciliation is done as part of a data-migration effort to make sure that the data at the target system contain all the data from the source, and that all differences have been identified and justified. An ordinary technique for reconciliation is to compare results of the same metric from two different systems, such as the classic example of balancing a checkbook.

In multi-domain MDM, there can be many situations where data reconciliation is necessary. For one, if data migration is needed as part of an initial population of the MDM hub, data will need to be reconciled to ensure that they were converted properly. As new sources of data are added, regular reconciliations are needed. Ultimately, MDM is constantly reconciling master data by integrating, matching, surviving, and synchronizing data across the enterprise to maintain master data consistent. A complete MDM solution should allow for inconsistent data to be reported and a mechanism to properly address any discrepancies.

Data Enrichment

Data enrichment or augmentation is the process of enhancing existing information by supplementing missing or incomplete data. Typically, data enrichment is achieved by using external data sources, but that is not always the case.

In large companies with multiple disparate systems and fragmented information, it is not unusual to enrich the information provided by one source with data from another. This is

particularly common during data migration, where customer information is fragmented among multiple systems and the data from one system are used to complement data from the other and form a more complete data record in the MDM repository.

As with any other data-quality effort, data enrichment must serve a business purpose. New requirements come along that may require data to be augmented. Here are some examples:

- A new marketing campaign requires nonexisting detail information about a set of customers, such as Standard Industry Code (SIC), annual sales information, company family information, etc.
- A new tax calculation process requires county information for all U.S. address records, or an extended format for U.S. postal code, which includes ZIP+4.
- A new legal requirement requires province information to be populated for Italian addresses.

Much of this additional information needs to come from an external reference source, such as Dun & Bradstreet or OneSource for customer data enrichment, postal code references for address augmentation, and so on.

It can be quite a challenge to enrich data. This process all starts with the quality of the existing data. If the existing information is incorrect or too incomplete, it may be impossible to match to a reference source to supplement what is missing. It can be quite expensive as well, since the majority of the reference sources will either require a subscription fee or charge by volume or specific regional data sets.

When matching data to another source, there is always the risk that the match will not be accurate. Most companies providing customer matching services with their sources will include an automated score representing their confidence level with the match. For example, a score of 90 means a confidence level of 90 percent that the match is good. Companies will need to work with their data vendors to determine what is acceptable for their business. Typically, there are three ranges:

- *Higher range:* For example, 80 percent and above, where matches are automatically accepted
- *Middle range:* For example, between 60 and 80 percent, where matches have to be manually analyzed to determine if they are good or not
- *Lower range:* For example, 60 percent and below, where matches are automatically refused

Once a match is deemed correct, the additional information provided by the reference source can be used to enrich the existing data. Address enrichment is very common, where the combination of some address elements is used to find what is missing. Examples include using postal code to figure out city and state, or using address line, city, and state to determine postal code.

The challenge comes when there is conflicting information. For example, let's say that city, state, and postal code are all populated. However, when trying to enrich county information, the postal code suggests one county, while the city and state suggest another. The final choice comes down to the confidence level of the original information. If the intent is to automate the matching process, it may be necessary to evaluate what information is usually populated more accurately according to that given system and associated business practice. If it is not possible to make that determination, a manual inspection is likely to be required for conflicting situations.

Data Classification or Categorization

Data classification is about categorizing and organizing data for better analysis and decision making. There are many ways to classify data. For example, categorizing data based on criticality to the business and frequency of use can be important to business process definitions. Classifying data based on compliance and regulations can be part of a risk management program. Data profiling is highly driven by data types and collections of data with similar content.

A particular type of data classification highly sought in a Customer MDM program is customer hierarchy. Customer hierarchy management entails managing customer data relationships to represent company organizational structures, for example. Another classification (in Product MDM, for instance), is product taxonomy, which is needed to organize products for a variety of purposes.

From a Business Intelligence (BI) perspective, hierarchical organization of data is essential. It allows a vastly superior understanding of market and industry segmentation. The volume of master data can be quite overwhelming. Classifying this data hierarchically is a critical first step to make the most sense of the information. The results can be applied to market campaigns, cross-selling, and up-selling.

From an operational perspective, hierarchy management is also critical in improving efforts to maintain master data. Some LOBs, such as sales, may have a vested interest in hierarchical organization for the purpose of territory segmentation and earning sales commissions.

It is doubtful that a single hierarchical representation for each data entity will meet the needs of all LOBs within any large company. Multiple representations are not uncommon but add to the project's cost. Maintaining a single hierarchy can be very challenging as it is. As the different perspectives grow, companies risk compromising the main reason that they engaged in an MDM project in the first place: an agreed-upon view of master data across the entire company. Finding the right balance is key.

Most likely, a multi-domain MDM repository will have relational relationships (e.g., a customer has multiple addresses or accounts, an asset is associated with a particular account,

and so on). These types of relationships are inherent to the structure of the repository and are conceptually different from hierarchy management. Furthermore, your MDM repository may not support hierarchy management. That might not be an issue to your implementation, as certain companies opt to do hierarchy management outside an operational multi-domain MDM hub, and inside an analytical environment. Expert guidance is always helpful to advice on the best approach.

Data Lineage and Traceability

Data lineage states where data is coming from, where it is going, and what transformations are applied to it as it flows through multiple processes. It helps understand the data life cycle. It is one of the most critical pieces of information from a metadata management point of view, as will be described in Chapter 10.

From data-quality and data-governance perspectives, it is important to understand data lineage to ensure that existing business rules exist where expected, calculation rules and other transformations are correct, and system inputs and outputs are compatible. Data traceability is the actual exercise to track access, values, and changes to the data as they flow through their lineage. Data traceability can be used for data validation and verification as well as data auditing. In summary, data lineage is the documentation of the data life cycle, while data traceability is the process of evaluating that the data is following its life cycle as expected.

Many data-quality projects will require data traceability to track information and ensure that its usage is proper. Newly deployed or replaced interfaces might benefit from a data traceability effort to verify that their role within the life cycle is seamless or evaluate whether it affects other intermediate components. Data traceability might also be required in an auditing project to demonstrate transparency, compliance, and adherence to regulations.

Implement and Validate

Much about data-quality activities was covered in the previous section. This step is about the realization of what was previously designed. The diversity of data-quality projects should be obvious by now. Data-quality efforts could be stand-alone, such as cleansing a specific data attribute, or they could be part of some bigger activity, such as a data migration project that encompasses improving the converted data. Obviously, multi-domain MDM requires many data-quality initiatives. Depending on the MDM approach, many of the following data-quality tasks are needed:

- Data integrity constraints as part of a data model implementation
- Data validation and prevention as part of a user interface customization
- Data cleansing, standardization, consolidation, enrichment, and reconciliation as part of a data migration implementation

- Implementation of business and data-quality rules to integrate, match, and survive a golden record from master data at multiple sources
- Realization of enterprise standardization for master data
- Addition of data validation to interfaces during data synchronization
- Integration with reference data for data enrichment

Data are everywhere, and so are opportunities for improving them. But besides implementing solutions, it is also important to test them meticulously. Companies are used to functionality tests, which do include different data scenarios, but they usually underestimate what it takes to truly test data quality.

Recall from Chapter 3 that the business has a shallow and narrow view of data because operational business processes only use a subset of all entity attributes at a time (narrow view of the data), and operational business processes use only a small number of rows of information at a time (shallow view of the data). To truly test data, it is important to widen and deepen that view. Therefore, when implementing data-quality improvements, understand how data will be changing in the future and how those changes will affect data integrity. Create data scenarios simulating those situations. Do not limit yourself to functional testing only.

Monitor and Control

Data-quality metrics falls into two main categories: (1) monitoring and (2) scorecards or dashboards. Monitors are used to detect violations that usually require immediate corrective action. Scorecards or dashboards allow numbers to be associated with the quality of the data and are more snapshot-in-time reports, as opposed to real-time triggers. Notice that results of monitor reports can be included in the overall calculation of scorecards and dashboards as well.

Data-quality metrics need to be aligned with business KPIs throughout the company. Each LOB will have a list of KPIs for its particular needs, which must be collected by the data-quality forum and properly implemented into a set of monitors, scorecards, or both.

Associating KPIs to metrics is critical for two reasons:

- As discussed previously, all data-quality activities need to serve a business purpose, and data-quality metrics are no different.
- KPIs are directly related to ROI. Metrics provide the underlying mechanism for associating numbers to KPIs, and consequently ROI. They become a powerful instrument for assessing the improvement achieved through a comprehensive data-quality ongoing effort, which is key to the overall success of an MDM program.

The actual techniques for measuring the quality of the data for both monitors and scorecards are virtually the same. The differences between them are primarily related to the time

necessary for the business to react. If a critical KPI is associated with a given metric, a monitor should be in place to quickly alert the business about any out-of-spec measurements.

Data-quality-level agreements (DQLAs) are an effective method to capture business requirements and establish proper expectations related to needed metrics. Well-documented requirements and well-communicated expectations can avoid undesirable situations and a stressed relationship between the data-quality team and the business and/or IT, which can be devastating to an overall companywide data-quality program.

The next two sections describe typical DQLA and report components for monitors and scorecards.

Monitors

Bad data exist in the system and are constantly being introduced by apparently inoffensive business operations that theoretically follow proper processes. Furthermore, system bugs and limitations can contribute to data-quality degradation as well.

But not all data-quality issues are equal. Some affect the business more than others. Certain issues can have a very direct business implication and need to be avoided at all costs. Monitors should be established against these sensitive attributes to alert the business about their occurrence so that proper action can be taken.

A typical DQLA between the business and the data-quality team will include the following information regarding each monitor to be implemented:

- *ID:* Data-quality monitor identification.
- *Title:* A unique title for the monitor.
- *Description:* A detailed description of what needs to be measured.
- *KPI:* The KPI associated with what is measured.
- *Data-quality dimension:* Helps organize and qualify the report into dimensions, such as completeness, accuracy, consistency, uniqueness, validity, timeliness, and so on.
- *LOB(s) affected:* A list of business areas affected by violations being monitored.
- *Measurement unit:* Specifies the expected unit of measurement, such as number or percentage of occurrences.
- *Target value:* Quality level expected.
- *Threshold:* Specifications for the lowest quality acceptable, potentially separated into ranges such as acceptable (green), warning (yellow), or critical (red).
- *Measurement frequency:* How often the monitor runs (e.g., daily or weekly).
- *Point of contact:* The primary person or group responsible for receiving the monitor report and taking any appropriate actions based on the results.
- *Root cause of the problem.* Explanation about what is causing the incident to occur (if known).

- *Whether the root cause has been addressed:* Prevention is always the best solution for data-quality problems. If a data issue can be avoided at reasonable costs, it should be pursued.

Table 9.1 describes a potential scenario in which a monitor is applicable. Notice the explanation of the root cause of the problem and the measures that are being taken to minimize the issue. Sometimes it is possible to address the root cause of the problem, and over time, eliminate the need of a monitor altogether. In these cases, monitors should be retired when they are no longer needed.

The monitor report result is best if it is presented graphically. The graph type should be picked according to the metric measured, but it is almost always relevant to include a trend analysis report to signal if the violation is getting better or worse with time.

Scorecards

Scorecards are typically useful to measure the aggregate quality of a given data set and classify it in data-quality dimensions. Numbers for a scorecard can be obtained from regularly executed data assessments. The individual scores can be organized in whatever ways are needed by the business and presented in a dashboard format.

Table 9.1: Sample Monitor DQLA

Id	DQ001
Title	Number of duplicate accounts per customer.
Description	Business rules require a single account to exist for a given customer. When duplicate accounts exist, users receive an error when trying to create or update a service contract transaction associated with one of the duplicated accounts. The probability of users running into duplicate accounts is linearly proportional to the percentage of duplicates. A 1% increase in duplicates translates into a 1% increase in the probability of running into an account error. Each account error delays the completion of the transaction by 4 h, which increases the cost by 200% per transaction. Keeping the number of duplicates at 5% helps lower the overall cost by 2%.
KPI	Lowers the overall cost of completing service contract bookings by 5% this quarter.
Dimension	Uniqueness
Affected LOBs	Services
Unit of measure	Percentage of duplicates
Target value	5%
Threshold	<= 10% is Green, between 10% and 20% is Yellow, >20% is Red
Frequency	Weekly
Contact	services_alias@company.com
Root cause	Duplicate accounts are a result of incorrect business practices, which are being addressed through proper training, communication, and appropriate business process update.
Fix in progress?	___Yes ___No _x_Mitigation ___N/A

Table 9.2 shows a sample scorecard. The objective is to obtain a score for a particular combination of context, entities, attributes, and data-quality dimension. Once the score is available, the scorecard report or dashboard can be organized in many different ways, such as the following:

- The aggregate score for a given context and entity in a certain dimension, such as 74 percent accuracy for addresses in the United States.
- The aggregate score for a given entity in a certain dimension, such as 62 percent completeness for all customer attributes.
- An overall score for a given data-quality dimension, such as 64 percent consistency.
- An overall score for all data-quality dimensions, which represents the overall score of the entire data set being measured. This associates a single number with the quality of all data measured, which becomes a reliable thermometer reflecting the data-quality efforts within the company.

The threshold should be set according to business needs. Data-quality issues represented by scores in the Red or the Yellow category should be the targets of specific data-quality projects. Furthermore, the scorecard itself will become an indicator of the improvements achieved.

The scorecard becomes a powerful tool for the following reasons:

- It assigns a number to the quality of the data, which is critical to determining if the data are getting better or suffering degradation.
- It can be used to assess the effect that a newly migrated source would have on the overall quality of the existing data.
- It clearly identifies areas that need improvement.

Table 9.2: Foundation Scores for the Data Quality Scorecard

				Threshold		
Entities	**Attributes**	**DQ Dimension**	**Score**	**Red**	**Yellow**	**Green**
Customer	Name	Completeness	98	<=95	>95 and <98	>=98
Customer	Social Security Number or Tax Identification Number	Completeness	60	<=55	>55 and <70	>=70
Customer	Social Security Number or Tax Identification Number	Conformity	70	<=80	>80 and <95	>=95
Address (U.S.)	Postal Code	Conformity	68	<=75	>75 and <90	>=90
Customer	Name/Country	Uniqueness	80	<=70	>70 and <90	>=90
Address	Address lines 1–4:City County State Postal Code Country	Accuracy	75	<=70	>70 and <85	>=85
Account	Account Type	Uniqueness	90	<=85	>85 and <95	>=95
Account	Account Number	Integrity	85	<=85	>85 and <95	>=95
Customer	Customer Type					

Notice that the scorecard alone may not be sufficient to determine the root cause of the problem or to plan a data-quality project in detail. The scorecard will highlight the area that needs improvement, as well as measuring enhancement and deterioration, but it might still be necessary to profile the data and perform root-cause analysis to clearly state the best way to solve the problem.

The DQLA for scorecards between the business and the data-quality team can follow a format similar to Table 9.1.

Conclusion

Data-quality activities are central to a multi-domain MDM implementation. This chapter covered how MDM fundamentally tackles data-quality problems by combining fragmented and disparate data from multiple sources, and distributes a single version of the truth in a timely fashion to multiple organizations within the enterprise. In a multi-domain MDM environment, certain data-quality activities need to be tailored to specific domains, but a single controlling body should guide the overall practice of DQM for more efficient execution and coordination.

There are clear benefits to establishing a centralized data-quality organization to coordinate the overall data-quality activities in support of a multi-domain MDM implementation. As a matter of fact, a DQM office should be structured to support a variety of data-quality efforts throughout the organization, working in conjunction with data governance to properly serve the most pressing business issues.

DQM needs to be managed as an ongoing practice, bridging business and IT regarding data-quality issues. A data-quality improvement model is important to ensure that the highly technical nature of most data-quality activities are properly founded by business needs, support, and validation.

Finally, data-quality needs to be measurable. The practices of monitor and control are necessary to measure the existing level of quality, as well as keeping track of any future degradation. Mature companies need to push the envelope toward proactive DQM, as opposed to only reactive efforts.