# Improving Data Quality Through Data Modeling

William McKnight

McKNIGHT
Consulting Group

www.ERwin.com

ca technologies | ERwin

# Table of Contents

# Introduction

Before you leave for the airport for an international trip, don't forget your passport. As our enterprises begin the inevitable journey to real-time competitiveness, the passport is data quality. Mostly, organizations have treated data quality as something that is assumed to be there. Prevention of quality shortcomings has always been the risk that organizations thought they could get away without. It is the most unanticipated reason for system failure since organizations began building systems. To say the time is now for implementing a data quality program to prepare your business for the present and the future would not be an understatement.

# The State of Enterprise Data Quality

Enterprise data is largely falling short of a standard that makes it possible to utilize in how business will be conducted in the next decade. Most enterprise data is "adequate" for basic operational needs today. It has been brought to standard, through long, slow processes of selectively upgrading the data intake functions of the organization. However, in the decade to come, business will care much more for the data than just ensuring it passes basic data entry standards. It will care that the data has true data quality across the enterprise.
This is based in large part on three certainties:

1. Information volume is exploding[1]— This is evident only in our accumulation of data, but also in the information flow into the business, including from third-party sources.
2. It is a real-time based business world— Opportunities must be realized early, often and accurately
3. Information is a key business asset— No matter what business you're in, you're in the business of information; information management is a strong competitive differentiator today.

In the developing business model, the timeliness of decisions is more paramount than ever. Latencies involved in capturing data, analyzing it, making decisions and then taking the requisite action must be eliminated. The appropriate action must flow immediately from the presentation of data into the enterprise. This "new information" must then be weighed against the backdrop of information selected from all the enterprise data ever collected and the appropriate pre-defined action chosen. When that action involves automation, that will be triggered at this point.

Having the correct information – that is, with data quality – available at all times is essential to this form of business conduct in which there is no time for latencies or extensive analysis. Analysis will become less real-time and more abstracted and will go into the development of systems. The analysis is performed in the engineering of the systems and the data they use. Analytics need to be embedded in processes and to accomplish real-time analytics, and to support this we must have data quality.

This is an event-driven, process-oriented world and quality data will be essential to success. To achieve this data quality, a data model is crucial to understanding the structure and meaning of information. Any data modeler and her models must be prepared on several levels such as the fact that models need to get up and running quicker and they are much more dispersed throughout the organization, well beyond the major operational and analytical systems. Modeling has never been more important. Neither has the quality of data within modeling's structures.

# Data Quality Defined

Data quality is the absence of intolerable defects. It is not the absence of defects. Every enterprise will have those. It is the absence of those defects that see us falling short of a standard that would have real, measurable negative business impact. Those negative effects could see us mistreating customers, stocking shelves erroneously, creating foolish campaigns or missing chances for expansion. Proper data quality management is also a value proposition that will ultimately fall short of perfection, yet provide more value than it costs.

[1]According to a study from research firm IDC and storage vendor EMC, data requirements are growing at an annual rate of 60 percent. Today, that figure tops 45 gigabytes for every person, or 281 exabytes total (equivalent to 281 billion GB).

All data quality defects fall into a set of 11 broad buckets:

1. Lacking integrity of reference between data values across the model
2. Entities lack unique identification
3. The expectations of the quantity of linked relationships (cardinality) of the data not met
4. Expectations of occurrences of field values based on other values (subtype/supertype) not met
5. Unreasonable values
6. Attributes are used for multiple meanings
7. Inconsistent formatting
8. Incorrect data
9. Missing data
10. Miscalculations
11. Data that falls outside of its intended codification

By using these categories alongside the articulated business interests, the data quality of a system, or an entire enterprise, can be measured and actions accrued for improvement. However, no such effort should begin with only a data cleanup, which only patches the bad data before the next wave of poor data comes through. So where does data lacking quality come from?

## The Causes of Poor Data Quality

In our haste to build the operational system, many times details such as the quality of the data in use is an afterthought. The goal is throughput of transactions. Putting gates up on the data entry process is not even a consideration. Yet, from an enterprise perspective, that entered data is more valuable than it ever was.

Information, as a competitive battleground, now has dozens of times over more uses than a couple of decades ago, when business was all about the transactions. Now, that single piece of data entry can be used in dozens of applications, whether or not it is physically replicated or not. The value of data entry to the initial receiving system is about 10% of the value of that data to downstream systems and the enterprise. However, most current data quality solutions focus on data cleanup.

This requires cooperation across the enterprise to meet enterprise goals. Data entry, more important than ever, now needs to be incented for quality measures instead of just quantity measures. Often, it just takes some oversight or cooperation to achieve data quality right at the point of entry. It takes influence from the data modeling discipline into the data entry process.

To be successful in the next decade requires planning for the next decade. It means planning what the organization will look like (markets, size, geography, etc.) and developing the commensurate plans to get there. These plans are lost without a hyper focus on the information of the organization. Such plans must have a strong focus on data modeling, data quality, data architecture and data entry. Data modeling and quality are attended to by organizations with a vision of success in the next decade.

# Modeling Cures for Data Quality Assurance

## Modeling for Data Entry

Fixing data quality problems through improved modeling of the data entry systems is effective and often overlooked. Data enters environments in a variety of ways. Data that is entered, received in any way, or calculated, needs to be properly modeled if data is to realize its potential for the enterprise.

Suppose a data entry analyst was under-trained and in a rush and standing in front of a customer taking the usual information when the customer springs something new into the profile – she lives in an apartment. This analyst looks for about two seconds across his screen, sees nothing that says ‘apartment number’, but remembers he has seen a field that he used to use but, as time has gone on and he has not seen the benefit, he has chosen personally to quit using. That field is *birthdate*. So now the apartment number is in the *birthdate* field, in violation of rule 11 above, the analyst is on to the next customer and he has sent a worse headache to headquarters than the usual headache he sends when he doesn’t fill out *birthdate*. One day the organization will want to become a real-time, analytical competitor in order to keep up with

the competition. It will want to know the ages of its customers. However, it will find it is extremely hampered by the practices it has let into its data entry systems.

The developer of the data entry screen had intended that apartment number simply be appended to the address field, as in "123 Main Street, Apt. 55". But this detail was not in the training and it was not evident by looking at the screen in the way it will be used in the real world. Given the sometimes low tenure and commitment of the analyst, it may be necessary to add some constraints in the data model to protect us from ourselves.

It used to be that operational systems were hands-off to any interests beyond support of quick data entry. Well-intended speed-up measures like removing drop-down lists and referential integrity had the predictable knock-on effect of lowering overall data quality. When the organization's lifeblood was those transactions, that made sense. Now that the transactions have slowed down or otherwise shifted to the competitors and information is the battleground, that strategy doesn't make sense. To those who are skittish about touching operational systems and performance, try harder. The ghost was given up far too easily in the past.

The major data modeling constructs relevant to data entry data quality, which relate to the data quality defect categories are:

1. Uniqueness – the enforcement that a column will have unique values in it
2. Check – guarantees that a column's value will fall in a predefined range or list
3. Key – forcing the integrity of desired references across entities like the existence of the customer before he places an order
4. Mandatory – forces a true value to be entered into a column
5. Default – setting a value when none is entered
6. Null – allowing null (no value) to be used in a column instead of forcing a value

Defaults and null constraints are usually more problematic to data quality when used than when not used because they allow for an abstract value (or null, which is no value) to be used in place of a customized, relevant value.

These major data modeling constructs relevant to data quality are enforced in the organization in its data models. Data models are the most leveragable place in the entire architecture to enforce change. If you get the data model correct by following the above constructs, the data has a much higher chance of having quality. Get the data model wrong and you spawn innumerable downstream workarounds and organization gyrations which run cover for the model shortcomings.

# Accepting External Source Data

The external data marketplace has never been more robust. Organizations are routinely going outside their organizational walls for competitive data, feedback data, market data, customer data, prospect data and as a replacement for the questionable quality of internal data. One major function of external data is to extend the understanding of the customer base by "reverse-appending" the limited/poor information on the customer base with the syndicated marketplace, yielding a more robust customer profile. This external, syndicated data can be purchased data, but there is also the web, which is increasingly a part of the source environment.

In addition to extending the attribute set for existing customers and building prospect lists, external data and logic can be used to deduplicate, or match and consolidate internal lists. Syndicators come in a wide range of sizes. There are large syndicators with mass consumer and business data designed to serve the need for up-to-date, common attributes. There are also numerous syndicators within various industries.

Regardless of the size of the syndicator(s) used or the number of syndicators, when addressing data quality within an organization now and for the next several years, addressing this data source is going to be increasingly important as it becomes a greater percentage of the information comprising the information asset of the organization. Organizations can be lulled into a sense of security with syndicated data and neglect necessary modeling principles at peril.

Syndicated data is still data and requires no less scrutiny than non-syndicated data. In many ways, much more scrutiny should be applied since less control is established over the information. This scrutiny begins by assessing the nature of how the syndicator arrives at the data that is sold and understanding what the generalizations are that they used in arriving at the data. More than internal data, syndicated data is subject to end user misinterpretation, implementation and integration errors and issues that stem from overlap with internal data.

Models serve the interest of the applications using them and should be built for purpose. Modeling for syndicated data should be built for use and not built with a hyperfocus on the source data from the external data provider. This modeling should follow these practices:

- The model should have assigned a unique value to each customer or other entity being sourced
- Use internally generated unique reference (surrogate keys) for each key value
- Do custom geography and time dimensions, do not necessarily accept these from a syndicator
- Utilize "Slowly changing dimensions" ("type 2 or 3") – which are techniques to track changes to data over time - to track changes in the entities - don't just rip/replace each time the vendor sends a batch
- Most syndicated data is in alphanumeric data type - model numeric metrics (i.e., salary, net worth, birthyear) as numeric
- Communicate, either by naming standards or internally generated definitions known as metadata, the full meaning of a syndicated data item - for example Acxiom's (syndicator)"net worth" is really "Acxiom derived net worth"
- Think twice before using syndicator field names
- Use confidence columns (explained below)

Confidence columns are internally populated columns that are associated with other "real" columns and are populated with the "confidence" that the organization should have in the associated column. It is a number on a scale of 0 to 1, with 0 meaning no confidence to 1 meaning full confidence. As opposed to "cleaning" the data, the idea is to tag the data (potentially in addition to cleaning) upon initial entry into the organization.

Confidence can be part of the column's metadata if the confidence is associated at the column level. This is appropriate when a discovery of the syndicator's method is the overriding factor. For example, you may source a "cigar aficionado" column as part of the customer profile, but have serious reservations about the manner in which the syndicator came to set this flag since you learned they set the flag based on subscription to Cigar Aficionado as opposed to firsthand survey data, based on the metadata definitions in the data model.

Confidence can also be paired with every existence of the column (each row) if the confidence is assessed at the value level. For example, in your analysis of a syndicator's data, you may assess that their data is much better for people on the East Coast than the West Coast.

Until the syndicated marketplace becomes the go-to high-quality source for the data it provides, this will be necessary. Less than full-quality data is valuable, but it is more or less valuable depending on the use of the information. For example, if you had a limited campaign budget, you may choose to market only to those records where the confidence level in the address was > .8.

Although confidence is a concept that is applicable to internal data, it is most appropriate with external data and should be part of any data model receiving valuable, but potentially suspect, data. The bottom line is regardless of data source; good modeling principles need to be adhered to in order to deliver data quality to the organization.

# Modeling for Master Data Management and Data Integration

A growing source of important information in our enterprises is found in master data management (MDM) and other data integration systems. Organizations are increasingly turning to MDM systems to improve data origination processes by

utilizing MDM's workflow, data quality and business rule capabilities. MDM systems provide for the management of complex hierarchies in the data, providing access to those hierarchies at any point in historical time. Other systems and projects that integrate data face similar issues, such as consolidation from mergers and acquisitions, data warehouses to support business intelligence, etc.

If well done, MDM systems generate the single version of the truth for the data it masters before any other systems gain access to the data. Then, those environments have systems that are working with the corporately adjudicated master data with high data quality, as opposed to environments where each system is responsible for their own data.

This architecture can have a dramatic effect on enterprise data quality. However, the MDM environment must be modeled well in order to achieve the benefits. You could be moving all kinds of interesting data around the organization with MDM, but if it does not adhere to a high standard of quality, it can all be for naught. Actually, that would be an MDM implementation that would not be worth doing at all.

How many applications could successfully execute their function with dirty data stemming from poorly designed data models? Could cross-selling and up-selling be effective if the customers were not unique and complete or the products at different levels of granularity or with incorrect attribution? Could you do credit card fraud detection or churn management correctly if you did not have the customer's transaction pattern correct? Many of these applications have failed in the past because they were not supported with clean, consistent data and supporting definitions of the business entities such as common definitions of customer, prospect, product, location, part, etc.

Of course, again, the best place to ensure data quality is at the original point of entry. Remediation efforts after that point are more costly and less effective. Many data entry systems, even MDM, allow for free-form data entry, which is a real inhibitor to system success. MDM systems, as much as any system of entry, should fully adhere to the Modeling for Data Entry section above. The modeling may be more important in a MDM system than any other system in the enterprise.

If MDM is not used as a system of entry, but rather as a collection and distribution point for the data, when the data does flow to MDM from its point of entry, data quality can be applied at that time and optionally/preferably shared back to the originating system as well as all subscribing systems.

Finally, it's important to note that given the expansive responsibility of the MDM model, it will be important to model it for the entire enterprise, as opposed to a point need(s). Modeling for the enterprise requires:

1. Modeling the data at the lowest grain possible – e.g. for every item purchased, not just the overall transaction
2. Modeling a superset of enterprise attributes – e.g. those interesting to Purchasing, Operations, Marketing, Sales, Service, Product Management, etc
3. Modeling business rules from across the enterprise

# Cleaning Up to Get Started

I have talked about the three major sources of modeling opportunities to improve enterprise data quality – data entry systems, syndicated data and master data management systems. I have also talked about the importance of implementing data quality early in the data lifecycle. It's no coincidence that the systems I focused on as having the major enterprise opportunities are those systems that meet data at its points of origination.

However, you may still have terabytes of information sitting in the environment, causing underperformance of systems across the enterprise. To enforce the "new" business rules consistently across all the data will require a cleanup effort. This brings us to the starting gate for physical changes to the data improve enterprise data quality.

Existing structures and their data must be analyzed for fit in the environment after the modeling cures are applied. It is highly desirous that the data quality be consistent across all history and into the future. This means remedying history data (defined as pre-cleansing modeling) to the go-forward standard. This also can mean a set of challenges.

If, for example, you want to constrain future data entry for last names to be at least 2 characters, but you have several that are only 1 character today, something must be done with that data to bring it to standard. If, for example, you wanted to constrain future birthdates to be less than the system date, but you have several that indicate a birth year of 2099, what are you going to do with that data?

Your choices are:

1. Clean the data
    a. Automatically – change the data based on rules (if x, then y)
    b. Manually – send the data to a human console to be studied for what to do
2. Nullify the data
3. Delete the record

Fields like name in a customer record that, if incomprehensible, would totally invalidate the rest of the record, you cannot just nullify the data. There is a level of poor quality data, such as name of "xxx" that leaves you with nowhere to go with it by itself. From here, you can:

1. Give up and delete the record
2. Salvage the rest of the record (if it's of high quality) by making the name something like "Unknown"
3. Pass the entire record to a human, who can do research with the other fields to come up with the name

No one field's fate should necessarily be decided based solely on that field's value. The adjacent fields must be considered as well.

There are other fields that you would expend less energy into necessary getting correct. One key overriding principle in data modeling for data quality is that you will never get it 100% perfect. You must balance perfection against timelines. A field like *automobile make* may not only be only mildly interesting, it may also be impossible to ascertain should the value be "xxx". For those invalid fields, you would nullify them.

When developing the clean up strategy, it is worth reiterating that the primary source of the modeling decisions in support of data quality is the business interests. These are best represented directly by those in the business with daily, direct triage with the data in question.

Once the cleanup strategy is determined and implemented and the modeling throttles are placed on new incoming data, all based on business interests and forethought tradeoff between timelines and quality, the enterprise will have data quality beyond operations. Effective data modeling chokes off data quality issues before they become mounting challenges to organization goals. The enterprise is ready for the business of the next decade – the business of information – and can channel the data into the timely analytical decision-making required.

*William McKnight, President of McKnight Consulting Group and very experienced at platform selection, has written more than 150 articles and white papers and given over 150 international keynotes and public seminars. His team's implementations from both IT and consultant positions have won Best Practices awards.*

To find out more about CA ERwin® visit www.erwin.com