White Paper

Bloor

Data Mastering

An 80/20 solution for MDM

A White Paper by Bloor Research
Author : Philip Howard
Publish date : February 2009

Data mastering is the process of getting your master data 'right' and delivering what is referred to in the Wikipedia definition as "an agreed-upon view" of the business entities under consideration.

Philip Howard

Introduction

It is easy to agree that getting your data 'right' is a necessary pre-requisite for Master Data Management (MDM), but:

1. few companies actually make realistic plans for that aspect of their MDM initiative and

2. even fewer stop to think about "if I already had data that was fit for purpose across my business, what would my remaining MDM requirements really be?"

This paper discusses the advantages of data mastering and, in particular, its relevance to MDM. Indeed, we suggest that it should be considered a pre-requisite and precursor to implementing MDM: in some cases even as a lower-budget substitute to a full MDM program, where much of the value can be delivered for a fraction of the cost (the 80/20 rule is a reasonable estimate). An additional factor is that traditional data quality products were designed to cater for name and address cleansing and matching and do not necessarily have architectures that lend themselves to cleansing product data, so if your MDM objective relates to product data (often referred to as PIM—Product Information Management), a different approach may be required, which we will explore later.

However, before going on to discuss these points we will first need to clarify what we mean by data mastering and explain how it differs from traditional approaches to MDM and data quality.

Data mastering defined

According to Wikipedia master data is: "that persistent, non-transactional data that defines a business entity for which there is, or should be, an agreed-upon view across the organisation." That is, the data that defines who a customer or supplier is (name and address, contact details and so forth), what a product is (code, description, weight, volume and so on), what contracts consist of, details of plant and machinery, company location information, and, generally, defining details of any business entity that is relevant to the company. The storage, processing and administration of this master data is typically (we will discuss this further in due course) the role of MDM or specific subsets of MDM such as customer data integration (CDI), product information management (PIM) and global supplier management (GSM). However, that leaves the question of what we mean by "data mastering".

In essence, data mastering is the process of getting your master data 'right' and delivering what is referred to in the Wikipedia definition as "an agreed-upon view" of the business entities under consideration. In other words, ensuring that there is a single, valid, as up-to-date as possible, set of data that defines each customer, product, contract, location or whatever. Note that this is a necessary precursor to any form of MDM: there's not much point in going to all the expense and upheaval of implementing MDM unless you have your data right in the first place.

To be specific, the core elements of a complete data mastering toolset that should precede and ultimately complement an MDM program are:

- Data discovery: you need to be able to identify the potential sources of data, analyse consistency, completeness, validity, overlaps and duplication, then select 'most trusted' sources. Most commonly, but not necessarily, this will be accomplished by some form of data profiling.

- Data quality: the ability to cleanse and match incoming data, identify errors, omissions and inconsistencies and to merge similar records based on business rules. This is the most obvious part of data mastering, but certainly not the whole solution. In addition, traditional approaches to data quality can have significant deficiencies depending on the data under consideration, as we will discuss below.

- Data integration: the ability to quickly map many dissimilar systems into a single consistent format and reconcile differences in structure, vocabulary, standards and even language. Without this essential 'smart plumbing', systems, including MDM repositories, cannot be effectively connected to each other. Traditional data integration can deal with the plumbing part but not the 'smart' part, which is usually attempted through hand coded scripts. A semantic approach to data mastering can alleviate this issue with a superior tolerance for data variability, as we will discuss below.

- Data governance: the ability to monitor the whole process and view metrics to determine basic health and identify opportunities for improvement. In essence, the control-centre for the data stewards who are working to enforce governance standards across the business.

- Data remediation: when reviewed against new standards, a significant percentage of data will usually fail in some way. In this case, identifying the error is not sufficient and there needs to be some sort of integrated mechanism, either manual or (ideally) automated, to remediate the data and bring it in line.

- Data repurposing: even in MDM, data is not needed in just one format so a complete data mastering toolset should be able to deliver the data in many different standards and even languages.

All these capabilities must be available in some shape or form to be able to deliver the core master data as and when required as part of either a one-off cleanup or as part of an ongoing MDM program.

Semantic-based data mastering

Semantic versus pattern-based technology

The core advantage to using a semantic-based approach to data mastering as opposed to a more traditional pattern-based approach is that semantic technology can more effectively handle the variability of less structured information. Whereas traditional pattern-based technologies rely on the presence of certain fixed syntax, semantic technologies are able to identify underlying meaning and extract pertinent facts irrespective of word order, punctuation or fixed vocabulary. This flexibility is imperative for almost any subject area beyond name and address validation and is particularly important for product-type data (products, items, SKUs, packages, assemblies, assets etc.) which rarely has a fixed syntax, comes from many sources and must be standardised, matched, merged and re-purposed from area to area.

Leaving aside that traditional data quality tools generally do not attempt to cover the full scope of a data mastering solution, traditional approaches to data quality are based on pattern recognition. That is, they will know that there is a specific pattern associated with a post code, for example, that is particular to the country you are concerned with. Then you can compare customer records (say) against this to see if their addresses match this pattern. If they don't then you have a data quality problem. More generally, you can attempt to see if what appear to be different customer records are actually the same by attempting to match against attributes such as name, address, postal code, phone number and so on, by comparing one customer's pattern with the next. Given that each of these individual comparisons may produce different match results then a combined score will provide you with a probability of a match. This is typically based on a combination of rules, which define the relative weightings of each individual part of the overall match process, and statistics.

In practice, while this sort of pattern matching is the de facto standard for customer data quality, it has a number of significant flaws. These include the fact that defining and maintaining rules is onerous (and doesn't improve over time) and that the allocation of weights is largely guesswork, amongst others. However, it is the implementation of the pattern matching within most data quality tools that is the problem rather than pattern matching per se. This is because there is, essentially, just one pattern for a name and address, for each country. However, when it comes to business entities such as products there is no single pattern to match against: there are hundreds, thousands, even tens of thousands of potential patterns. As a result, the whole pattern matching approach breaks down for anything other than the simplest types of product.

The basic problem is that product data is much more complex than name and address data and that there are no consistent ways of presenting the data. For example, the following are all descriptions of the same motor:

- 10hp motor 115V Yoke mount
- MOT-10,115V, 48YZ,YOKE
- mtr, ac(115) 10 horsepower 115volts with a 5 year warranty
- This 10hp yoke mounted motor is rated for 115V
- 10 Caballos, Motor, 115 Voltios
- TEAO HP = 10.0 1725RPM 115V 48YZ YOKE MTR
- Motor, TEAO, 1725 RPM, 48YZ, 15 Voltios, Montaje de Yugo, hp = 10

As can be seen, we are far removed from a single pattern: we have different abbreviations, different languages and the data presented in different orders. Not to mention additional information that is included in some descriptions. As a result, matching and cleansing this data requires a different approach that mandates semantic awareness that is focused on meaning (for instance, that mtr = MOT = Motor) rather than patterns across multiple languages, and is sequence independent. This is the advantage of a semantic approach to understanding data.

Note that it is not that pattern-based products cannot cope with product data. The vendors of such solutions have recently added text parsing and other capabilities to enable such functionality. However, what they cannot do is to cope well. Because traditional solutions do not understand semantics they are less likely to find matches—in a complex product scenario typically a third less than when using semantic approaches—which means that you require more human intervention and therefore cost. Similarly, weights and other elements of the environment have to be defined in rules that require human interaction as opposed to semantic solutions where iterative self-learning eliminates this necessity. Thus pattern-based solutions are less efficient (particularly for complex product environments) and more expensive, while semantic data mastering products are not only more efficient but less costly and, moreover, the costs per record processed will reduce over time as human intervention becomes less necessary, as the systems learns more of the rules of each data-set (which could be manually tuned or self-taught through inference—but that's another white paper).

The data mastering approach to MDM

It should be self-evident that high quality data is a necessary first condition for supporting MDM implementations: there simply is no point in MDM if the master data that you are managing is not fit-for-purpose. Yes, you may be able to share customer, product or other data across the enterprise on a consistent basis but if that consistency is based on invalid information then the benefit to the business will be zero. Indeed, it may even be damaging if people assume that the information is correct when it is not.

The question that might arise, however, is the time relationship that exists between data mastering and MDM: do you implement your MDM system first and then worry about quality, or is it a simultaneous process, or should you worry about data quality first?

To apply the Socratic method: what is the fastest route to value? We would suggest that it is in focusing on data mastering in the first instance. If you implement MDM first you won't get any value until you have also ensured your data quality, as we have already discussed; and the same applies if you attempt a parallel implementation. On the other hand there are direct business benefits that derive from having reliable, as opposed to error-prone, information, so if you start with data mastering the time to value on your investment will be reduced. Moreover, we could also argue that the greatest value in such a project actually accrues from the higher quality of data provided rather than through the ongoing management and provision of master data. Indeed, as much as 80% of the business value could accrue from only 20% of the cost, effort and time of a 'full' MDM program: certainly this will be true in the short to medium term and may also be true in the long term as well.

There is another reason why it makes sense to worry about data quality first and MDM second. This is that actually having top quality data may change your MDM priorities. If you are considering MDM and data quality holistically then one set of solutions may appeal to you whereas if you regard data mastering and MDM separately then the reasons for adopting a particular approach to MDM (for example, a lightweight registry-style approach as opposed to a more complex and time-consuming MDM hub) may change: the cost-benefit equation may be altered.

To be more specific, we are suggesting that it makes sense to consider the value of data mastering as something that is independent from MDM and, conversely, that the benefits of MDM should be evaluated excluding data mastering. If you do this you may find that your priorities are altered: you may decide that PIM is more important than CDI, for example. Indeed, in these recessionary times you may even decide to defer, alter or cancel broad MDM plans. In particular, if the data mastering solution chosen has facilities to support data governance in general and data stewards specifically, then it could well be that data mastering is all you need.

Finally, a word of caution: providers of MDM solutions have varied approaches to data quality and the broader notion of data mastering. Some companies provide their own, typically older generation, pattern-based data quality solutions while others either leave it to you to sort out for yourself or support a partner ecosystem. We strongly recommend that you apply due diligence to the selection of both the data mastering aspects of your solution as well as the MDM component and, as we have indicated, the former should really be dealt with first and implemented first.

Conclusion

We are not arguing in this paper for the benefits of implementing data quality or indeed data mastering solutions. We believe that the costs associated with poor levels of data quality are (or should be) so well-known by now that the adoption of some sort of formalised approach to data quality or data mastering should be considered mandatory. On the other hand what we are arguing is that that semantically aware data mastering tools should be preferred to conventional pattern-based approaches, especially when it comes to complex matching and cleansing requirements such as those involving product data; and that, at a minimum, data mastering should be treated as a necessary precursor to all Master Data Management initiatives.

Further Information

Further information about this subject is available from http://www.BloorResearch.com/update/1004

Bloor Research overview

Bloor Research has spent the last decade developing what is recognised as Europe's leading independent IT research organisation. With its core research activities underpinning a range of services, from research and consulting to events and publishing, Bloor Research is committed to turning knowledge into client value across all of its products and engagements. Our objectives are:

- Save clients' time by providing comparison and analysis that is clear and succinct.
- Update clients' expertise, enabling them to have a clear understanding of IT issues and facts and validate existing technology strategies.
- Bring an independent perspective, minimising the inherent risks of product selection and decision-making.
- Communicate our visionary perspective of the future of IT.

Founded in 1989, Bloor Research is one of the world's leading IT research, analysis and consultancy organisations—distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services and consultancy projects.

About the author

Philip Howard
Research Director - Data

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up what is now P3ST (Wordsmiths) Ltd in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director. His practice area encompasses anything to do with data and content and he has five further analysts working with him in this area. While maintaining an overview of the whole space Philip himself specialises in databases, data management, data integration, data quality, data federation, master data management, data governance and data warehousing. He also has an interest in event stream/complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to www.IT-Director.com and www.IT-Analysis.com and was previously the editor of both "Application Development News" and "Operating System News" on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and published a number of reports published by companies such as CMI and The Financial Times.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master) and walking the dog.

Copyright & disclaimer

2nd Floor,
145–157 St John Street
LONDON,
EC1V 4PY, United Kingdom

Tel: +44 (0)207 043 9750
Fax: +44 (0)207 043 9748
Web: www.BloorResearch.com
email: info@BloorResearch.com