



BroadstreetData

Sustainable
Data
Intelligence



master data management and data integration: complementary but distinct

A White Paper by Bloor Research
Author : Philip Howard
Review date : October 2006

Put simply, if you ignore data integration or do not treat it as being of sufficient importance then your MDM project will fail.

Philip Howard

This white paper discusses the relationship between master data management (MDM) and data integration. However, before we do that, we need to understand what we mean by MDM on the one hand and data integration on the other. We also need to understand the role of technologies such as customer data integration (CDI), product information management (PIM) and global supplier management (GSM).

MDM is an increasingly popular technology because it enables the enterprise to understand customers, products, suppliers and other business entities in a more complete and holistic manner, thereby enabling more advantageous interactions with external bodies and better control over internal company assets.

In most large organisations there are multiple applications that deal with the same sort of information, whether that is customers, suppliers, products or whatever. However, that information is often not synchronised and is distributed across the organisation. MDM aims to bring this information together (as we shall see, there are various ways to do this) so that you can see all details about customer A, for example, or product B, in one place at any one time.

With respect to CDI, PIM and GSM, these are all specific instances of MDM. That is, master data management is the term used for a generic platform that supports all of these solution types and it is also used as an umbrella term without necessarily implying any such platform.

Moving on to data integration, this consists of several technologies. However, between them they basically do two types of things: they either move information from one place to another or they check and assure the validity of the data. It should be clear that MDM cannot work without some means of moving data, as otherwise you could not form a

consolidated view of your customers across, say, 15 different sales order processing systems. Note that this is regardless of whether you move all of the data into one place first and then consult the consolidated view or if you do this dynamically, collating this view when, say, the customer calls into a call centre.

Secondly, it should also be obvious that if some of the information you have about your customers is invalid then it is effectively useless. Indeed, it may even be worse than useless since it may lead to making a sales proposal that puts the customer off rather than encourages him. In addition, the validity of your data is essential not only for best business practice but also to meet regulatory requirements such as Sarbanes-Oxley.

So, it should be clear that data integration, both for moving the data and ensuring its quality, is a fundamental requirement for any successful MDM implementation. This is not in dispute. However, there are two issues: the first is the role that data integration plays within MDM, which we believe to be too easily ignored or not fully appreciated and, secondly, it seems to us that less thought has been put into the relationship between data integration and MDM, and whether they should be treated as an integrated whole or if it would be better to treat them as separate layers within a stack. While either approach can work in any one implementation, this white paper contends that from a general-purpose perspective it is better to treat data integration as a set of enabling technologies for implementing MDM rather than as a part of MDM per se.

In this paper we will discuss the different ways in which MDM may be implemented, the role that data integration has to play (which may differ according to the approach taken to master data) and expand on why we believe that data integration should be regarded as a separate set of technologies.

While the executive summary for this report was intended for the general-purpose reader, this section, together with those that follow, is intended for readers that are already familiar with the concepts behind MDM and are aware of its benefits. We will not, therefore, be discussing why you might want to implement MDM or what its underlying concepts are but instead dive right in and start by discussing different types of MDM solution.

Breadth of data

As we have intimated previously, there are different ways that MDM may be implemented, depending upon what you want to do. In particular, these differences can be categorised into three main groups, as follows:

- **Master Identity** where management is of the system of record for keys only, with the inclusion of a few specific attributes solely to aid the determination of uniqueness.
- **Master Record** where management is of the unique key and the definitive best record of master attributes that has been resolved from the composition of one or more defined source authors.
- **Master Data** where management is of unique keys, the definitive best record of master attributes and additional transactional information related to the entity.

For an example record, like a customer, product or other entity

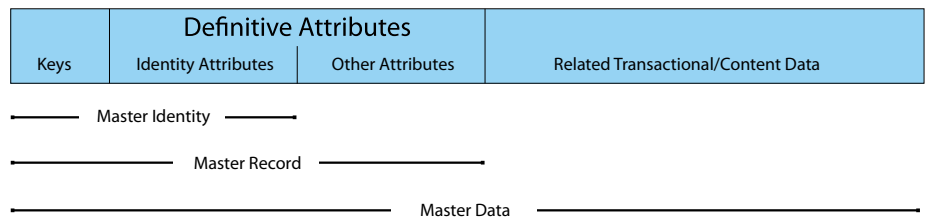


Figure 1: MDM solutions differ in data scope

Note that hierarchies and relationships between entities are a key requirement (no pun intended) and are assumed to be included in all three categories of master information.

Architectural approach

Just as there are multiple acronyms for MDM (CDI, PIM and so on) so are there multiple terms emerging to describe the architectural approach of how data is authored, stored and accessed. We believe there are 3 fundamental architectures to MDM:

- Master Registry:** This MDM approach holds a central registry of unique identities and a reference mapping to the participating applications that own entities. In this situation, the participating applications retain authorship for creating and updating the individual entities. This approach physically holds the smallest amount of master data, the key, with cross-reference mapping to enable other tools and applications to leverage the unique identity in translation and to access or bring other data together via the unique identity.
- Master Repository:** This MDM approach holds a central repository of unique identities together with the resolved 'best record' of attributes from participating applications that own entities. In this situation, the participating applications author segments of the best record, while the repository creates its own new composite master record that is used by the organisation as the definitive unique definition. This approach includes the registry and physically stores more master data than the registry.
- Master Hub:** This MDM approach takes authorship for master data away from the participating applications and creates a central hub of identities, attributes and transactional data associated with each unique definition. In this situation, the hub may replace participating applications and/or provide the basis for new applications to be built.

	Registry	Repository	Hub
Data Scope	<ul style="list-style-type: none"> • Master Identity 	<ul style="list-style-type: none"> • Master Record 	<ul style="list-style-type: none"> • Master Data
Authoring	<ul style="list-style-type: none"> • De-centralised • Other applications author entity records and their data • Other applications remain independent to MDM 	<ul style="list-style-type: none"> • Hybrid • MDM creates new composite record and synchronises • Co-dependency between other applications and MDM 	<ul style="list-style-type: none"> • Centralised • MDM authors master record for other applications • Others are replaced or dependent on MDM
Strengths	<ul style="list-style-type: none"> • Non-intrusive • Light footprint • Identity is high value 	<ul style="list-style-type: none"> • Master record readily available • Trusted "best" source • Harmonises master record with other applications 	<ul style="list-style-type: none"> • Forces reconciliation of sources • All master data readily available • Eliminates on-going source discrepancies
Challenges	<ul style="list-style-type: none"> • Requires EII to gain broad customer view • Fails to address authoring inaccuracy at source 	<ul style="list-style-type: none"> • Potential conflict of rules • Hard to keep current • Requires EII beyond the core set of attributes 	<ul style="list-style-type: none"> • Disruptive technology • Blurs boundary to transactional purpose • End benefit must outweigh the large endeavour
Deployment	<ul style="list-style-type: none"> • Reference for others 	<ul style="list-style-type: none"> • Augments others 	<ul style="list-style-type: none"> • Replaces others
Primary Uses	<ul style="list-style-type: none"> • analytical • Look-up reference 	<ul style="list-style-type: none"> • synchronisation • Consistency across data stores 	<ul style="list-style-type: none"> • operational • Composite Applications

Figure 2: Approaches to MDM and their attributes

No architectural approach is better than another as their applicability is dependent on the MDM purpose, the existing IT infrastructure in place, and the propensity of the business to absorb significant change. However, a simple effort versus return equation suggests the law of diminishing returns as the volume of data associated to the master grows.

Implementation

It should be clear from the preceding discussions that there is no right way or wrong way to implement MDM. Moreover, it should also be apparent that there are problems with all of these different approaches, even though these may pale into insignificance compared to the benefits to be derived from implementing MDM. However, it is important to understand these drawbacks because they impinge on the use of data integration within the context of MDM.

To put it briefly (the details are provided in Figure 2) both registry and repository-based approaches suffer from insufficiency. That is, they do not do all that you might like to do, such as addressing authoring inaccuracy, resolving potential rules conflicts and so on. On the other hand, deploying a hub-based approach is a much more costly and lengthy exercise. Typically, a hub-based implementation will be measured in months if not years. A year to 18 months might be an appropriate timescale. Conversely, a registry or repository-based solution can be implemented within a matter of weeks: six to eight might be a reasonable estimate. In other words, a hub-based approach will typically take an order of magnitude longer than either of the others and this probably applies to the total costs involved also.

We should say that this is not an argument against hubs: they provide a more complete solution and you pay for that 'more complete' solution. Nevertheless, what this does suggest is that some companies—in our view, many companies—may opt to start with a registry or repository and then, if appropriate, may migrate to a hub later. That way, you get initial benefits in the short term to justify the whole project.

However, if this two-phased approach becomes commonplace, which we believe it will, then there is a natural corollary. This is that organisations may invest in a registry or repository now and then a hub at some time in the future but, and this is the key point, the company may not choose a hub solution from the same vendor that provided the initial software.

There are, basically, two approaches:

1. Invest in a single MDM project that starts with a registry or repository and then moves straight on to a full hub implementation. In other words the registry/repository is simply a phase in a larger process. We think that this will be a minority approach.
2. Invest in a registry or repository now and revisit the issue of a hub in due course. This has the advantage that you can consider the use of products today that do not have hub-based capabilities (though they may do so in the future). In other words, you can choose a best-of-breed solution and not be limited to the relatively few vendors that offer hubs (and also assuming that they support registry/repository style approaches).

We believe that the arguments in favour of the second approach are sufficiently strong that this is the course that most companies will follow. This has important implications for data integration. If there is the possibility that you may choose a different vendor for your hub then if data integration capabilities are embedded within your MDM solutions then you will have to redo all the integration work that you did initially, because it will not be portable to the new environment. Note too, that in some cases you may not have the intention of adopting a hub at all but subsequent actions (a takeover or merger, for example) may change your position so that a hub becomes more attractive.

We believe that this reuse argument is compelling. However, it is not the only such argument.

A further MDM trend

While we think that a major trend in MDM will be towards starting slowly and moving iteratively towards a hub-based approach rather than going for a 'big bang' approach, there is also another major trend within the MDM market that we need to explore. This starts from the fact that, at present, almost all users and vendors are adopting MDM on a piecemeal basis. That is, they are implementing or offering customer data integration or product information management or whatever. What almost no-one is doing is to implement a platform with generic MDM capabilities that supports relevant master data services for customers, products, suppliers et al. In other words, current implementations consist of siloed applications that are effectively divorced from one another. This is very dangerous and needs to be stopped. Fortunately, we (as an IT industry) have done this enough times in the past that this time the danger signs have already been recognised and the major hub providers, at least, have all stated that they are moving towards this platform-based approach. While this is encouraging, it does pose a number of questions:

- What happens if you have different MDM solutions from different vendors? Clearly, it will be unlikely that you will be able to buy a common platform. However, there is no reason why all solutions cannot be built on a common infrastructure, namely a common data integration platform where the same data movement, data federation and data quality rules and procedures may be applied.
- How long will it take to develop and introduce a common platform and how easy will it be to migrate to it from the existing software? While we can't answer the first part of this question we can say that it won't be soon. As far as migration is concerned, this is actually a broader issue: it doesn't just apply to migrating from one version of an application to another (which can be demanding enough) but also migration between systems after an acquisition, say, when the two companies may have significantly different MDM solutions.
- In view of the potential issues raised in the previous point would it make sense to opt for a simpler registry or repository-based solution now and wait before implementing a hub until after a common platform becomes available? If this option is selected then the implications for data integration are exactly as depicted in the previous section.

As an instance of this we know of one company that is now on its third MDM solution. First, it developed its own solution. Then it licensed a registry/repository solution from a vendor. Now it is implementing a hub-based approach from a different supplier. Fortunately, the company had taken the view that data integration was a separate topic from master data and it was able to reuse the work that it had done in this area across its different implementations.

- How widely will you deploy master data? Potentially, every distributed application in your organisation is suitable for the use of MDM. Clearly the benefits associated with such an implementation will vary widely by both application and organisation but it is likely that there will be more and more use of MDM across the enterprise as the years go by. This has important implications: first, you might well decide that some applications merit a hub while others warrant only a registry. In other words, we can expect to see hybrid environments with different data integration requirements. Secondly, there will be many specialist providers of solutions for MDM in particular application areas. This suggests that many organisations will end up with multiple providers of MDM solutions. Where this is the case, then a common MDM platform will not be deployable universally whereas a common data integration platform could be.

To conclude this section: there are a variety of ways in which MDM may be implemented and all of them involve some degree of uncertainty over the future. Even if you think you know what you are doing today, that may not be the case tomorrow. These uncertainties militate towards adopting a solution, or series of solutions, that are as flexible as possible. Such flexibility can only come about by recognising that as much of the infrastructure as possible should be divorced from the MDM solution per se. In other words, data integration (in its widest sense) should be seen as an enabling (and necessary) technology for MDM but not a part of MDM. Nevertheless this essentially negative argument is not the only one that is relevant to our taking this position: there are also issues such as data governance and data access.

Data governance is the general term used for that set of technologies, methods and procedures that ensures that all of the organisation's data is fit for purpose, accurate, secure, audited, and suitably managed so that its use conforms to relevant government regulations (including, but not limited to, Data Protection legislation and Sarbanes-Oxley).

In practice, data governance is much broader than IT. It typically involves the establishment of a data governance council, the appointment of both data and domain stewards, and the creation of relevant policies and procedures that ensure the quality and conformity of data. On the technology front it involves a number of IT functions that include data security and monitoring, data auditing, data quality and data lineage amongst others. Master data management can also be considered to represent best practice as a part of data governance though data governance can, at least in theory, be achieved without it.

In so far as this paper is concerned, there are three aspects of data governance with which data integration platforms are concerned: quality, auditing and lineage. We will deal with each of these in turn.

Data quality

Data quality is sometimes built directly into MDM solutions rather than treated as a part of the infrastructure. In terms of ensuring the quality of the data within the MDM environment there is no theoretical reason why this shouldn't work absolutely fine. However, the MDM environment is not an island. The purpose of data governance is to establish high quality data in the first instance and then to maintain it. There is no point in cleansing all of your data and then having to cleanse it all over again, and then again and again. What is required is a set of processes that prevents, or at least minimises, incorrect, invalid or incomplete data in the first place. This needs to work in conjunction with the data quality tools themselves and is one of the main reasons for establishing a data governance project in the first place. Data quality should continue to be active on a permanent basis but it should, once proper procedures are in place, be in tick-over mode rather than running at full bore so to speak. However, data quality embedded within MDM treats this subject as if it was isolated and only had its own thing to do and was not part of a greater whole.

Data auditing

Data auditing is the recording of who did what to the data and when. In terms of the current discussions both MDM applications and data integration solutions would be expected to provide auditing capability. However, the levels at which this auditing will be applied will typically be different. This is because applications, necessarily, are concerned with auditing at the application level whereas data integration is more interested in where data came from: that is, its provenance. For a full auditing solution, you really need both. Of course, it is feasible that an MDM solution provider with data integration built-in rather than as a separate part of the infrastructure might build this sort of capability but there must be a strong likelihood that it will not.

Data lineage

Data lineage is really an extension to data auditing, at least in integration terms, because it is not just concerned with where the data came from but also what was done to it en route. That is, all the aggregations, transformations, changes and so forth that were applied to get the data into the state in which it currently resides. This is a necessary requirement to meet regulations such as Sarbanes-Oxley. The question here is whether the vendor of an 'all-in-one' solution will go to this necessary depth in order to provide such capability. It should also be noted that without data lineage you cannot do such things as impact analysis (for example, "if I change this what else will be affected?") or offer where-used capabilities.

If uncertainty over MDM directions and the role of data governance are two major issues to be borne in mind when considering the relationship between master data and data integration, then access to the data is a third. There are really three points to be considered here:

1. A requirement, particularly in hub-based environments, is that you need to be able to update the master data from source systems easily and quickly. You could, of course, write programs to do this. But this would be complex, expensive and unnecessary. A more practical method would be to use change data capture (CDC) as a mechanism for this purpose. This works by reading the logs of the source systems and whenever a relevant change is detected this can be automatically propagated to the MDM system. However, the people who specialise in this, and have the widest range of such adapters, are the data integration vendors.
2. One of the requirements for registry and repository style MDM solutions listed in figure 2 is EII (enterprise information integration), also known as data federation, which is the ability to access heterogeneous front-end data sources in real-time. This is an integral function for most leading data integration platforms. Because data integration vendors are intent on addressing as many data sources as possible they typically have a wide range of such sources that they can address, which would not be limited to the usual suspects. This may not be the case if these facilities are built into the MDM solution.
3. Master data management solutions typically assume that all existing data is in relational databases. This may not always be the case. You may, for example, have important information that you need to use, which is held in a spreadsheet. This is common in product lifecycle management, for example. So unless the MDM solution has the ability to synchronise with this and other, unstructured and semi-structured, data sources then it may be limited in its utility. A comprehensive data integration platform, on the other hand, is likely to have such capabilities.

Data integration has much broader applicability than master data management. It is used for moving data into data warehouses either in batch mode or real-time, for real-time query support using data federation, for synchronisation and replication, for migration of both applications and data, for conversion of semi-structured data (for example, from an EDI message format to XML) and so forth. If you are a large enterprise you should, ideally, have standardised on a particular data integration platform (because of economies of scale, reduced decision times, minimised re-training requirements and so on). If that is the case, why not use it to support your MDM solution? Even if you haven't, does it make sense to introduce yet another integration capability into your organisation, with all the ensuing costs that that implies?

There are two things that are important to appreciate in so far as this paper is concerned. The first is how important data integration is to master data management. Put simply, if you ignore data integration or do not treat it as being of sufficient importance then your MDM project will fail. That will cost you a significant sum of money. In particular, take heed if you are trying to implement a hub because that sum of money will probably be so large that it will quite possibly cost you your job.

The second thing that is important is to understand the relationship between master data management and data integration and why the latter should be treated as a set of technologies in its own right rather than simply as a part of MDM. It is not that you cannot successfully implement master data based upon a solution in which data integration is embedded but that such a solution will be less flexible going forward. As we have seen, flexibility is not all there is to this argument, but we believe that in these days of the agile enterprise any lack of flexibility is to be avoided at all costs. By all means consider an investment in a data integration platform at the same time as you ponder upon your MDM solution provider, just do not confuse the two: as long as they work together in an appropriate manner then they should be two separate decisions.

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

This document is subject to copyright. No part of this publication may be reproduced by any method whatsoever without the prior consent of Bloor Research.

Due to the nature of this material, numerous hardware and software products have been mentioned by name. In the majority, if not all, of the cases, these product names are claimed as trademarks by the companies that manufacture the products. It is not Bloor Research's intent to claim these names or trademarks as our own.

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.



BroadstreetData

Sustainable
Data
Intelligence



Suite 4, Town Hall,
86 Watling Street East
TOWCESTER,
Northamptonshire,
NN12 6BS, United Kingdom

Tel: +44 (0)870 345 9911
Fax: +44 (0)870 345 9922
Web: www.bloor-research.com
email: info@bloor-research.com