

# White Paper



## Pervasive Data Quality

Improving business processes with high quality data

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

For data quality to be truly pervasive it cannot be limited to a single business domain such as customer, financial, product or asset data but must be applied across the organisation to all areas

**Philip Howard**

## Executive summary

---

The problems associated with poor data quality, which include lack of trust in the data, poorly supported decision making and significant costs, have been well known for some time. However, even within those companies that have recognised the problem and which have taken steps to cleanse their data, there are still major data quality issues. This is for two major reasons. Firstly, most data quality programs are retro-active, which means that there is a time lag between the origination of the data and its validation. In an increasingly real-time and 24 x 7 world this is no longer acceptable. Secondly, the actual process of cleansing the data is usually the responsibility of IT, but IT does not understand either the meaning or the value of the data to the business. As a result, IT has to refer back to business analysts and others that do understand the data, on an on-going basis, which creates further delay.

In this paper we argue that a more progressive and pervasive approach to data quality initiatives is what is required, in order to overcome the delays and associated costs involved with current processes. To begin with, organisations need to expand the number of people with a remit for data quality, in the spirit of many hands making light work. For example, it is the business that owns the data, that uses it and which derives value from it. We therefore believe that business people should take some responsibility for data quality and get more involved in the process of managing it. In addition, organisations need to broaden their data quality coverage, by addressing more types of data across more applications and systems.

Expanding the “data quality team” to include more people who actually use the data day-to-day should be developed in two ways. Firstly, business analysts need to be able to fulfil data cleansing functions on the information that they are concerned with. Secondly, line of business managers need to know that the information they are using is fit for purpose or, if it is not, they need to be able to identify that fact and work with business analysts to get the quality of that data improved. To enable this, business managers will need a dashboard that can show the manager that quality is up to an appropriate level (for the data specific to that manager’s function), so that they can trust the data they are using.

Broadening data quality coverage needs to address the twin challenges of more data types and more applications. A more pervasive data quality program that spans all data domains, including customer, product, financial and asset data types, will ultimately deliver more benefits for the organisation. The benefits will be even more obvious if the program is applied to all applications. This sort of facility has actually been available for some time, but we would argue that this was merely a first step in moving towards pervasive data quality. Data quality needs to be assessed at the point of entry into applications, and inaccurate data needs to be rejected and/or cleansed at that stage, rather than leaving it to pass through the system. What is important is that prevention is better than a cure.

Needless to say, all these capabilities need to be provided with minimal (preferably no) impact on existing processes and with a focus on ease-of-use, in order to prevent training requirements from spiralling out of control. Currently, in IT-centric data quality implementations data quality tools and capabilities are used by perhaps five or ten people. In this new environment, on the other hand, we expect hundreds of members of staff to touch (if ever so lightly) data quality processes and functions. With wider involvement comes greater awareness and more attention; the net effect should be improved quality on the one hand and reduced costs on the other.

Put this altogether and we have what we refer to in this paper as pervasive data quality. In this paper we will consider what we mean by this, what it involves, how it is used and by whom, and what sort of characteristics of a data quality solution are required over and above standard facilities for supporting quality initiatives. For data quality to be truly pervasive it cannot be limited to a single business domain such as customer, financial, product or asset data but must be applied across the organisation to all areas. Note that this is not intended to be a technical paper and we do not discuss detailed technical requirements. We do, however, assume a working knowledge of the concepts behind, and the tools used for, data quality purposes.

## What is pervasive data quality?

---

The verb 'to pervade' means to diffuse. So the idea of pervasive data quality is that an awareness of the importance of data quality is diffused throughout the organisation and, as a corollary, that there are appropriate facilities in place to ensure that quality. There is thus a cultural element to pervasiveness: staff at all levels within the company need to be aware of the importance of having good quality data, although any discussion of how to achieve this is outside the scope of this paper, at least in so far as this is a political and cultural issue within the organisation, rather than a technical one. However, it is worth commenting that some businesses actively work against their own interests in this respect. For example, it is common that workers in call centres are incentivised on the basis of the number of calls they get through. If this is the sole measure of their effectiveness then it will actively work against ensuring good quality data: it is faster to enter a default value, when that is permitted, compared to actually collecting and entering true values. If data quality is important to your organisation then one of the metrics used for remunerating call-centre operatives should be accuracy of data collection.

The principle behind pervasive data quality is that you should be able to trust all of your data, for whatever purpose you are using it (that is, in any application) and in whatever (financial, customer, product and so on) domain, on a permanent and on-going basis. Achieving this is not simply a cultural and organisational issue: it is also important to understand how it applies in different areas within the company. As we shall see in due course, what is an appropriate level of data quality in one department may be unacceptable in another. Generally, outside of the data stewardship and other specialised roles that need sophisticated data cleansing, profiling and other capabilities, what business users need is just enough data quality functionality to support them in their day-to-day activities. Moreover, this should be deployable without requiring any specialised training and without interfering in any way with the way that the user does his or her job.

## What is entailed?

Before we discuss data quality requirements it is pertinent to point out that there is no such thing (in most instances) as 100% data quality. Even if it was possible to attain such a level at reasonable cost it would be impossible to maintain it. This is because things change. For example, it is estimated that customer data deteriorates at 1.4% per month. This is due to the fact that people die, get married and change their names, change email addresses, move house and so on, without telling you. In this context knowing how good or bad your data is, and being able to share that information across the organisation, is essential. In the sections that follow we will briefly discuss this collaboration (enabled by monitoring) followed by the four major aspects of data that need to be focused on: accuracy, completeness, consistency and timeliness.

### Data quality monitoring

Robust data quality monitoring capabilities are necessary if data quality is to become pervasive. This typically means the deployment of dashboards, both for data stewards and line-of-business managers, where the former is general across the domain managed by the data steward and the latter is specific to the manager's remit. This means that you can estimate the reliability of the data you (or your applications or processes) are basing your decisions upon and the extent to which you can actually use your data: for example, if you are conducting an email marketing campaign then you need an email address and if 20% of these are missing then you can only mail 80% of your customers. Similarly, you will want to know how timely your data is, how up-to-date it is, and so on.

### Accuracy

This is primarily a business function. IT can establish that it is likely (with an appropriate probability) that two customers are one and the same but only the business can say, ultimately, whether this is in fact the case. Similarly, it is product managers that will know that product x in the United States (say the Ford Escape) is the same as product y in Europe (the Ford Maverick) rather than these being distinct products. Accuracy is also about ensuring that descriptions are correct, that product hierarchies are valid and so on. When data quality processes are first introduced

into an organisation, the task of checking potentially large numbers of matches (we have seen users overestimate their actual number of customers by as much as 65%) is onerous. However, once initial cleansing processes have been completed, on-going requirements may be delegated to line management but more usually will remain the responsibility of a data steward or business analyst. We should also mention that there are some details that you cannot easily measure for accuracy, at least using your own resources. For example, you can check that a customer address is a valid address but that doesn't confirm that the customer actually lives there. We will come back to this point.

Unless you delegate on-going quality checking to line management the main difference between conventional and pervasive data accuracy is that the latter includes preventative capabilities such as data quality processes that are run at the same time as data is entered into an SAP R/3 application, for example. This helps to prevent quality issues getting into your system in the first place.

### Completeness

Do we have all the information we need about a customer, product or other entity? You can automate (an IT function) the discovery of missing, default or invalid data. Note that there is a distinction between invalid data and inaccurate data. The latter is valid (it is a real address) but may be inaccurate (the customer doesn't live there anymore). Also, because data is invalid it doesn't necessarily mean that it is not useful. For example, a phone number in an address field is invalid but you still want to know about it and the software should be able to recognise that this entry is formatted as a phone number. A secondary function of data quality software is to support what is known as enrichment. That is, to retrieve information from third party sources about the entity you are interested. This might mean accessing Dun & Bradstreet for corporate information, Experian for credit data, GS1 or other bodies for industry specific terminology and codes, geospatial sources for location-based information, and so on.

While some companies have already implemented data checking at the point of entry (mostly in marketing systems), many have not. Pervasive data quality requires that processes

## What is entailed?

---

are in place to ensure the completeness of the data, across all domains, when it is being entered rather than after the event. At the same time you can prevent things such as the entry of default values.

### Consistency

This is about ensuring that when you have multiple applications dealing with the same data, that that data is consistent across those applications. For example, you have customer data in CRM systems, sales order processing applications and so on. If these different applications are reliant on disparate data then when you run queries or reports against these applications you are likely to get different results. When people see different results returned from different systems then they won't know which to trust and will end up not trusting either: this is especially relevant to business intelligence, finance and compliance reporting.

This is primarily the responsibility of data stewards but it is also likely that business analysts will want to monitor that their applications are synchronised with others within the organisation that use the same data.

### Timeliness

Ensuring information is up-to-date is, to a certain extent, impossible: you cannot ensure that customers will tell you when they move house. However, you can implement rules that will address both business and technical requirements. For example, in this case, it might make sense to have a business rule that requires you to contact each customer at least once every six months: thus ensuring that customer information is at least up-to-date to that extent. At a technical (IT) level you would expect data integration tools, for instance, to report on a failure to load last night's data into the data warehouse, thus alerting you to the fact that data is not up-to-date for daily reporting purposes (say). Note that this emphasises the relationship between data integration and data quality: data integration processes need to deliver trusted data, and data quality needs data integration to deliver the data on time. More generally, and in a pervasive sense, timeliness is a matter of IT collaborating with the business.

While some metrics will be black and white (for example, has this customer been contacted in the last six months: yes or no) others are more nebulous. Moreover, you have to decide what level of accuracy is appropriate for your needs. This may be department specific. For instance, delivery needs 100% accuracy on delivery addresses (which is attainable, because customers want to actually receive their goods) but marketing may only require 90% or 95% for customer names and addresses. Thus you need metrics that are customisable by functional requirement at the business level. Note that best practice means identifying where poor data quality has the biggest business impact and addressing those issues (in implementation terms) first.

## Roles, responsibilities and roadblocks

We know what we have to achieve and we know that it needs to be pervasive throughout the organisation but what is actually required in terms of whom or what is involved, and what barriers are there to realising this? There are a series of considerations that we will discuss in turn. However, before doing so it is worth pointing out the contrast that pervasive data quality implies. Currently, in IT-centric data quality implementations data quality tools and capabilities are used by perhaps five or ten people. In a pervasive environment, on the other hand, we expect hundreds of members of staff to touch (if ever so lightly) data quality processes and functions. As this is, to a certain extent, a cultural issue, this may well require an educational programme that explains how poor data quality impacts on the enterprise at both a department and organisational level. This should mean a broader awareness of the importance of data quality and, in turn, should mean that greater attention is paid to it. The net effect should be improved quality on the one hand and reduced costs on the other.

Of course, the impact of a broader appreciation of data quality issues will vary according to role. Business analysts and data stewards will have less of the burden of ensuring good data quality bearing upon them, because more poor quality information will be filtered out before it reaches them, leaving them to focus on the more difficult data quality issues. On the other hand, while analysts and stewards may have less work to do, line-of-business managers will have marginally more with respect to data quality per se (because they will now be able to see the impact of poor data quality on their business processes) but this should be more than offset by the fact that they can immediately act (via the data stewards) to address critical issues that arise, and this in turn should lead to improved business processes and reduced inefficiencies. Thus the total burden on line-of-business managers should reduce. Even more broadly, the applications that you use to run your business will be more reliable, not just for line managers, but for anyone who has to use the data provided by these applications.

### IT

IT's responsibilities in terms of data quality are well understood. In particular, whatever can be automated is the domain of IT. This includes profiling the data to discover invalid and incomplete data, enrichment to move towards

more complete information, the discovery of relationships that exist across data sources in order to support consistency, facilities to support rules checking, the automated calculation of metrics and the initial process of matching. In this last case, some records can be automatically matched but there are often uncertainties involved which will need to be referred to the business for reconciliation. In addition, IT will collaborate with the business to establish data quality rules and implementation of those rules, regardless of whether they operate in batch or real-time. Further, IT should optimise these rules so that they run as fast as necessary within your data integration infrastructure.

### Business analysts

The role of business analysts in supporting pervasive data quality is primarily with respect to exception matching. That is, it is business analysts who will typically be responsible for making the reconciliation referred to in the previous section: for example, determining that 3M is the same as the Minnesota Mining and Manufacturing Co (or any of its other 200 or so variants). What they need for this function is to be provided with the relevant details of the records that may match by IT, preferably via a web-based front-end that allows them to access and update the exceptions that need to be matched for their domain. Since matching exceptions occur across multiple domains the software will need to know which analysts are responsible for which sets of data and it will also be necessary to alert the analysts whenever there are exceptions that they need to deal with.

### Data stewards

While data stewards are concerned primarily with governance, their actual role in the organisation, and their relationship with business analysts, is likely to vary by company. The main responsibilities of data stewards are concerned with governance (setting targets, defining data quality business rules, data profiling and monitoring data quality), ensuring the quality of reference data (which has similar requirements to standard data quality but is more static) and cross-application consistency. Of these, governance is the most contentious. Technically, governance can be taken to include data discovery, data quality, data security, information lifecycle management and compliance, and monitoring of all of

## Roles, responsibilities and roadblocks

these. However, in many enterprises, compliance and security, in particular, may be handled separately within other departments that have cross-enterprise responsibility for these functions. Whatever the specific characteristics of stewardship within your organisation, it should be clear that it involves considerable collaboration, with IT on the one hand and with business departments on the other.

### Auditors and compliance officers

It is not enough to be compliant. At least with respect to external regulations you need to be able to prove that your processes and procedures are compliant. In terms of data quality this means that you need to be able to provide features such as data lineage so that auditors (whether internal or external) can track how information has been processed. Similarly, you will need to be able to prove compliance with data protection legislation and that only authorised personnel have access to sensitive data. Where unauthorised people need to have access to data (such as outsourced development teams) then facilities to mask sensitive data will be required and you will need to be able to demonstrate that business rules are built into your software that ensures masking is applied whenever necessary.

There are also specific issues that relate to auditing in particular industries. For example, in the banking sector data quality may directly relate to the capital adequacy requirements mandated by Basel II and we have seen banks significantly reduce their capital requirements through being able to prove (thanks to monitoring and metrics) how good their data quality is, thus allowing improved risk assessment. The same thing is likely to occur within the Insurance sector with the introduction of the Solvency II regulations.

In general, what auditors and compliance officers require is access to monitoring tools. Thus, for the examples in the previous paragraph, they will want to see your metrics. Moreover, in the first instance they will also likely want you to demonstrate how those metrics are calculated. Note that the more you can provide this in an automated fashion, then the easier and faster that they can do their jobs and the lower your auditing fees are likely to be.

### Line-of-business managers

Line-of-business managers, whether application owners or business process owners (such as order to cash, shipping, marketing or other processes) need just enough information about data quality to support their functional responsibilities. Ideally, this means that data quality has reached an acceptable level (say, 95% accuracy) and that the manager's involvement is simply a question of checking that quality remains at this level, and that appropriate rules (contacting customers every six months) are being followed. If quality drops then the manager needs to understand how this will affect his business processes and he will then liaise with relevant business analysts or data stewards, or, if business rules are involved that are the direct responsibility of the manager, he or she will need to take direct action. In so far as provisioning quality information is concerned this will be limited to providing dashboards showing relevant metrics, which will enable managers to see not only what quality issues there are but how much that is costing them.

### Domains

For data quality to be truly pervasive it cannot be limited to a single business domain such as customer, financial, product or asset data but must be applied across the organisation to all areas. Historically, most vendors have focused on the first of these so it is important that any selected supplier can offer the sorts of facilities (which can be specialised) needed to support these other areas.

In a recent (2009) survey conducted by Information Difference it found that 81% of respondents had data quality initiatives focused more widely than just "name and address". Moreover, when prioritising domains of interest the results showed product data and financial data were of more interest than name and address data.

## Roles, responsibilities and roadblocks

---

“It is easier, faster, and less expensive to avoid errors at the outset than to try to get rid of them once they are in a computer file.”

Malachy J Foley, University of North Carolina

“The longer incorrect records remain in a database, the greater the financial impact. This point is illustrated by the 1-10-100 rule: It takes \$1 to verify a record as it’s entered, \$10 to cleanse and de-dupe it and \$100 if nothing is done, as the ramifications of the mistakes are felt “over and over again.”

2009 survey from SeriusDecisions

### Key applications

Pervasive data quality is not just about people but also about applications. Ideally, relevant applications should not allow incorrect or invalid data to be entered. While facilities to ensure that only good quality data is acceptable should be built into newly developed applications, in practice most companies use off-the-shelf software from leading providers that have no such provision. However, it remains important to prevent poor data quality. This is because it is much cheaper and easier to prevent problems rather than to try to fix them later. We therefore strongly recommend that you apply data quality at the point of entry, when you are entering data into your ERP, CRM or other applications. This typically involves a plug-in to your application provided by a data quality vendor that will check in real-time that a default value is not being entered, that the phone number put in is a valid one, that the product code is a real one, and so on. If data is being collected for entry via a call centre then it will also be useful if the software can suggest that this customer may actually be the same as an existing customer on the database (that is, a potential match) so that this can be checked directly with the customer.

### Inter-application firewalls

This requirement is similar to the previous one except that instead of ensuring data quality when a person is entering data into an application we are here concerned with the matter of passing information from one application to another.

### Third party and partner data

Here we do not have control over the person or thing entering data into our systems but the concerns are the same: we want to prevent, as far as possible, poor quality data from entering our environment. While you probably cannot interact directly with the third party (to establish matching, say) you certainly can implement checks for valid addresses, phone numbers and so forth.

## What do you need?

---

While the foregoing discussions should give a fairly clear idea of the sorts of facilities you need from a data quality provider in order to support pervasive data quality it is a good idea to specify these more precisely.

### Flexibility

In addition to a broad range of functionality for both profiling and cleansing it is also important that the tool suite is flexible. Particularly, when multiple people are interacting it is important that you can profile when you like, cleanse and match when you like, enrich the data when you like and so on. Moreover, data quality processes are not linear; you will often want to profile or re-profile the data, for example, several times within a single process. You thus need tools that are flexible enough to allow this.

### Comprehensive metrics

It will be useful if some metrics are provided out of the box but these need to be customisable and customers should be able to develop their own. In addition, metrics are not just about measuring data quality but also compliance with business rules that you have established as best practice (for example, contact each customer at least very six months). Further, the software should be able to provide trend information: is data quality improving, static or decreasing? If the last of these, you should be able to drill down to discover where there is an issue.

In some instances it may be useful to be able to derive risk performance indicators as a particular form of metric.

### Rules and exceptions

We have already noted that it is necessary to be able to monitor business rules but it is also important to have policies in place for such things as exception handling. There needs to be a formal process for handling matching, for example, where this cannot be done automatically, by passing the information over to a business analyst. This process itself needs to be monitored so that you can ensure that policies are being met and the software will likely need some sort of workflow support to enable this.

### Role-based tools

In an environment that supports pervasive data quality different people have different responsibilities. Not only do you need to monitor their interaction, you also need to provide each individual with the right tool for the job. This will vary from a detailed capability in IT, through business analysts and data stewards who have a significant involvement in the processes involved, to business managers whose involvement is relatively lightweight. Browser-based access, with no need for any installation, will be important for line-of-business managers in particular.

### Glossary

When you have managers, business analysts, data stewards and IT personnel all working in collaboration it is important that they can all understand each other. The difficulty is that they tend to use different terminology. It will be useful if a business glossary or semantic capabilities are provided that will enable each department to use its own terminology with automatic mapping between business and IT terms.

## Conclusion

---

There are a number of key points to be made about pervasive data quality:

- The importance of good quality data needs to be understood and appreciated throughout the organisation.
- Poor quality data should be prevented from entering the business environment wherever possible.
- Business managers need to understand how poor quality data impacts on their ability to perform.
- Business managers need to understand what level of quality is necessary for them to do their jobs most efficiently and what business rules are needed to ensure good quality data.
- Business managers need dashboards that enable them to monitor the quality of the data that impacts upon them, and easy-to-use and low-impact tools that enable them to define business rules where appropriate.
- Appropriate tools, such as workflow and glossaries, are needed to enable collaboration between the business and IT.

More generally, pervasive data quality is an idea that makes obvious sense. While traditional approaches to data quality represent a significant move forward, compared to having no control over quality, you are always going to be one step (or more) behind. That is, you are always reacting to poor quality data and you can't rely on that data until after it has been cleansed. By implementing a pervasive approach the intent is to prevent poor data quality where you can and, where you cannot, at least let you know how reliable or unreliable the data is. In other words you can approach your data on the basis of knowledge about its characteristics rather than guesswork. To achieve this means involving the business. Previously, such an approach to data quality has not really been possible since the necessary capabilities were not provided by the various vendors of data quality products. However, that is changing: we are now in a position where suppliers are recognising the need for pervasiveness and bringing to market appropriate tools to support this approach. We believe that pervasive data quality is the next step forward for all organisations concerned with quality issues and, given that every company should be so concerned, that means you.

### Further Information

Further information about this subject is available from  
<http://www.BloorResearch.com/update/1056>

## Bloor Research overview

---

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the whole story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the whole story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter "noise" and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channel.

Founded in 1989, we have spent over two decades distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

## 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](http://www.IT-Director.com) and [www.IT-Analysis.com](http://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 copyright © 2009 Bloor Research. 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. Likewise, company logos, graphics or screen shots have been reproduced with the consent of the owner and are subject to that owner's copyright.

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.



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](http://www.BloorResearch.com)  
email: [info@BloorResearch.com](mailto:info@BloorResearch.com)