Bloor

HyperReport

**HyperReport** by Bloor Research
Author **Philip Howard**
Publish date **February 2020**

# Data Assurance

# Data Assurance

# Executive summary

Assuring the quality and provenance of your data is vital for both operational and analytic purposes. As we use data more and more, and especially as we increasingly automate the processing of that data, we need the assurance that the data is both accurate, timely and complete. In other words we need to be able to trust the data we are using. However, there are two aspects to this: from an internal corporate perspective you need to be able to trust your data to support decision making as well as operational processes; but on the other hand customers need to be able to trust you to keep their data secure and to know that you will not misuse that data. This hyper-report considers the first of these aspects while a companion report on *"Managing Sensitive Data"* is forthcoming.

> **"The mission of American Family Insurance is to progress on its journey from a *"data-rich"* to a *"data-driven" company…* this transformation is fundamentally about a shift in thinking, *"We are helping make data thinking become business thinking."*"**

Historically, the focus when discussing data assurance has been the costs associated with poor data quality. For example, Gartner has suggested that, on average, each organisation loses $13.3m per annum thanks to poor data quality, while a third of companies cannot even estimate this figure because they don't track data quality. IBM has estimated that $3.1 trillion is lost, in the US alone, due to the downstream impact of making less than optimal decisions based on bad data. Research suggests that poor data quality also adversely affects both operational expense and labour productivity. However, these figures relate to what we might call *"traditional"* data quality concerns rather than considering a broader context. Today, many organisations are going through a digital transformation so ask yourself this question: *"if you can't trust the data that underpins your organisation, how can you transform your business?"*

In technical terms, since the advent of big data and the widespread introduction of data lakes, these same data trust issues have spread beyond conventional environments and solutions. In particular, data assurance is now also an issue for data scientists and others using data preparation tools. Moreover, there are also potential problems when it comes to machine learning and artificial intelligence, not only with respect to the reliability of decision making but also with the risk of reputational loss.

The problem, of course, is how to ensure that your data is trustworthy and, unfortunately, ensuring that that is the case is not a simple matter: there are a plethora of different technologies – sometimes overlapping, sometimes not – that can be used to support data assurance, and choosing the right tool for the right task is not a trivial process. Some of the areas these tools address, are relatively mature, while others are new and emerging. In particular, the rise of big data and, more recently, machine learning, have transformed the market, along with increased regulation and compliance driving risk-averse behaviour. Ethical issues are also increasingly raising concerns.

In practice, many data assurance issues are approached from the perspective of resolving a particular problem or use case. In such environments, the adoption of only one or two related technologies may be appropriate. However, this is a tactical approach where more strategic thinking may lead to a broader consideration of data assurance from a holistic perspective.

Data assurance has several major elements:

- **Discovering your data and its context.** Firstly, you need an inventory of all your data assets that supports the classification of those assets. There are several ways to do this but data cataloguing is an increasingly popular methodology. Secondly, you need to understand where directly related data is stored. For example, if you have half a dozen different databases storing customer data then you need to understand how the different – if it is different – customer data in each data store is related. This is the role of the data discovery aspect of data profiling technologies. And finally, you would also like to know how this customer data, say, is related to other types of data such as product data. Knowledge graphs are increasingly being used for this purpose.

- **Discovering poor quality data.** This is the role of data profiling or of the data profiling aspects of data preparation tools, which will look for things such as missing values, values out of range or other discrepancies that contravene data quality rules that you establish. In the case of spreadsheets and other end user computing resources, spreadsheet management and governance tools provide this sort of capability.
- **Ensuring the quality of your data.** There is a distinction here between matching and cleansing. The former is intended to resolve issues where duplicate or near-duplicate records exist, while cleansing is concerned with issues arising from broken data quality rules. Remediating these issues both fall within the realm of data quality tools as well as the data quality aspects of data preparation technologies. They also provide the ability to enrich data records with additional data from external sources. Note that product data quality, for example, has different characteristics from customer data quality. The same is also true of sensor data in Internet of Things (IoT) and industrial environments.
- **Monitoring data quality.** Data assurance is not a one-off activity. It needs continuous monitoring to ensure that data quality is maintained. Data profiling tools typically provide this capability and work with the data stewardship modules that vendors of data quality tools usually provide.
- **Assuring consistency.** When you have multiple systems involved with, say, customer data, then you need appropriate tooling to create a single customer view (SCV) or so-called *"customer 360°"*. However, if that SCV spans data across multiple data sources then you need to ensure that each of those sources is updated in a consistent manner whenever any one of those sources is updated. This is typically the role of Master Data Management (MDM).
- **Governance.** This is a broader concept acting as an umbrella function for all of the above, but also covering business functions such as how you introduce a new product or onboard a new customer. It also encompasses compliance and compliance reporting. While data and spreadsheet governance cover much of this, there are also specialised data lineage tools that provide capabilities that most data governance tools do not.

Note that this categorisation is not definitive: there are overlaps within these elements. For example, data preparation products typically include profiling capabilities and some elements of governance, as well as collaborative capabilities. Moreover, we have not mentioned bias in training data for machine learning environments, which is a subject all to itself. Whichever way you look at it, this is a complex environment, and this has implications for product selection. When situations are simple it is easy to opt for best-of-breed solutions. However, the more complicated the milieu the more you should be thinking about tightly integrated tools and platforms. Indeed, there are vendors that do not necessarily excel in any one particular discipline, but whose solutions fall more broadly under the category of information management suites that you might select because of their breadth of capabilities and the advantages that accrue from a single, holistic solution.

Regardless of the position you are in, this paper will discuss the relevant technology landscape and all of the product areas discussed above. However, this is not intended as a detailed research document in the sense of providing guidance as to the individual products within each space. These are mentioned, but we are more concerned with the fundamental requirements of each technology under consideration. Links are provided to more detailed analyses published by Bloor Research and others.

# Data discovery

Like many other terms used within IT, data discovery is used in more than one context. In particular, some business intelligence vendors have historically described their capabilities as providing data discovery, though we would categorise this as – more accurately we feel – insight discovery. However, it is also used in a more literal sense, to discover where your data is, how data in different data sources is related and what those relationships are, and whether there are any dependencies that exist between data elements regardless of whether those elements exist with a single database or across multiple, potentially heterogeneous, data sources. While there are stand-alone data discovery tools within the sensitive data space, within the context of data assurance data discovery has historically been built into either data profiling or data cataloguing tools, and we will discuss these in turn. We will also discuss knowledge graphs, not because they help to assure the quality of your data, but because they provide context for that data.

## Data catalogues

Data catalogues provide the business analyst or data scientist with information about what data is available to them. A widely quoted figure is that 80% of the time required to develop analytic processes is spent in preparing the data for those analyses but it is also true that a large part of that period (another 80%?) is actually spent looking for the data you need. Data catalogues were originally developed, by companies such as Alation and Waterline, to abbreviate that process in data lakes, which were (and are) otherwise in danger of becoming data swamps.

Put simply, a data catalogue is a repository of information about a company's data assets: what data is held, what format it is in, within which (business) domains that data is relevant, and where it is (in which databases and/or files). The information within a data catalogue may be classified further by geography, time, sensitivity, access control (who can see the data) and so on. Data catalogues are indexed and searchable, and support self-service. They can be created in a similar manner to the way the Google provides a "catalogue" of web documents by using web spiders or other technologies to create a fully searchable experience. Business specific terminology can be derived from business glossaries.

To improve the quality of the catalogue "crowd sourcing" allows users to tag, comment upon or otherwise annotate data in the catalogue. Some products support the ability for users to add star ratings as to the usefulness of a data asset. The software will monitor who accesses what data and employs it in which reports, models or processes. If a user starts to search against the catalogue for a particular term, the software will make suggestions to the user about related data that other users have also looked at in conjunction with that term. Catalogues can also be useful in identifying redundant, out-of-date and trivial (ROTten) data that should be removed.

> *The data catalogue is extremely important because our model is to focus on self-service. It's really a matter of connecting the researchers and the scientists directly to the data.*
>
> GSK

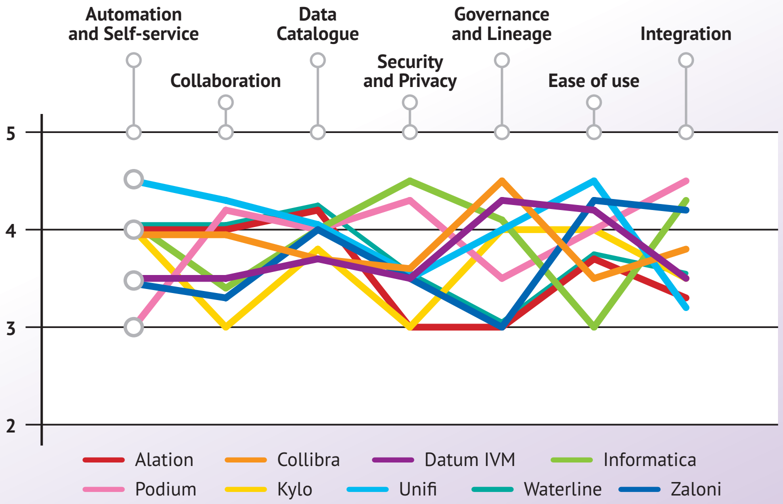


*Figure 1: Data Lake Management Tools – see also Figure 9*

As we have noted, data cataloguing was initially introduced specifically to support data lakes but the provision of catalogues has proliferated across both business intelligence/ analytic vendors and suppliers of data governance. Early providers include but are not limited to those shown in ***Figure 1***, a diagram extracted from Bloor Research's 2017 report on data lake management (see ***www.bloorresearch.com/research/data-lake-management-p1/***). Many more vendors have since entered this space, and since this report was published, DATUM (a data governance vendor) has been acquired by Infogix and Podium by Qlik. Kylo is no longer marketed. There has been a split in the market between those companies offering point solutions, either within their own products or being limited to data lake environments, and vendors such as Alation, IBM and Informatica that are focusing on enterprise catalogues that span corporate database environments.

In the context of data catalogues as an enterprise and not just a data lake resource, we should comment that some of these products have been expanded to address this space, challenging the traditional approach of using a metadata repository, enterprise data model (with tools such as Erwin, ER/Studio or Power Designer), a front-ended data dictionary and a query tool. These are heavyweight and complex approaches and data catalogues represent a significantly simpler and easier approach. For those companies with existing investments in metadata repositories, you might consider using a graph database in place of the enterprise model, dictionary and query tool.

> *(Using knowledge graphs) if you start looking at what kind of documents you have and how you're able to transform those into actionable knowledge for your end users, you can improve your decision making.*
>
> **NASA**

One problem with this proliferation of catalogues is that most of these are incompatible with one another and there is the danger of creating silos of metadata. In response to this, there are ongoing initiatives designed to overcome this issue. The most advanced of these is ODPi Egeria, which is an open source project running under the aegis of the Linux Foundation. Egeria provides an open metadata standard that allows metadata exchange across all relevant catalogues and repositories. There is a built-in graph database capability so that you can visualise metadata relationships across your computing estate and this in turn means that you should be able to define, for example, a data governance rule once and then apply that rule across your entire environment.

There have been multiple attempts to solve the metadata management problem, starting with AD/Cycle back in the 80s. These have all failed due to lack of buy-in by vendors, but Egeria has significant backing. It is currently supported by IBM, ING, China Mobile, SAP, SAS, Attunity (now part of Qlik), Syncsort, HortonWorks (Cloudera) and various others. We understand that some other notable vendors are on the verge of announcing support for Egeria also. The existing code is available on GitHub though it is arguably not ready for prime time yet. Nevertheless, we wish it every success.

## Data Profiling as discovery tool

Data profiling tools provide a range of functionality that includes data discovery, statistical analysis of data values, the identification of data quality issues, and the monitoring thereof. Here we are concerned with the first of these capabilities.

There are actually two things that you want to with data discovery. Firstly, you may need to discover data of a particular type. This is most obvious when you need to discover sensitive data but can also be used to find data that is subject to particular data governance rules (for example, discovering Legal Entity Identifiers [LEIs] or, more prosaically, customer or supplier data). Despite the demonstrations often proved by vendors, this is a non-trivial process: column names are often indecipherable and scanning through content and metadata may not help, even when semantic capability is provided. Additional techniques such as the ability to introspect stored procedures may be useful, but ultimately complete discovery will require manual intervention from domain experts.

Secondly, as noted previously, you may want to discover relationships between different data elements, for example to discover foreign key relationships across tables in different sources. More generally, features you would like to have include exception detection against discovered or pre-defined business and transformation rules, data validation, dependency analysis, overlap analysis, precedence analysis, the discovery of cross-source binding conditions, matching key evaluation, outlier analysis, clustering, sub-schema and sub-type profiling, recognition of join key values that match multiple times (which is an often overlooked reason for unexpected data duplication) and so on. Needless to say, a number of these requirements are only relevant in multi-source environments and, in this respect, support for non-database sources (Excel spreadsheets, CSV files, COBOL copybooks and so forth) is also important.

In addition to these uses there is also a question of scale. In large enterprises there may well be hundreds if not thousands of databases. Indeed, some organisations have tens of thousands of databases. In such environments it is often the case that no-one has any good idea what data is held where. Data discovery at scale is necessary to clarify these sorts of environments and you also need to be able to easily visualise the results. However, very few providers of data profiling software (Global IDs is one exception and Informatica is also targeting this market) – detailed in more depth later – focus on this level of scale.

***Figure 2*** illustrates an example of visualising data relationships across multiple databases, using Global IDs technology.

## Knowledge graphs

There is no agreed definition of a knowledge graph and how it differs from a regular graph on the one hand, nor how it differs from a knowledge base on the other. Several sources have suggested that knowledge graphs are domain specific but if you look at ***Figure 3*** (taken from a presentation by Neo4j, illustrating knowledge graph use cases) you can see that this is not necessarily the case. Here the graph encompasses the customer, product and supply domains but, and this is the key point, it recognises that these domains are linked, and that both customers and suppliers, for example, provide context to product data. If, for example, you want to build a Customer 360º application (a simplification of single customer view – see later – and another knowledge graph use case) you need to know what products they have bought in the past, and you would like to know which ones they have looked at on your website, and if they are influenced via social media, and who by. In other words, a knowledge graph is about discovering and understanding the relationships that exist between different data elements. It isn't about the quality of your data but it is about the context within which that data resides, Bloor Research would argue that understanding that context is as important to the analytic and operational processing of domain-specific data as is the quality of that data. We therefore define a knowledge graph as *"a graph that provides contextual information and understanding with respect to specific domain and cross-domain data."*

As far as we know, there have been no analytic reports comparing different providers of knowledge graphs. It is technically possible to support knowledge graphs without using a graph database but, to our minds, a graph database is the obvious place to start. That said, some graph database providers focus more on this area than others. Examples would include Grakn.ai, Neo4j, Ontotext and Stardog, amongst others. Given that this is a report on data assurance rather than on graph databases we are not including a discussion of these. However, for a detailed exploration of relevant graph database technologies, see ***www.bloorresearch.com/research/graph-database-market-update-2019/***.

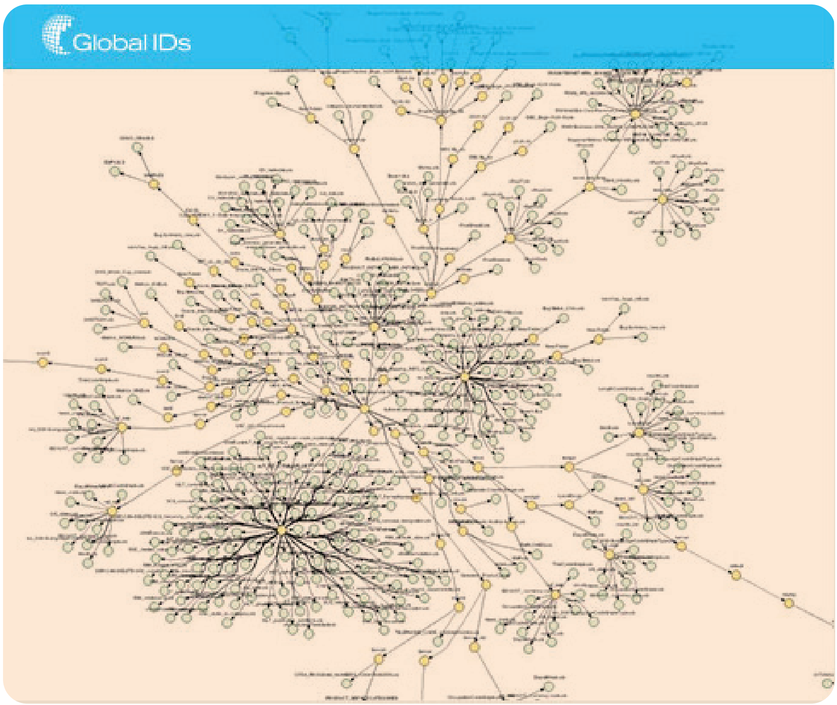


***Figure 2: Visualising data relationships across large IT estates***

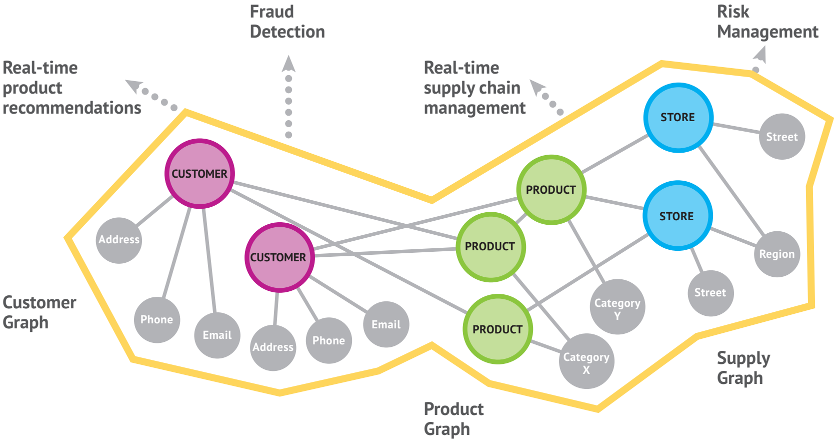
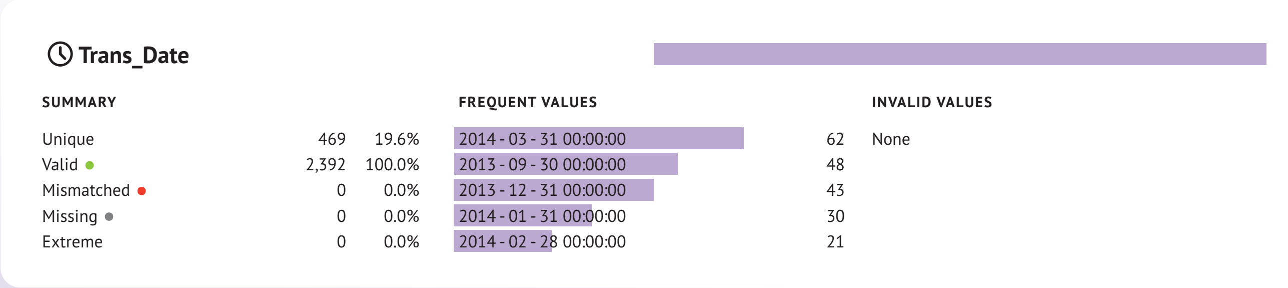


***Figure 3: Example conceptual view and use cases for knowledge graphs***

# Data profiling

Apart from the discovery capabilities already discussed, the primary role of data profiling tools is to statistically analyse – or *"profile"* – the data in whichever data source or sources you are investigating. It does this primarily through column analysis that generates a full-frequency analysis of the values within each column (maximum/minimum and so on), together with primary and foreign key analyses. Analyses are typically presented as histograms or bar charts – as in ***Figure 4*** – and these will highlight the uniqueness of distinct values, missing or null data, data that doesn't match the defined datatype for that column, invalid values that break business rules and so on and so forth.



**Trans_Date**

| SUMMARY | | | FREQUENT VALUES | | INVALID VALUES |
|---|---|---|---|---|---|
| Unique | 469 | 19.6% | 2014 - 03 - 31 00:00:00 | 62 | None |
| Valid ● | 2,392 | 100.0% | 2013 - 09 - 30 00:00:00 | 48 | |
| Mismatched ● | 0 | 0.0% | 2013 - 12 - 31 00:00:00 | 43 | |
| Missing ● | 0 | 0.0% | 2014 - 01 - 31 00:00:00 | 30 | |
| Extreme | 0 | 0.0% | 2014 - 02 - 28 00:00:00 | 21 | |

***Figure 4: Example bar chart showing the results of data profiling***

More technical considerations are concerned with where the profiling takes place, and against which sets of data. Ideally, you would like the option of profiling in situ or by extracting the data, with discovery run against all of the data, or a sample, as required. Which is most suitable will depend on the number of sources, their complexity and the task you are trying to achieve. Flexibility will mean that the tool is more suitable for a wider range of tasks. If you are going to use data profiling as a part of broader data quality initiatives, then you should be able to run data cleansing and matching routines without having to re-parse the information that you have already parsed for profiling purposes.

Data profiling is, or can be, an important collaborative tool. It is typically business analysts and domain experts who are best placed to validate business rules, for example, but on the other hand much of the information that is uncovered by data profiling is also of value directly to developers and to data management. It will be helpful therefore, if the product has functionality that will assist both of these constituencies. Support for a business glossary, an understanding of semantics, the discovery of attributes (constraints, reference data and so forth) that may be of value to an analyst, workflow capability and the ability to visualise discovered relationships through entity-relationship diagrams (or something similar) will be useful.

On the statistics side, you would like to be able to distinguish between hidden sub-types. By way of illustration suppose that you have a table of financial instruments containing data on both bonds and equities, including a column for maturity date. A bond has a maturity date and thus must not be null, but an equity doesn't, so it must be null. Simply reporting the number of nulls is not enough.

Another major issue is that if you are checking rules about your data then most tools will simply tell you about any exceptions that have occurred. However, some tools cannot cope with multiple rule violations. What you would really like to know is what percentage of records have no violations, one violation, two violations, and so on. Going a step further, you would also like to monitor this over time and be able to compare these figures with a baseline to get comparative confidence levels for the data.

There are multiple aspects of data assurance: the identification of potential issues with the data, the cleansing of that data, record matching and curation, enrichment and verification, and the ongoing monitoring and stewardship of the data. Historically, data profiling tools have provided the first and last of these, while *"data quality"* products have supported cleansing, matching and enrichment. Moreover, there have been a number of vendors offering one or the other but not both.

These have all but disappeared and the trend is towards providing an integrated experience across all of these capabilities. Nevertheless, it is worth considering the requirements for data profiling separately from those of data quality tools.

Finally, the statistical analysis that profiling provides means that it can be used to monitor data quality on an on-going basis. For example, you may decide to cut-over a data migration project only after data quality metrics have exceeded a particular threshold: in this case you will therefore also need dashboard capability and the ability to capture or use quality metrics. This is commonly provided by relevant vendors to support data stewardship. While there used to be stand-alone data profiling vendors, these have all but disappeared (Datiris is an exception) and data profiling nowadays is built into either data preparation tools or data quality suites, both of which are discussed below.

# Data quality

The measure of a data's quality is the extent to which it meets the requirements of the business processes that it supports. Not having high quality data can be extremely expensive: see the Executive Summary. However, measures of quality are not fixed and are use-case dependent: in a nuclear power plant data must be 100% reliable, when it comes to mailing out marketing material it is not quite so critical. Moreover, there are many scenarios where it is simply not possible to ensure completely accurate data: people move from one house to another and forget to tell you, while sensors can lose power and fail to transmit data when they should. In practice this means that user departments need to establish acceptable rules for the quality of data, depending on what that data is used for.

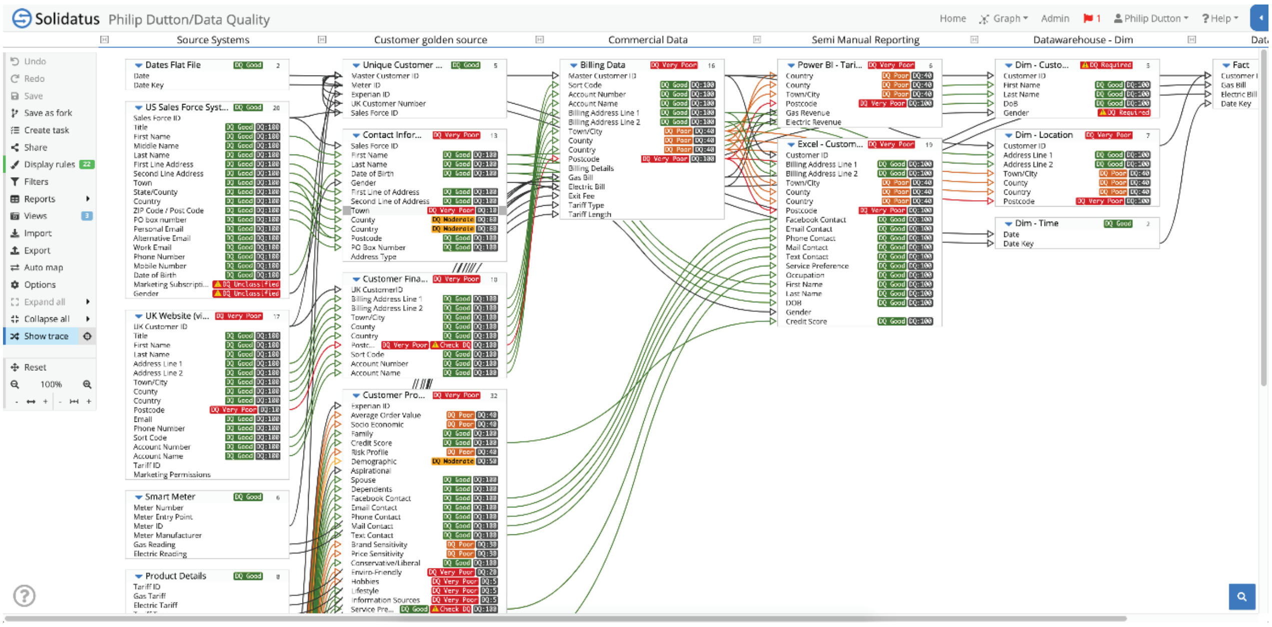
> *Data is as much an asset of our business as our aircraft, our routes and our brand; the quality of that data is everything to us.*
>
> BA

However, we have not defined what we mean by *"quality"*. This is not just the accuracy of the data from a technical perspective but also its completeness, timeliness, validity (the data may fit the format of an email address but that doesn't mean it is valid), consistency (contradictory data will not help in decision making) and availability (if you can't access the data it doesn't have much value).

As mentioned above, data profiling tools are commonly used to discover and assess the quality of data. They provide statistical analysis of the data within each column of a table, thereby enabling the identification of errors (null values, invalid entries and so forth), plus these tools typically also provide monitoring facilities (dashboards) that allow you to see your current state of data quality and how it is evolving over time. Note, however, that data profiling tools are not the only way to do this. Data lineage tools – see later – can also provide this sort of data quality tracking. For example, ***Figure 5*** illustrates data quality tracking from Solidatus.

As far as data quality products are concerned, these typically include a variety of capabilities. However, there is a difference between the sort of capabilities you need when you are dealing with people, for example, and products. And there are also different requirements when dealing with sensor or other IoT-based data. We need to consider each of these separately.

*Figure 5: Tracking data quality across your environment*



## People data

This is the most common use case for data quality tools. Further, data profiling is typically integrated into the solution to enable the discovery and monitoring of data quality, with specific quality capabilities that include:

- Data cleansing, which is the process of rectifying the quality issues, such as missing values, values outside an allowed range or of an incorrect datatype, and so on; all of which should have been identified through profiling.
- Data matching, which is the process of identifying and then merging (semi)-duplicate records to create what is sometimes known as a "*golden record*". Note that matching should be supported across data sources and not just within a single source. The performance of match engines is a significant differentiator when comparing products and it is also an area ripe for the implementation of machine learning to help to reduce false positives and negatives. Graphs can play an interesting role here. There can be value in leaving multiple entities in place and linking them, optionally with a "*certainty*" weight on the relationship. This is useful when the multiple records represent different physical identities (Twitter, Facebook, email, and so forth) or are spread across different lines of business, but which you want to relate to a logical entity.
- Entity matching. This is  complementary to data matching and applies when you need to match across files or systems where no join key is available. It applies to any sort of entity, not just people and, as with data matching, machine learning has a major role to play.
- Data verification (referred to above as validity and consistency). You would verify your data in two ways: either against a set of rules based on business definitions and regular expressions or against reference data sets such as national postal addresses, domain registries, phone companies, watch lists, etc. These types of verification give data stewards the ability to measure their world and there is a clear link to data governance (see later).
- Data enrichment. It is often the case that you will want to add further information about customers, available from third-party sources. This might include company reference data from Dun & Bradstreet, credit information from Experian, geo-spatial data such as latitude and longitude, or other similar information provided by various bodies. For example, in the UK, **www.Gov.UK** registers are available that provide *"structured datasets of government information"*. However, while enrichment has traditionally been regarded as a function of data quality tools, it is increasingly common for organisations to use APIs to access this sort of data, so it is arguable that enrichment functionality within data quality products is no longer required (though it may be nice to have).

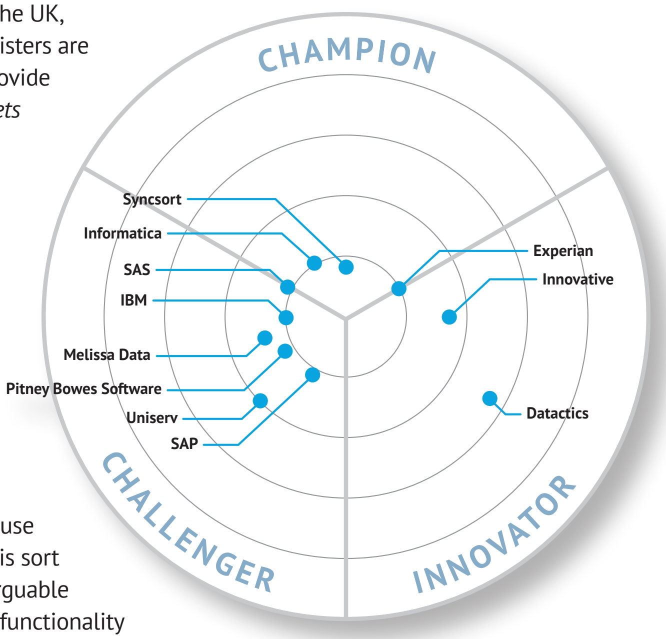


*Figure 6: Bullseye diagram for Data Quality vendors*

- Data steward capabilities. Many vendors also provide specific capabilities to support the activity of data stewards who monitor and improve data quality. This may be in form of a separate module or there may be persona-based capabilities.

Although we discuss product-specific capabilities in the next section, we should point out that most vendors have at least some relevant functionality that supports product quality information. Bloor Research has recently (2019) published its annual Market Update on tools within this sector, with ***Figure 6*** detailing the results. The full report can be found at (***www.bloorresearch.com/research/data-quality-2/***). Note, however, that there are a great many suppliers in this space. Others include Ab Initio, Ataccama, Data Mentors, HCL (Actian), Human Inference, Microsoft, Oracle, Pentaho, Talend and many others. Infoglide, as well as one or two others, specialises in identity resolution.

## Product data

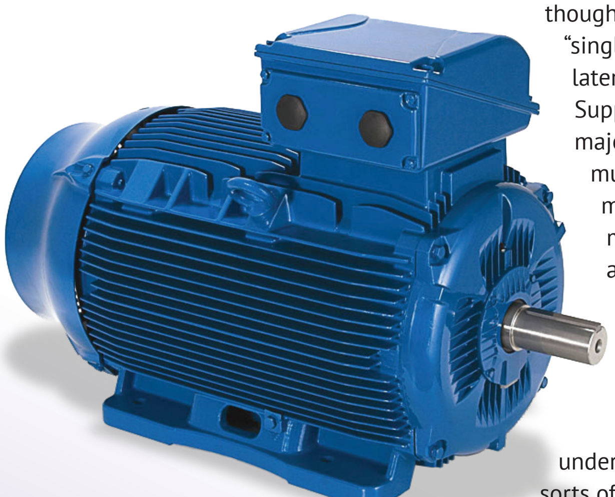


*Figure 7: 10 Horsepower motor*

Issues with product data can recur in other environments dealing with "things" but it is most typical when dealing with products. We can best describe the issues that can arise – essentially matching issues, though arguably relevant to "single product view" (see later) – by way of example. Suppose that you are a major manufacturer with multiple facilities in multiple countries, with multiple descriptions and part numbers for the same product. Then how do you rationalise your environment? And how do you sensibly understand your sales? The sorts of issues that arise with the products that you manufacture can be illustrated through the following real-world example, where you are producing the motor illustrated in *Figure 7*. Depending on the origin of this motor, all of the following might be valid descriptions:

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

| | |
|---|---|
| CLASSIFICATION | Motor |
| ITEM | 26101600 |
| POWER | 10 horsepower |
| VOLTAGE | 115 |
| MOUNTING | Yoke |

*Figure 8: Golden record*

While a human being can recognise that all these descriptions are probably about the same product the challenge is how to get software to recognise this on an automated basis and without, or at least minimal, manual intervention?

Note that we haven't bothered to include product codes against these descriptions. If they were consistent across all manufacturing plants in all countries then identifying the equivalence of these motors would not be an issue. In practice, however, this will not be the case and product codes will be so wildly different as to be useless for matching purposes. The issue then is how to determine that all of these descriptions equate to the table shown in *Figure 8*.

Now, one approach to this problem would be to write a business rule, using a conventional approach to data matching, that said that Motor = MOT = MTR and another one that HP = horsepower = Caballos and a third that says that volts = V = voltios and so on and so forth. However, that would not only be extraordinarily time consuming it would never be completely accurate. For example, what if you also manufacture or supply motor oils? Then V might stand for viscosity instead of volts. You could try writing conditional rules but you couldn't do it on the basis that the description would also include MTR or MOT or motor, because you would expect these in both. What you need is a much broader approach that understands the context within which each product described, with the ability to parse strings of text (thereby implying semntic capabilities), whereby input is validated, vocabulary checked, relationships identified and so on. This will result in the software recognising that this is in fact a motor with the characteristics shown in *Figure 8*. Note that in this example, only the power, voltage and mounting information is necessary to establish a match. Other details can be treated as optional to which you can assign weightings - if appropriate – and use these to calculate relevant match scores for each candidate, which can be presented to the relevant business analyst or data steward if a manual decision has to be taken.

Two further points need to be made. First, note that once the motor has been identified it can be enriched with a UNSPSC (or similar) industry classification. Secondly, note that similar issues arise for distributors (and some retailers) that deal with complex products like this. In these cases, the problem outlined can get even worse, because your company may distribute motors sourced from various manufacturers, in which case you may have further complications with identifying supplier names.

Finally, we should add that while the example above applies to discrete manufacturing, exactly the same issues can arise in process manufacturing and it is also an issue for branding, where you have the same product marketed under different names in different geographies.

### Sensor data

Sensor data is subject to a different set of quality issues. When cleansing sensor data, for example, that is streaming into your environment, then profiling may not be available or appropriate. However, you can still get null values, for instance, if something has gone wrong with a sensor; and you can also get sensor readings that are out of sequence. Ideally, your streaming platform would have appropriate cleansing capabilities to do deal with these issues. In addition, you can also get duplicate readings. For example, mobile phone records are frequently duplicated when data is collected from more than one cell tower. This effectively means that you need matching. However, this should be automated. When matching customer data, it is often necessary to preserve the original records (because your match decision may be mistaken and you will need to roll back) but this is not the case when a single event, such as a call detail record is recorded twice: you simply want a single record.

One further aspect of sensor data quality results from something known as sensor drift. This happens when readings are inaccurate, but in a consistent manner. It is commonplace when sensors are in extreme environmental conditions, such as down an oil well. What happens is that a sensor starts to record temperatures, say, being 5% higher than the really are. Now, this could be because there is a problem that needs to be fixed or it could be because of sensor drift, and you need to know the difference. Drift can usually be compensated for by having multiple sensors, since they are just as likely to drift upwards as downwards, and you can then aggregate the results. This is an issue that companies need to be aware of.

# Data preparation

Data preparation tools provide profiling (but not discovery) and cleansing and joining capabilities, typically within a data lake environment though, if combined with enterprise data catalogues, they should be capable of supporting preparation across IT ecosystems. Indeed, while data preparation is popularly associated with data lakes and analytics, in practice there are many companies using the data preparation capabilities provided by data quality vendors for operational purposes.

The aim with data preparation tools is to bring together data from relevant sources and get that data ready for analysis. This process – finding the data you are interested in and then preparing it – is estimated to amount to 80% of the total time involved in analysing data, whether for purposes such as predictive and prescriptive analytics or to support machine learning or other AI functions.

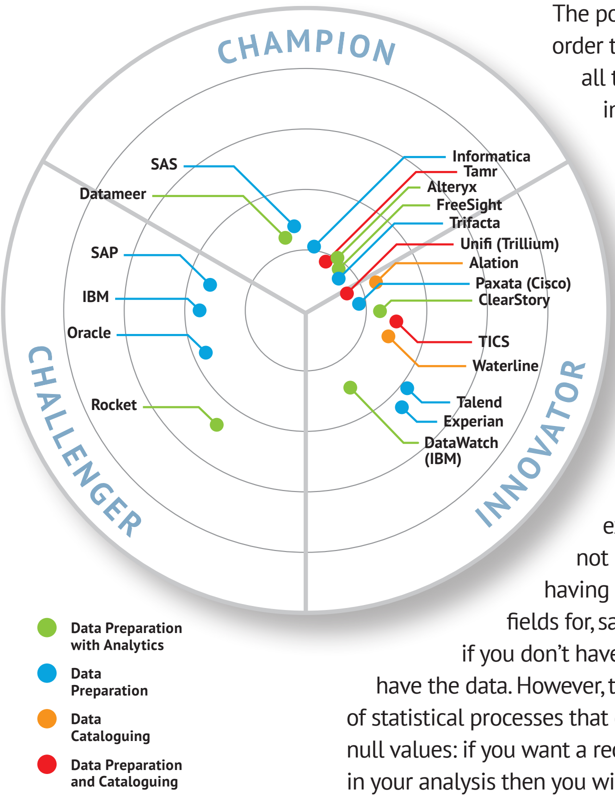
The point here is that in order to perform analyses all the data needs to be in a consistent format (hence the need to join different records) and it needs to conform to data quality principles. These are typically the same as with conventional data quality but there are some differences. For example, you might not have a problem with having null values in some fields for, say, marketing purposes: if you don't have the data, you don't have the data. However, there are a number of statistical processes that cannot run with null values: if you want a record to be included in your analysis then you will have to assign a default value to anything with no value or omit it completely. Another issue that can arise from a data quality perspective is with respect to inferring values. For example, suppose you want to map locations using Tableau: where is Boston? Is it in any of the dozen or so states in America that has a town or city called Boston, or is it in one of the even larger number of other countries that have places called Boston?

> "With (data preparation), we've been able to accelerate the existing processes of understanding our data, surface data quality issues, and wrangle data sets in preparation for further analysis. This... provides us with significant productivity gains and help us assess our data more accurately."
>
> Bell Canada

The second thing to note about data preparation tools is that they are intended for use by either business analysts or data scientists rather than IT people. They therefore need to enable self-service and include collaborative capabilities and features that support recommendations for such things as what to use for a join key when joining different sets of data. Some data governance functions are typically built-in though these are largely invisible to the user. While outside the scope of this paper, data preparation tools should include masking capability (not just encryption). This is because the business analysts and data scientists who use data preparation tools will not normally be permitted to see sensitive data that is subject to GDPR and other such regulations. This is discussed further in Bloor Research's forthcoming paper on "Managing Sensitive Data".

As far as vendors are concerned, Bloor Research published a detailed analysis of this market in 2016 and the landscape diagram from that Market Update is presented in *Figure 9*. The market for data preparation has moved on significantly since then, most notably in that virtually every business intelligence and analytic vendor has now introduced data preparation capabilities within their own tools. In other words, there are a plethora of suppliers and products: too many to seriously evaluate. Various other analyst houses have produced quadrants, waves or other diagrammatic representations of this market, which are readily available on the Internet, but they all omit a significant number of vendors. Moreover, as mentioned previously, data preparation is also used in operational environments and the research referred to invariably ignores the fact that Experian, for example, provides operational data preparation capabilities.

*Figure 9: Bullseye diagram for early providers of data preparation and data cataloguing tools*



# SCV and MDM

Wikipedia defines a single customer view (SCV) as *"an aggregated, consistent and holistic representation of the data held by an organisation about its customers that can be viewed in one place, such as a single page."* It also states that it is also known as propensity modelling. This is incorrect: an SCV may be used for propensity modelling but it is also useful for other purposes such as ensuring compliance with GDPR. Moreover, the use of the word *"customer"* in SCV is a generalisation: it not only applies to clients, criminals, patients and employees but exactly the same requirements – and the technologies to support them – can be required by products (for example consider the motors discussed above) and other non-human assets.

What are the elements required to establish an SCV? Firstly, you need to discover all the data sources in which relevant data is located. Secondly, you need to establish the accuracy of the data in each of those sources and you need to remove duplicate records. In other words, you need the sorts of data assurance capabilities that we have been discussing. In addition, you need to ensure that the data in these sources, having been cleansed and de-duplicated, remains consistent and that data errors do not creep back in. The latter can be accomplished through data profiling tools but maintaining consistency is more of an issue, and its solution will depend on how you create and manage your SCV.
So, in addition to data quality tools (or you can use data preparation tools such as Tamr) you will require at least one of:

- A hub-based MDM implementation with all the data gathered together in one place.
- A registry-based MDM system that supports the propagation of changes to all relevant source systems.
- CRM (customer relationship management) or similar (if you are not dealing with customers). If there is a single, centralised CRM system then this will effectively act like a hub-based MDM environment. However, there may be an issue if you have multiple CRM systems that are not linked together in some way. If you are (you should be) using data profiling to monitor data quality then you should be able to detect when related data becomes out of sync, and it then becomes the task of the data steward to update other data sources appropriately.

> ***The single customer view means that whatever touchpoint that customer comes to us across, whether it be the website, whether it be guest services on site, whether it be the arrival lodge when they arrive, we can engage with them directly, we can make them feel valued.***
>
> **Center Parcs**

The Information Difference has published recent research into the MDM market. As one might expect, it comments that MDM solutions are gradually moving towards cloud and hybrid cloud implementations, but that progress is relatively slow. It also mentions that its research indicates that people costs are four times those of the software costs, when implementing MDM. Interestingly, there has been some consolidation in the MDM market recently with TIBCO acquiring Orchestra Networks, Magnitude buying Agility Multichannel and Informatica acquiring AllSight (for customer 360° insights).

## MDM

Moving on to MDM specifically, The Information Difference reports that the average enterprise has 15 data sources competing with each other to represent that organisation's "master" copy of the data, which is why you need MDM.

For full details of The Information Differences research see ***www.informationdifference.com***. ***Figure 10*** is copied from that report, where the size of each bubble represents the size of the relevant customer bases.

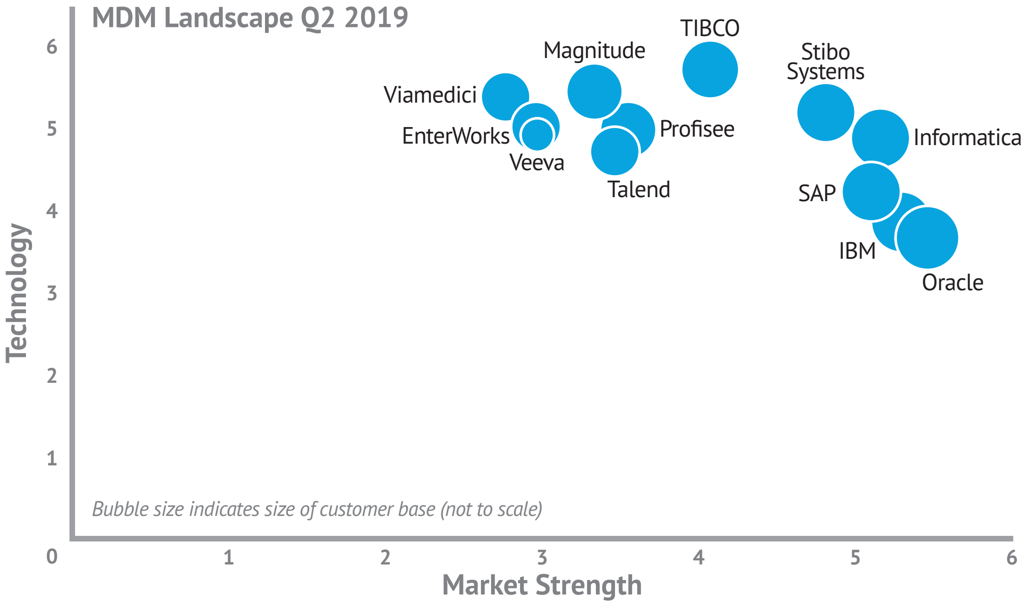
One thing that this report does not discuss is the difference between the various terms bandied about with respect to MDM, most notably PIM (product information management) and PXM (product experience management).

> ***Our primary goal is to drive pipeline for sales and we've been able to tie a 20% lead conversion improvement with MDM. The holistic view of data and the ability to perform predictive analytics will give us almost fortune teller-like abilities.***
>
> **Citrix**

Ben Rund at Riversand has created ***Table 1***, which we reproduce here to help to clarify the differences between these things. As can be seen, it is MDM and neither PIM nor PXM that is concerned with data assurance. However, we should make clear that MDM in this context includes product master data such as the motor discussed above. In particular, it is worth commenting that product, and part, hierarchies are commonplace when it comes to product data. Moreover, bills of materials, for example, are often not strictly hierarchical, which is why a number of vendors in the MDM space are basing their solutions on graph databases.

The Information Difference graphic does not distinguish between MDM, PIM and PXM. For example, InRiver, QAD and ViaMedici (manufacturing) are all primarily PIM vendors. Other vendors in the MDM space not named in ***Figure 10*** (and some not mentioned in the report at all) include Ataccama, Purisma (D&B), Global IDs, iWay, Lansa, Pitney Bowes, Reltio, Riversand, Semarchy and Software AG amongst others.

***Figure 10: DM Landscape provided by The Information Difference***



| *Table 1* | What is Master Data Management (MDM)? | What is Product Information Management (PIM)? | What is Product Experience Management (PIM)? |
|---|---|---|---|
| **Definition** | MDM creates a 360° view across any kind of master data such as parties, places and things and connects the dots across them. | PIM streamlines the product information supply chain from creation to sales across different users. | PXM makes great product information and customer experience smarter and more relevant. |
| **Examples of critical capabilities** | ● Data Governance<br>● Stewardship<br>● Hierarchy Management<br>● Integration<br>● Match and Merge<br>● Analytics<br>● Data Quality | ● Collaboration and Workflows<br>● Enrichment<br>● Catalog Management<br>● Vendor Portal<br>● Print Catalogs<br>● DAM<br>● Translation Management<br>● E-Commerce integrations | ● Product Syndication<br>● Contextual Offerings<br>● Channel Intelligence<br>● Automation by ML/AI<br>● Product content analysis<br>● Personalisation |
| **Examples of business outcomes** | ● Trust worthy data<br>● Comply with regulations<br>● Smarter decision making | ● Faster time to market<br>● Improve operational efficiency<br>● Higher conversions& higher cart value<br>● Less product returns | ● Enables the next best offer<br>● Improves SEO<br>● Good recommendations |

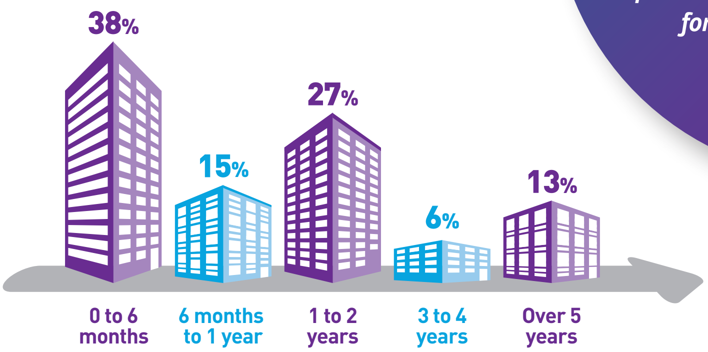
# Data governance

While data cleansing products typically include the ability to define data quality rules, data governance products extend these capabilities in a number of ways. For example, supporting rules about the processes involved in onboarding a new customer or introducing a new product. In this sense, data governance is not just an umbrella for the technologies discussed in this paper, and its companion on sensitive data, but is much broader. These tools also enable the implementation of corporate rules rather than those specifically related to data quality: a top-down rather than a bottom-up approach. They also support the monitoring of compliance though this is more often about sensitive data and security.

> *With analytics-enabled data governance, machine learning algorithms can monitor and improve data quality across an enterprise, self-learning as issues are resolved. Improved data quality increases user trust in data reliability, and therefore increases data utilization for analysis. Machine learning can also play a vital role in compliance efforts, with automatic monitoring for potential non-compliance.*
>
> Infogix

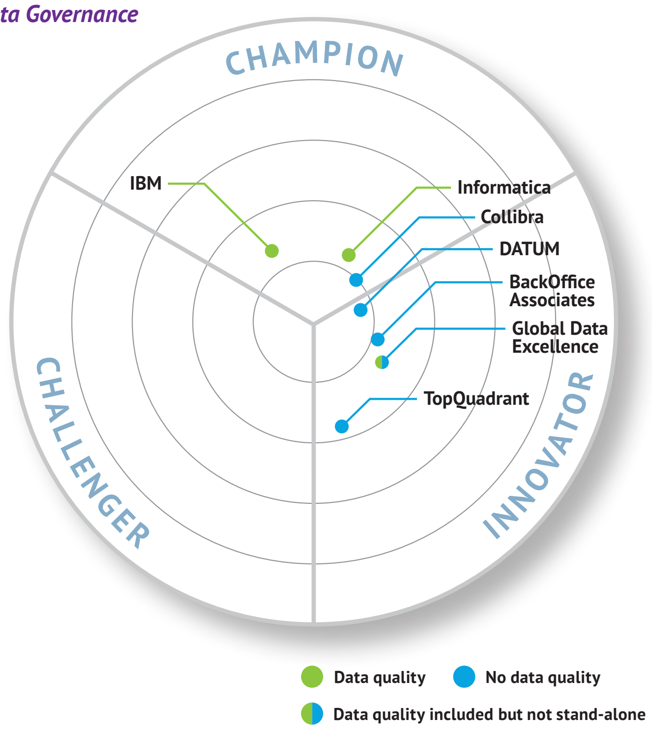
A major driver for the implementation of data governance has been the introduction of GDPR and other regulations concerning personal data but, nevertheless, there remain a significant number of companies that have not, or are only beginning to, implement data governance. ***Figure 11*** represents the results of a survey conducted by CIO Watercooler and it illustrates this point well.

***Figure 11: Plans to implement data governance***



As can be seen in the accompanying diagram (***Figure 12***), which has been extracted from Bloor Research's 2017 report Market Update on Data Governance (see ***www.bloorresearch.com/research/data-governance-2017/***), some data governance vendors also offer data quality and some do not. What we have not shown is that suppliers in this market usually offer data stewardship capabilities that will overlap with data quality products. The same is also true for data catalogues, which are now provided by all, or almost all, data governance vendors. Thus, in addition to policy management – a major differentiator – the ability to integrate across these different technologies is important, as well as things such as providing or integrating with business glossaries.

***Figure 12: Bullseye diagram for Data Governance***



As far as the market is concerned the major change has been the acquisition of DATUM by Infogix. Alex Solutions, an Australian company, has also emerged with a strong technical solution, Global IDs is also a player, while erwin has also entered this space.

## Data Lineage

Data lineage is the ability to track how data moves through your organisation over time: in other words, what is its origin, what happens (is it transformed, aggregated with other data and so on) to it and where does it end up? Apart from formal governance and compliance data lineage is also relevant to understanding mainframe environments, API suites, and interactions within hybrid cloud environments, amongst others. In any case, particularly when the environment is complex. For example, anti-money laundering requires traceability across legislation, processes, data structures and policies. Needless to say, visualisation is key to data lineage. For example, ***Figure 13*** shows data lineage to support Comprehensive Capital Analysis and Review (CCAR) provided by Solidatus. Here the issue is that you need to document and track model development, implementation, use, and validation so that you can prove that your data is *"complete"* and that you can defend the accuracy of your estimates.

> ***"We used... to rapidly visualise and redesign our global overnight batch process. We imported the existing flow... and re-modelled it in just 3 weeks. We delivered what was expected to take 3 months, making savings of at least 70%."***

Technically speaking, providing an understanding of data lineage is a subsidiary requirement for data governance and most, if not all, data governance tools provide lineage capabilities. That said, the last couple of years has seen the emergence of several pure-play vendors in this market, which suggests that at least some of the data governance providers are not doing a good enough job. This is illustrated by the fact that Solidatus actually partners with Collibra. These new vendors include Manta, Solidatus and Conto (with Octopai), though Manta and Octopai are primarily visualisation tools whereas with Solidatus you can actually manipulate the data. Indeed, Solidatus is a partner of Manta's (as is Stardog: see the knowledge graph section). You can also use it for associated functions such as data quality tracking (see ***Figure 5***) or for tracking masked data in a similar way to Informatica's Secure@Source. As far as we know, no industry analysts – including Bloor Research – has done any comparative analysis on products in this space.

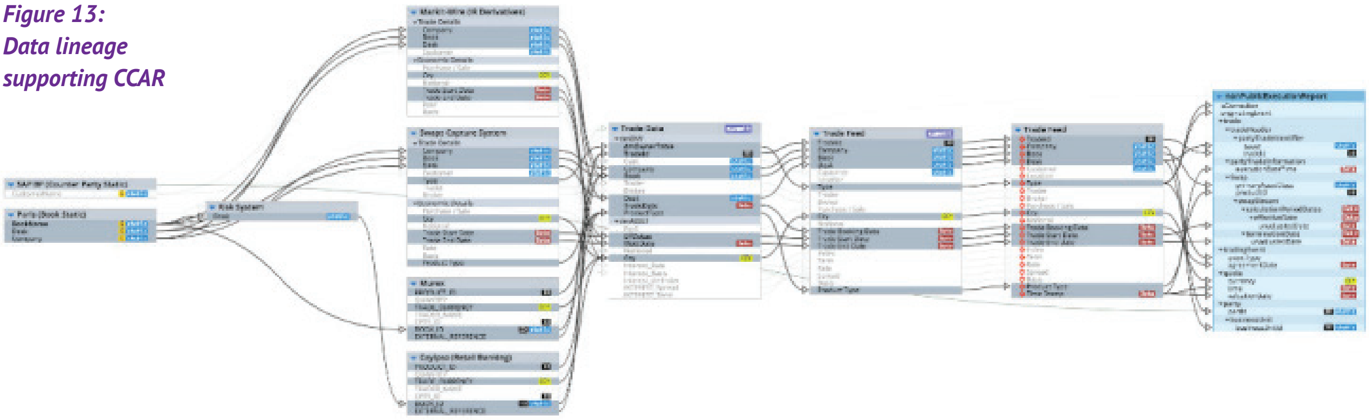


***Figure 13: Data lineage supporting CCAR***

## Spreadsheets

There are NO complete and comprehensive data quality or data governance tools, though there is one vendor (Freesight) that provides spreadsheet management along with data preparation. This is because none of them cater to the quality of data in spreadsheets or the governance thereof. Given the prevalence of spreadsheets and the costs of calculation and other errors (see ***http://www.eusprig.org/horror-stories.htm*** for some examples) this means that any company seriously interested in data assurance will need a separate solution for managing spreadsheets.

Just as with data in databases, there are three phases to spreadsheet management: discovery, quality and governance. The first thing that you need to do is to discover what spreadsheets exist and where they reside on the network. However, it is not enough to treat this as a one-off exercise because new spreadsheets are created all the time. This means that you need discover new spreadsheets on an iterative basis. Secondly, while finding spreadsheets is one thing you will also need to collect as much information (metadata) as possible about these spreadsheets. This is for a variety of purposes: to understand who owns this spreadsheet, who has access rights to it and what those rights are; to support versioning; and to identify links (broken links are a major cause of spreadsheet issues). Once you know what spreadsheets you have, the next issue is to determine which are the most critical. There are two basic elements in this. Firstly, how complex is the spreadsheet? How likely is the spreadsheet to be broken or have errors in it? And, secondly, there is materiality: does the information in the cells of the spreadsheet suggest that this is an important spreadsheet? This is important for assessing the risk that any particular spreadsheet poses. In this context it is worth commenting that a number of spreadsheet management offerings integrate with GRC (governance, risk and compliance) products such as IBM OpenPages, RSA Archer and MetricStream.

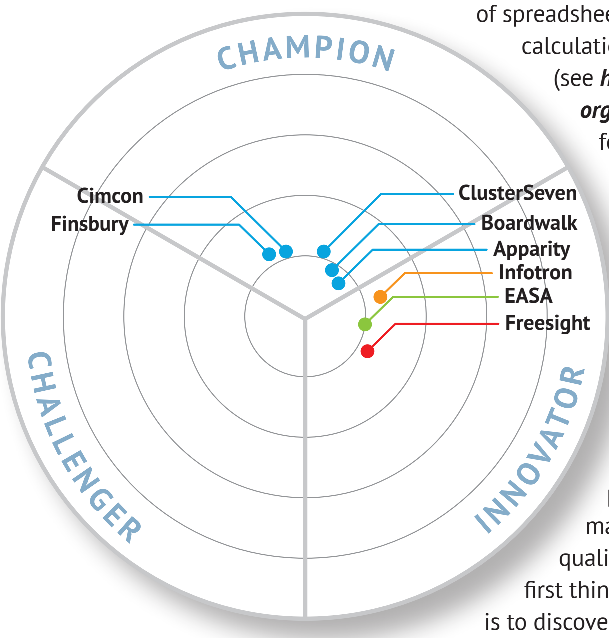


*Figure 14: Bullseye for Spreadsheet Governance*

As far as quality is concerned, these are better described as errors. There are tools available that are solely targeted at errors. Common features used for this purpose, include spreadsheet comparisons, either between two versions of the same spreadsheet or different spreadsheets; formula mapping, which is the ability to see how formulae have been copied (or not) across cells; precedent and dependent mapping, to see relationships and references across spreadsheets; detection of formula and other errors such as text in a data field, a sum that is adding up non-numeric fields or range checking; facilities to understand formulae more easily; data lineage, *"where did this number come from?"*; circular reference detection; and various fraudulent uses of spreadsheets.

As far as governance is concerned this means putting in place a management framework for the provisioning, development and versioning of spreadsheets, plus the governance of what users are allowed to do with spreadsheets. That is, can they change this value, amend that formula, and so on? Historically, the process of putting discovered spreadsheets under management control meant moving them to and hosting them in a central repository or database. However, more modern tools do not mandate this process and allow control without the disruption of moving spreadsheets.

The foregoing is by no means a comprehensive discussion of the requirements of spreadsheet management and for more information refer to the various reports that Bloor Research has published on this topic. These include our most recent report: see ***www.bloorresearch.com/research/spreadsheet-management-and-governance/*** published in 2018, from which ***Figure 14*** has been extracted. Vendors not shown here include Incisive Software, Hub85 and DataRails.

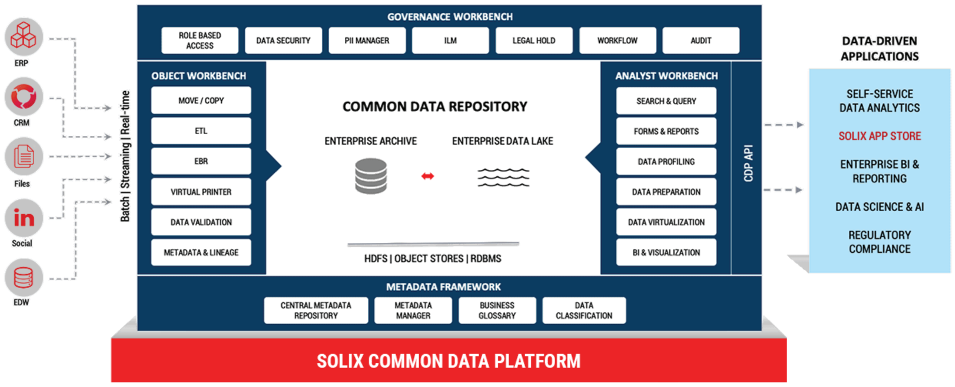
# Information Management Suites

Companies in this space, such as Informatica, IBM, Solix, Ab Initio, Global IDs, Talend and others, often refer to their offerings as data management platforms. However, the term "data management platform" is also used, according to Wikipedia, to refer to a *"technology platform used for collecting and managing data, mainly for digital marketing purposes, (providing the ability) to generate audience segments, which are used to target specific users in online advertising campaigns,"* so we will refer to these as information management suites.

As a general rule, an information management suite will encompass all of the subject areas previously discussed, with the possible exception of data preparation and the probable exception of spreadsheet governance. A typical example is shown in ***Figure 12***. On the other hand, they may well include additional capabilities such as test data management, sensitive data discovery, search and data integration functions. They may or may not offer best-of-breed capabilities in any particular area but their real strength lies in the breadth of capability that they offer, with a common user experience, collaborative capabilities that link across subject areas, and so on. Solix, for example, is quite clear that it's Common Data Platform is targeted at companies that want a broad solution rather than any particular point solution. This is not necessarily true for other vendors of information management suites, where there may be individual areas in which the supplier excels, but it is generally true.

As far as we know, there have been no comparative studies made of any providers of information management suites.

***Figure 12:***



# Assured data to support machine learning

Machine learning can be deployed in one of two ways: either pre-trained or intended to *"learn on the job",* which we might describe as self-trained. If you build a machine learning model for your business you will normally pre-train it. Vendor provided products, on the other hand will almost always include some degree of self-learning. Matching engines, for example, might be delivered with significant amounts of pre-training, because matching customer names and addresses is a common problem and there is lots of available training data. On the other hand, predicting customer churn will very much depend on your products, your service and your company, so most if not all training will need to occur on the job, unless you feed that model with lots of historic data first.

Training data, applicable when there is any degree of pre-training, needs to be of high quality for reasons discussed previously. It also needs to be free from bias, because biased machine learning models will either be sub-optimal from the business' perspective or unfair to customers or clients, or both.

There are more than 180 identified biases. We don't have space here to review them all but they range from racial, gender and ideological biases through to more technical issues such as sample bias, which is where the training data is based on a sample of records that does not accurately reflect the environment in which the model will be deployed.

In addition to biases in training data, models in deployment can also become biased over time. In effect, they can become self-reinforcing. A model might suggest a course of action, you take that action and it is successful, so the model becomes even more strongly convinced that this is the best approach and so on and so forth. You might call this a spiral bias.

There are very few vendors that are currently offering bias detection and remediation and there certainly have not been any analyst reports delving into this area in detail. Of the suppliers we are aware of, both Google and Databricks have launched appropriate tooling for bias in training data with respect to their own environments (TensorFlow and Spark MLlib respectively). Neither of these have capabilities supporting bias discovery and remediation in production. As far as we know IBM is the only company offering these functions for both training and in deployment through its Watson OpenScale product. This supports a variety of environments (including TensorFlow and Spark MLlib but also others) and also includes a number of other capabilities such as traceability and explainability.

# Conclusion

In practice there are two types of company interested in data assurance. The first consists of organisations that are trying to solve very specific problems that will potentially make a huge difference to their business. To do this, they may only need to dedupe an SCV in Salesforce or keep a debt collection system up to date with contact details, or on-board dirty data from brokers, suppliers, or resellers more efficiently. In these cases, most of what they need they already have: it's just the final one or two parts of the jigsaw that they need and therefore a best-of-breed solution, provided it fits into their jigsaw puzzle, may be the most sensible solution. However, that proviso in the last sentence is important. If that best-of-breed solution doesn't easily integrate with other pieces of the puzzle then it may not be the best solution.

More generally, there are a lot of things to consider with respect to data assurance. If you have a major project that spans several of these areas then it may make sense to consider a platform-based solution rather than selecting several individual products from the plethora of point tools that are available. For example, if you are going to implement MDM then it will often make sense to use data quality and profiling tools from the same vendor so that you can avoid any integration issues. We therefore place a premium on suppliers that offer to resolve multiple issues within a single tool or toolset. Leaving aside spreadsheet governance, specialised data lineage requirements and issues with training data, this means looking at vendor solutions that span more than one of the areas considered in this paper.

## About the author

**PHILIP HOWARD**
**Research Director / Information Management**

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 his own company 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, focused on Information Management.

Information management includes anything that refers to the management, movement, governance and storage of data, as well as access to and analysis of that data. It involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration, data quality, master data management, data governance, data migration, metadata management, and data preparation and analytics.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to ***IT-Director.com*** and ***IT-Analysis.com*** and was previously 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 written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

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

## Copyright and disclaimer

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.

## Bloor overview

Established 30 years ago, Bloor has become one of Europe's leading independent IT research, analysis and consultancy firms.

Bloor is widely respected for providing actionable strategic insight through its innovative independent technology research, advisory and consulting services. Bloor assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

Underpinning Bloor's whole ideology is that digital and business transformation isn't a serial 'one and done'. Being a Mutable Business™ in a state of permanent reinvention; evolving business models, people and resources with technology is the key to securing long term survival.

## Bloor Consulting

Bloor Consulting is here to guide your organisation on its journey to be a Mutable Business™ Our senior-level consultants leverage in-house research, best practice data and vast industry experience to meet and exceed our client's expectations.

Bloor offers a range of packaged or custom consulting services. Short, or long-term, consulting engagements. Depending on where you are in your journey, these are just some of the areas we can advise your organisation.

For a full list see ***bloorresearch.com/consulting-service/***

### Disruption & Change

Bloor can help you assess and evaluate opportunities and threats to your business (and marketplace) from the emergence of new organisations and technologies. If you are looking to the future and considering how best to innovate, Bloor can help fully evaluate the technologies and opportunities arising, and create an adoption roadmap, risk assessment and change plan to support your innovation.

### Strategy

Bloor's unique blend of technology competency, delivery, business, and commercial experience enables us to support your strategic planning, ensuring that realistic expectations are set, and risks are appropriately managed. As well as ensuring the impact of change is fully reflected in your business plans.

### Business Creation & review

Bloor can assist you to create a practical vision of how developing technologies and emerging working practices will create opportunities for your business. We can introduce you to organisations developing the innovation as well as transferring our knowledge and experience to your business.

### Cybersecurity

According to the latest World Economic Forum poll, cyber-attacks are seen in 2019 as the most pressing risk for CEOs in Europe and North America, including six of the ten largest economies in the world. Cybersecurity risk is a whole board agenda item and our experts can advise you on the latest and best ways to protect your organisation

### Leadership

We can undertake a review of your Mutable Business™ journey so far and help your team to understand the impact and full consequences of the change needed. Our team can work with you to create a vision of the future resulting from the change.

### Advisory and Review

If your business has a high dependence on external 'partners' to deliver major service contracts, are you concerned in ensuring you can control their cost and change processes? We can provide advisory and review services and support, to ensure you can keep their costs, and your agenda, on track.

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

tel: **+44 (0)1494 291 992**
web: **www.Bloorresearch.com**
email: **info@Bloorresearch.com**