

Data Cleansing

Introduction

This is the third of four Market Updates on data discovery, data profiling, data cleansing, and data quality platforms respectively. Within “data cleansing” we include matching, standardisation, enrichment and other facilities normally associated with data cleansing. We do not include data profiling, facilities to support data migration, data governance capabilities and similar facilities, which will be included along with data cleansing in the fourth of these Market Updates. That said, we do discuss identity resolution as distinct from conventional name and address matching in this paper, and we also consider the needs of product cleansing and matching as opposed to name and address matching.

Key market issues

Of course there are generic issues that distinguish between data cleansing products, such as performance, scalability, ease of use, and so on. There are also factors such as Unicode and multi-lingual support—which are not the same thing—the former applies to the data being cleansed, the latter to the user interface. However, leaving those aside the key issues that distinguish products in this market are mostly dependent on the markets that they address.

The name and address market

This is the standard market that all but one vendor in this report is focused on. Suppliers in this sector also typically claim that they address the markets for identity resolution and product cleansing. With the exception of those that have specialised facilities in these areas, we would go no further than saying that these vendors can do an adequate rather than a good job though some, of course, do better (or perhaps we should say worse) than others. Insofar as name and address cleansing is concerned there are obvious factors such as country support for postal address files but, more specifically, major issues are as follows:

Does the product support geospatial enrichment? This goes beyond just latitude and longitude and includes spatial analysis

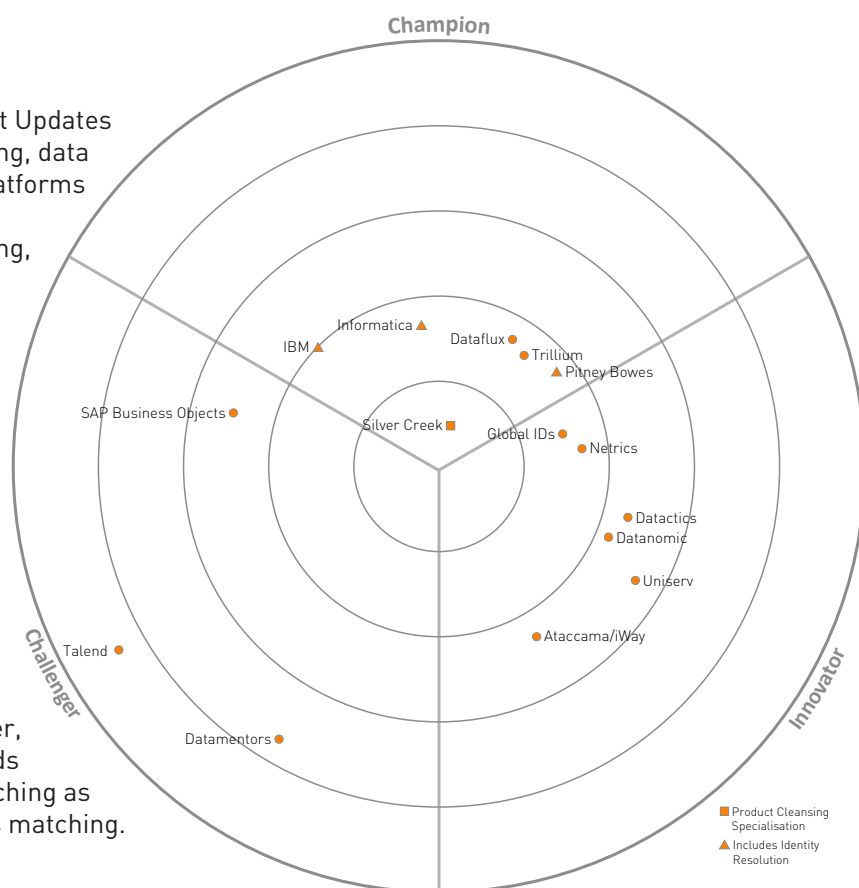


Figure 1: The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator segment if their innovation rating is over 2.5 and Challenger if it is less than 2.5. The exact position in each segment is calculated based on their combined innovation and overall score.

to determine point-in-polygon, closest site to features with boundary data for things such as flood plains, telco public service answering points, earthquake zones, hurricane tracks and so on. There is an increasing requirement for this sort of data in a number of vertical sectors. Note the difference between geo-coding and geospatial capabilities.

Business glossary. A populated business glossary helps both business and technical staff to collaborate on data quality, using commonly understood terminology. However, while a number of suppliers claim the ability to support such glossaries (that is, “you could build your own if you wanted to and we could host it”) few actually deliver them in any sort of pre-built fashion.

Are there vertical industry-specific features built-in? This could encompass both of the previous points but also extends to cater for industry-specific terminology. This is growing in popularity and more than half the vendors in this Update support, or plan to support, such capability.

How good is the matching engine anyway? Most products in the market are getting long in the tooth and have engines that are a number of years (in some cases more than a decade) old. So, to make them better, suppliers have increasingly focused on adding more and more algorithms that will make the old engine better. In our view, it is not more algorithms that are needed but new engines.

Self-learning. You want the process of data cleansing and matching to get easier as time goes by, so that it becomes less and less manual. The traditional way of supporting this is by creating additional rules that the software will process—but rules need maintenance and they aren't foolproof. A more productive approach is to use some form of machine learning that will automate the process of learning.

Preventative functions. There is a trend (and about time too) towards preventing data quality problems rather than detecting and resolving them. Of course, all sorts of data cleansing products need to integrate with the outside world but this is a particularly acute problem when preventing problems, because you need tight integration with the relevant application suite from SAP, Oracle, Sage or whoever. You also need appropriate management of the processes involved in preventative data quality.

The product market

Name and address data is typically in a defined format with first name, last name, address lines, and so on. It is therefore relatively easy to compare different names and addresses: just compare the elements of the name, the elements of the address and any other attributes. This is how the engines of most leading products were built. However, if the data is not structured, as is often the case with product data (typically, a long string of data with no recognisable sequence) and sometimes with name and addresses then this sort of field-by-field matching won't work very well.

There are two possible ways to overcome this issue. The first is to compare every element within the fields being matched with every other element but this requires a next generation matching engine that has appropriate performance characteristics. Moreover, this approach won't resolve the sort of ambiguities that can occur when, for example, the same abbreviation can be used for different attributes. The second is to standardise the data before you match it. In other words you extract the fields to use for matching purposes. This is the approach taken by semantically-based products, which can much more easily handle ambiguities. There

are also conventional products with attribute extraction algorithms that effectively do the same sort of pre-standardisation though they will still be rules-based when it comes to handling ambiguities.

The identity resolution market

Most vendors who don't have specific identity resolution capabilities don't appreciate (or pretend to not appreciate) that their products are not ideally designed to match the names of things (typically people and companies). The problem is similar to that for products. For example, in many countries people put the surname before the forename. Worse, migrants to foreign countries will often anglicise their name or 'germanify' it, or whatever, as appropriate. Then again, in some cultures there are parts of the name, often titles or semi-titles, which may or may not be dropped depending on circumstances. In criminal circles you also have the issue of aliases. So, you potentially have many versions of a name, all of which are valid, and you need to be able to recognise that they all related to the same person. Similar considerations can apply to company names: there are several hundred valid representations of 3M, for example.

To cater for all of these issues you really need a special-purpose identity resolution product and there are a number of such on the market. For relatively simple identity matching you could use a next generation engine, as described in the previous section, but this will not have the ability to cater for such things as aliases and will not have the in-built cultural understanding (forename first or last, the use of patronymics and matronymics and so on) you would expect.

Vendor landscape

As is usual with reports of this nature our research has coincided with releases and re-organisations from various companies in an untimely manner, meaning that a number of vendors have declined to be involved in this Update. The most notable of these are Microsoft and Oracle, however the loss in the latter case is not great as the company resells software from Silver Creek, Trillium and Identity Systems (now part of Informatica). Some of these omissions are discussed in this paper even though they are not represented on our Landscape diagram.

The nature of market dynamics means that most users will opt for a data cleansing solution from one of the leading and most well-known providers. These include SAP Business Objects, IBM, Dataflux, Trillium, Informatica, Microsoft, Oracle and, arguably, should also include Pitney Bowes.

Of these vendors, all of them take the same fundamental approach to data cleansing, using a traditional matching engine. However, Microsoft acquired Zoomix last year, which employs a semantic approach to data matching. The company declined to take part in this report because it is too early for it to disclose what it is doing with Zoomix but it could well be that it is planning to break ranks with its most well-known competitors.

The problem with the traditional approach to matching engines can be demonstrated just by looking at the number of vendors who are trying to do something about it. For example, Uniserv uses a traditional matching engine but has added knowledge bases and expert systems capability to try to get over its problems. Ataccama (whose software is resold by iWay and embedded within that company's real-time EIM product) has also implemented some innovative technology for the same reason. More radically, Netrics has introduced a matching engine that supports algorithms based on bipartite graphs (which provide all-versus-all comparisons) and this is also sold by Global IDs, which embeds Netrics' software. Taking a totally different approach, Silver Creek, which specialises in product cleansing and matching, is semantically based. Pervasive is also worth special mention since its DataRush engine (which underpins DataMatcher) has been designed to support parallelism across multi-core architectures.

Clavis is a new start-up brought to the market by the people who founded Similarity (now part of Informatica). It is interesting for two reasons: firstly, because it is SaaS-based and secondly because its emphasis is on preventative data quality and, particularly, the processes involved in that. As far as the latter point is concerned this is also an area of strength for Trillium, amongst others. In terms of SaaS capabilities this is an area that everybody is exploring and many are offering. Most interesting is Pitney Bowes: because its product is highly modular it allows you to license some elements directly and use SaaS for others, in a very granular way.

Summary and conclusions

What we would really like is Netrics' matching engine with Silver Creek's semantics and self-learning, Pitney Bowes' geospatial capabilities (though this is currently only available for some countries) and SaaS-based flexibility, IBM's Business Glossary, Trillium's process management, Datanomic's industry focus and ease of use, and some smattering of bits and pieces from a variety of other suppliers. Unfortunately, you cannot have that.

For name and address cleansing we particularly like Netrics and Global IDs, and we are also favourably impressed with Ataccama. Of course, all the mainstream vendors will do a decent job in this area. For product data quality, Silver Creek is head and shoulders above the rest, especially thanks to its remediation capabilities, with Dataflux (thanks to its attribute extraction algorithms) and Datactics also worth consideration. If you need both name and address and product cleaning/matching then Dataflux, along with Trillium, is a good bet, as is Global IDs, or you could take the approach of using Silver Creek along with one of its partners.

To get a head start on your data cleansing in particular vertical sectors there are a number of relevant vendors: Silver Creek in a number of areas for product data and various suppliers for financial services. Less common capabilities include Informatica for CPG and utilities, Datanomic in telecommunications and Datactics for the automotive, heavy equipment manufacturing and electronics sectors.

Finally, insofar as identity resolution is concerned: IBM, Informatica and Pitney Bowes are the only three companies offering this, with IBM's offering being more extensive (integrating with entity analytics) but the latter two arguably more tightly integrated into their overall offering. If you want identity resolution but don't like any of these vendors then the only other specialised vendor in this area that we are aware of is Infoglide.

*Philip Howard
Data Management
May 2009*

The logo for Bloor Research, featuring a stylized 'B' icon followed by the word 'Bloor' in a bold, sans-serif font.

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

Tel: +44 (0)207 043 9750
Fax: +44 (0)207 043 9748

Web: www.BloorResearch.com
email: info@BloorResearch.com