

Data Profiling

Underpinning Quality

Data Management

Overview

The current economic climate forces all businesses to compete more effectively. The ever increasing pressure to reduce margins in recent years has forced organisations to look inwardly at solutions based on cutting costs and maximising customer profitability. These drivers have resulted in a reliance on IT systems supporting the likes of: Enterprise Resource Planning, Customer Relationship Management, Supply Chain Management, Manufacturing, Stock Control, Logistics, and Business Intelligence Solutions, to name but a few.

However, all of these solutions will only deliver value if the data they depend on is accurate, complete and consistent. This information feeds and supports everything from the simplest business process to the highest level strategic decision. Hence, the quality of an organisation's data assets has a very real impact on the bottom line.

A recent survey commissioned by the Data Warehousing Institute estimated that poor data quality costs US businesses alone \$600bn a year. In our experience, the greatest losses are derived from:

- Wasted investment into new systems that do not deliver a return or concrete benefits
- Excessive overheads due to inflexible processes reliant on data that is not "fit for purpose"
- Lost revenue caused by poor quality customer data
- Flawed strategic business decisions based on inaccurate or incomplete data

The need to analyse data has been at the foundation of every effective data management strategy since the dawn of modern information systems. Whether auditing your data assets to assess quality, ensure regulatory compliance, gain a better understanding of your information needs, or to embark upon new systems implementations, Data Profiling can deliver a deeper and broader insight in a fraction of the time required by traditional approaches to data analysis.

Data Analysis vs. Data Profiling

Traditional data analysis techniques are simply unable to cope with the scale of today's data management challenges, which inevitably stretch the analysis resources available to the point where radical reductions of scope become a necessity. These factors contribute to the near certainty of missing completely, or managing incorrectly, critical data quality problems.

More often than not, data quality issues remain hidden until they surface at a time when they have the most negative impact on business performance. In a project context, this frequently occurs during loading, final testing or go-live situations where the costs of retrospectively fixing the issues have risen exponentially (sometimes to the point where all the business value of the project is negated). It is impossible to guess in advance where these problems may lie. Only comprehensive and regular data audits can identify all possible dangers before they adversely affect the business or become unmanageable.

"Unfortunately, conventional methods for analyzing real data take a great deal of time, involve only small samples of the data and fail to deliver a complete understanding of the source data. Manual or semi-automated processing techniques cannot possibly compare the thousands of attributes and millions of values necessary to uncover the relationships. The answer is a new category of software called data profiling [...] which offers a fast, accurate and automated way to understand your data. It enables a small, focused team of technical and business users to quickly perform the highly complex tasks necessary to achieve a thorough understanding of source [or target] data. This level of understanding cannot be achieved through conventional approaches." - Craig Olson, Data Management Review, March 2000

Data profiling technology vastly improves the scope and depth of data analysis in three key ways:

- Automation of traditional analysis techniques – it is not uncommon to see analysis time cut by 90% while still providing a better understanding.
- Ability to apply a brute force approach to analysing data – analysts are no longer limited to working just with sample data. Terabytes of data can be profiled effectively and completely. Sometimes the smallest anomalies can have the greatest impact.
- Assessment of rules that govern data which cannot easily be discerned via manual coding and inspection – eg. pattern generation, dependency testing, join analysis.

Data Profiling is the crucial first step that should be undertaken at the start of any data-driven initiative, whether that be a new data warehouse, system migration, data integration project, or the implementation of a corporate data quality strategy. Without the level of insight Data Profiling can provide, the risk of hitting data quality issues remain unacceptably high for modern businesses.

The Problems Data Profiling Addresses

The old adage, “garbage in, garbage out” applies more today than ever before as data-centric systems support every aspect of running a business. If data is of a poor quality, or managed in structures that cannot be integrated to meet the needs of the enterprise, business processes will inevitably suffer, as will decision-making and profitability.

Data quality issues are endemic in most organisations – duplication, incompleteness, inconsistency, and countless other potential anomalies can all play a part in undermining operational efficiency. Furthermore, any gaps in one’s understanding of the business rules that govern that data can result in new processes and systems being rolled out that fail to meet the ever-changing needs of the organisation, thus compounding the existing problems, compromising trust in the data and increasing risk.

Without a thorough understanding of data quality issues, it is almost inevitable that development costs will spiral and projects will overrun, or even fail outright. A clear, up-front picture of all the potential issues is essential to plan projects effectively. Data cleansing and transformation requirements need to be understood before timescales and costs are finalised, not after. Traditional approaches to data analysis simply cannot answer all the questions that need to be asked. Especially when it is not clear what those questions should be in the first place.

It is not always easy to quantify the exact costs of having projects and business processes undermined by data quality issues, but it is accepted that the impact can be huge on profitability across an enterprise. Most disturbing of all is that many organisations have few metrics in place to assess the extent to which poor data quality affects the bottom line, whether that be from lost revenue, excessive overheads, or unsound business decisions based on misleading business intelligence. Ultimately, the success or failure of a business can depend on the quality of its data assets.

Although often hidden within departmental budgets, what is easier to put a value to is the ongoing cost of “fire-fighting” systems exhibiting data quality issues. These costs break down principally to the associated requirements of:

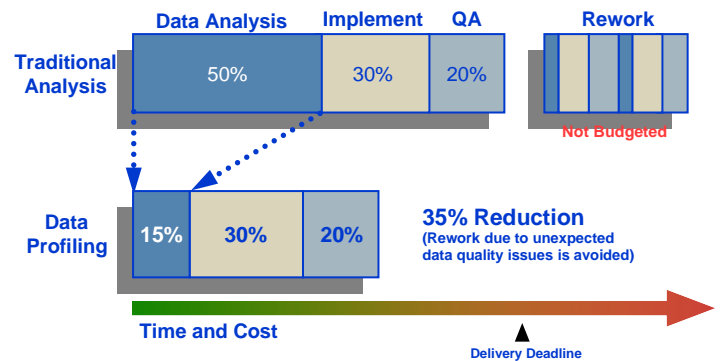
- Identifying data quality issues through thorough data analysis
- Managing the priorities, statuses, and allocations of remedial activities and resources
- Implementing, validating and sustaining data quality improvements

Even taken on its own, being able to reduce or eliminate these overheads almost always justifies the adoption of Data Profiling technology, but the consequent benefits across the enterprise are often many times greater.

Implementing a Data Profiling led approach as a first step in all data-driven projects will radically improve the performance of activities reliant on data management and significantly reduce risks, costs and timescales.

When used, in tandem, to support a strategic data quality initiative, Data Profiling can also identify gaps early to ensure that high performance is sustained across the organisation.

Typical Project Based Scenario

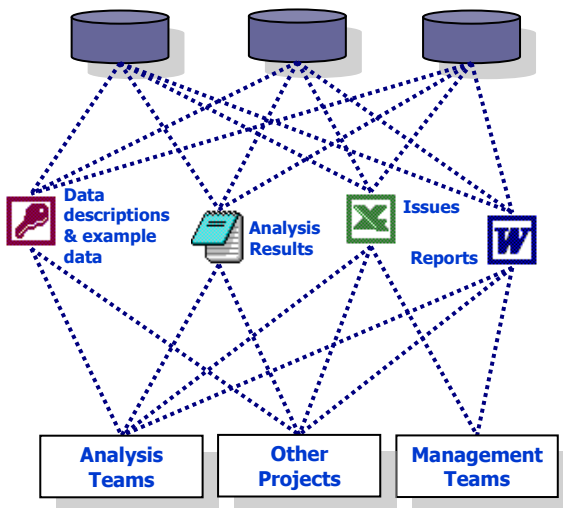


How does Data Profiling promote better Data Quality?

Delivering better data quality relies first and foremost on understanding the data you manage and the rules that govern it. Without this knowledge, no effective data management plan can be formulated. Profiling data provides both the framework and roadmap to improved data quality, smoother running systems, more efficient business processes, and ultimately, the performance of the enterprise itself.

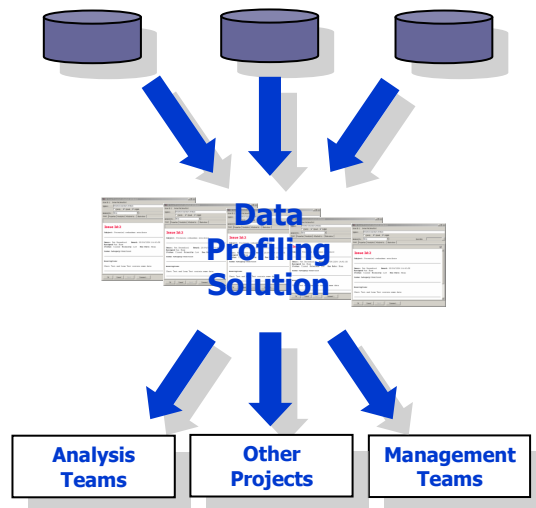
Compared to manual analysis techniques, Data Profiling technology can significantly improve the organisation’s ability to meet the challenge of managing data quality.

Traditional Analysis Approach



- No centralisation of information
- Tools applied on ‘ad hoc’ basis
- High data management overheads
- Increased effort and greater risk

Data Profiling Led Approach



- Single point of reference
- Designed for analysis activities
- Reduced management overheads
- Increased confidence and lower risk

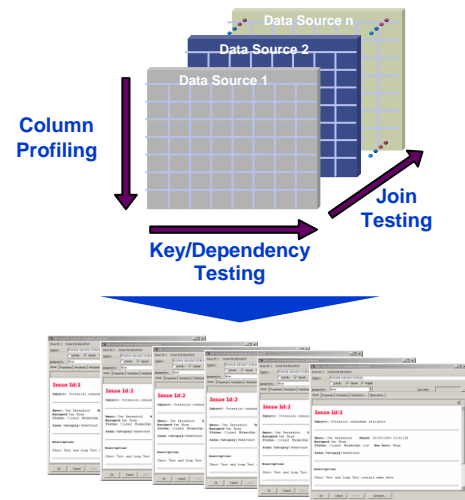
Core Features of Data Profiling Technology

A complete Data Profiling solution delivers “3-dimensional” analysis rather than being limited largely to what is achievable through manual techniques, namely just Column Profiling (see below).

Ideally, the solution should also support a collaborative approach that centralises analytical effort, metadata management for structures and rules, issue tracking workflow, process management, documentation, and reporting.

Automated
Data Analysis

Issue
Management



Metadata Management

Managing metadata (i.e. the library of definitions describing the countless data structures and rules) is a non-trivial task. More often than not, this crucial information is both not accessible to those who need it, and out of date. Typically, these definitions are captured only when systems are implemented and stored in text documents that have no physical links to the systems themselves. Dynamically changing business needs lead to constant evolution, not only in the way existing structures are used and interpreted, but also in the behaviour of the data itself. New structures are added and some existing structures become redundant. Similar or related information is often stored in several parts of the organisation. Maintaining accurate cross-references is pivotal to ensuring synchronisation and consistency.

Attribute	Data type	Dominant pattern	Percentage null
Customer ID	SMALLINT	9(3)	0
Contact First Name	VARCHAR(16)	Cc(4)	0
Contact Last Name	VARCHAR(14)	Cc(6)	0
Contact Title	VARCHAR(4)	Cc	0
Address1	VARCHAR(40)	9(4)WCC(4)WCC(5)	0
Address2	VARCHAR(19)		0
City	VARCHAR(16)	Cc(5)	0
Region	VARCHAR(26)	C(2)	0
Country	VARCHAR(22)	C(3)	0
Postal Code	VARCHAR(9)	9(5)	0
Phone	VARCHAR(16)	9(3)-9(3)-9(4)	0
Fax	VARCHAR(16)	9(3)-9(3)-9(4)	0
On hold	VARCHAR(1)		0
Number of orders	INTEGER	9	0
Customer type	SMALLINT	9	0
Notes	VARCHAR(49)		0

16 items 101 ms Completed

Ok Cancel Apply

The advantage of Data Profiling technology is that it does not rely on existing system documentation or pools of expertise hidden in the business. Instead, insight into the structure and rules governing data assets is derived from the data directly, thus avoiding incorrect assumptions and overlooked issues.

Comprehensive Data Profiling solutions will provide a framework for capturing and maintaining an accurate Data Dictionary, driven by the data resources themselves, across all source platforms.

Ideally, they should additionally provide the option of managing snapshots of data sources (along with analysis results) offline from your operational systems so that data quality histories can be reviewed over time to identify trends and spot unexpected anomalies as soon as they develop.

Column Profiling

Column profiling provides the first cut in understanding data. This is the area where traditional manual data analysis techniques focus most of their attention, looking at each data attribute (column) in turn to evaluate basic features such as dominant data type, percentage population, uniqueness, value ranges, and field lengths.

However, unlike manual analysis, Data Profiling provides several additional capabilities that cannot easily be achieved using traditional, hand-coded approaches.

- The ability to process the entire source rather than a subset of rows or columns limited by how much time is available to the team.
- All data features across the source can be summarised and reviewed rather than just focusing on the issues that are already known.
- Generation and analysis of patterns in data. For instance, many items such as postal codes and customer identifiers will conform to a defined set of standard alphanumeric structures. Pattern analysis enables immediate identification of non-standard values.
- Most Data Profiling tools will automatically generate value-frequency lists allowing analysts to quickly review the range and spread of data values as well as identifying duplicates and outliers where they were not expected.
- Instant “drill-down” access to relevant records based on identified features (e.g. a value or pattern) that require further analysis.

The following screenshot illustrates a column profile view where duplicate Customer ID numbers have been identified. The bottom-left pane contains the drill-down to affected records.

The screenshot shows a software interface for data profiling. The main window is titled 'Profile 7. CRM System.CUSTOMER.csv'. It displays a 'General profile assessment' table with columns: Attribute, In scope?, Analysed?, Percentage distinct, Inferred type, and Minimum. The 'Customer ID' attribute is highlighted, showing a percentage distinct of 99.25 and an inferred type of SMALLINT. To the right, a 'Values [Custome]' table shows the frequency of values for 'Customer ID', with values 1 and 7 both having a frequency of 2. Below this, a 'Data where [Customer ID] is selected values' table shows the records for Customer ID 7: Christine Manley, Kristina Chester, and Alex Smith. To the right of this table, a 'Patterns [Custon]' table shows patterns like 9(2) and 9(3) with their respective frequencies.

Attribute	In scope?	Analysed?	Percentage distinct	Inferred type	Minimum
Customer ID	N	N	99.25	SMALLINT	1
Contact First Name	N	N	86.29	VARCHAR(16)	Aaron
Contact Last Name	N	N	94.81	VARCHAR(14)	Abbibi
Contact Title	N	N	2.22	VARCHAR(4)	Dr.
Address1	N	N	100	VARCHAR(40)	1 Featherstone W.
Address2	N	N	26.29	VARCHAR(19)	

Customer ID	Frequency	Pattern
1	2	
7	2	
3	1	9
4	1	9
5	1	9
6	1	9
9	1	9

Customer ID	Contact First Name	Contact Last Name	Contact Title	Address1
1	Christine	Manley	Miss	410 Eighth Avenue
1	Chris	Christianson	Mr	7464 South Kingsway
7	Kristina	Chester	Miss	3802 Georgia Court
7	Alex	Smith	Mr	8194 Peter Avenue

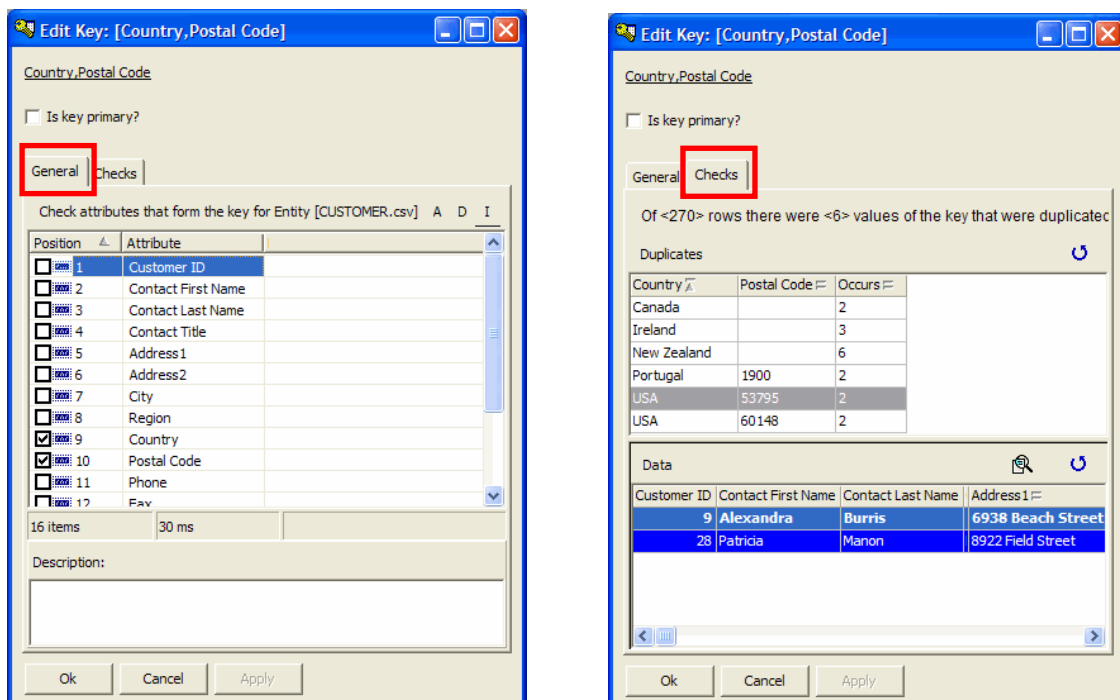
Pattern	Frequency
9	9
9(2)	90
9(3)	171

Many products that claim to be Data Profiling solutions only provide column profiling capabilities. To achieve a complete understanding of your data, it is essential to look beyond the behaviour of attributes in isolation so that relationships between attributes within and across tables/systems can be assessed and tested for anomalies.

Key Testing

Key testing allows the analyst to verify the uniqueness of primary and candidate keys within an entity (table). Since keys uniquely identify the data in other non-key fields, it is crucial that they perform the intended role. In the last example, we saw that the key, Customer ID was compromised because, in two cases, the same ID identified two distinctly different customers.

Key testing provides an explicit tool for assessing more complex situations – in particular, where keys are defined using combinations of more than one attribute. In the following example, we show an alternative key combination being tested using Country and Post Code (perhaps to assess compatibility with another system containing addresses, but no Customer IDs). Again, note how Data Profiling allows the user to interactively define and test the key and drill down to exceptions in a fraction of the time it would take to script an equivalent query and run it against the database.



Dependency Testing

Like key testing, dependency testing focuses on the relationships between data values across attributes (columns) in a single entity (table). However, here the emphasis is not on uniqueness, but consistency. For instance, if we believe that a combination of two field values [a,b] should always specifically match values in three other fields [x,y,z], we can test this correlation regardless of rows where the same values are duplicated. We are only interested in exceptions such as [a,b] determines both [x,y,z] and [x,y,y].

Dependency testing has particular relevance when restructuring data. The majority of older systems relied on merging tables (“de-normalisation”) to enhance performance. This results in the values in some columns not being directly dependent on the primary key, but on other column(s) in the table which, in turn, are dependent on the key (“transitive dependencies”). Although this technique offers performance benefits, it can also lead to anomalies creeping into the data which will cause problems if the data is subsequently moved into more optimised structures (say in a Data Warehouse or new ERP system), or even if the data is just being used in new ways and assumptions are being made about its consistency.

The following screenshot shows the result of a dependency test where the combination of specific Country and Postal Code values are always expected to return consistent City and

Region values. For instance, a combination of Country *UK* and Postcode *MKI* should always occur together with the City *Milton Keynes*. The example shows that City and Region values correlate only 97.04% and 97.78% respectively. Both relationships are thus classified as “Potential” and all compatible and incompatible value combinations are shown in the panes on the right. The highlighted rows show a drill-down to a specific pair of inconsistent rows. It is clear from looking at this example that the primary cause of the inconsistency is that there are missing Postal Codes leading to ambiguity in the dependency being tested. One might chose to retest this relationship after the issue of missing postal codes is addressed.

The screenshot shows a software window titled "Test 1. CRM System.CUSTOMER.csv[Country,Postal Code>City...]" with a timestamp of 10/07/2005 21:30:43. The interface includes tabs for "General", "Notes", and "Issues".

Summary Results:

Determinant	Dependent	Valid?	% Coverage	Rows	Compatible Rows	% Compatible Rows
Country,Postal Code	City	Potential	100	270	262	97.04
Country,Postal Code	Region	Potential	100	270	264	97.78

Compatible Values:

Country	Postal Code	City
Argentina	6755 56	Buenos Aires
Aruba	655456	Oranjestad
Australia	2061	Sydney
Australia	5353	Canberra
Australia	2155	Melbourne
Australia	4774	Churchill

Incompatible Values:

Country	Postal Code	City
Canada		Port Coquitlam
Canada		Vancouver
New Zealand		Christchurch
New Zealand		Hamilton
New Zealand		New Plymouth
New Zealand		Auckland

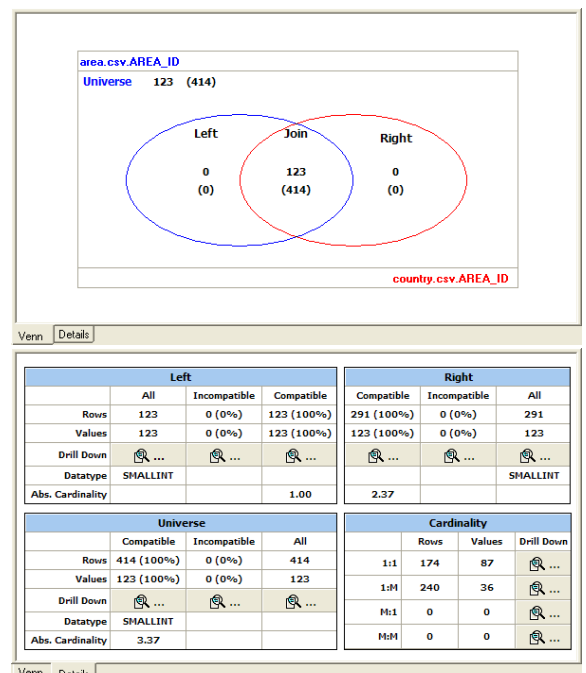
Records having selected values:

Country	Postal Code	City	Customer ID	Contact First Name	Contact Last Name
Canada		Port Coquitlam	24	Tony	Garneau
Canada		Vancouver	60	Dave	Elkins

Join Testing

Join testing is a vital step in any project relying on the ability to integrate data from separate sources. Consider a scenario where a data warehouse is envisaged to create a single view of customer data. The requirements may dictate that data is drawn from previously unconnected systems managing customer relations, accounts, ordering, and deliveries. Any assumptions about the common data (such as customer and account codes) that will be used to merge the sources must be thoroughly tested as part of assessing the feasibility of the project. There follow two illustrations.

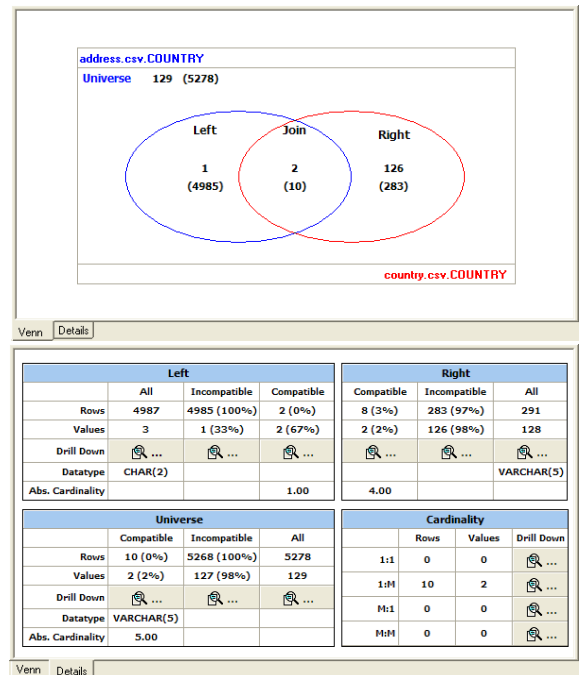
The first (right) shows a clean result from a join test where there is a complete overlap of AREA_ID values between a Sales Area table containing 123 records (all with unique values) and a Country lookup table containing 291 records (containing the same 123 unique values). The fact that there are duplicates in the Country lookup is not a concern as long as each Area matches a Country. The cardinalities in the Details table indicate a



consistent relationship where records either match 1-to-1 or 1-to-Many (1:M overall). Note that, even with a full overlap of values in the Venn Diagram, any combinations of values exhibiting both 1:M and M:1, or any M:M, relationships would make integrating this data almost impossible since it would not be clear which Area record matched which Country record when alternatives exist on both sides.

This second example shows a very different picture. In this case, there is a radical mismatch of COUNTRY values between an Address table and the same Country lookup we used above. Only 2 values are common (accounting for a mere 10 records between the tables).

We may not be surprised to see that there are orphan COUNTRY codes if we already know that we only do business with a handful of countries in the lookup. However, there is an assumed business rule that all addresses we keep should match a country. This test has proved that all but 2 do not. Using drill-downs we soon establish that the orphan COUNTRY value in the Address table is **UK**. Similarly, we uncover an unused orphan in the Country table of **GB**. Although these values have slightly different semantic meanings, we conclude that there two systems were implemented using different standards and these values should be assumed equivalent for the purposes of the project.



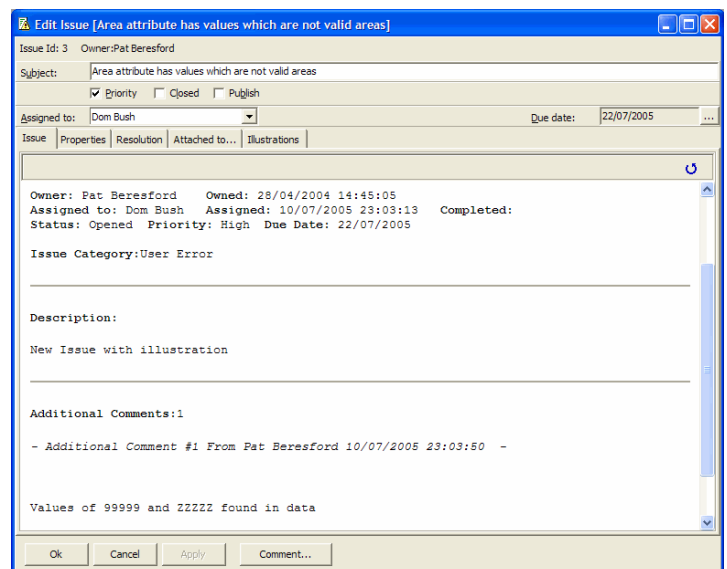
A prudent choice in the light of this discovery would be to standardise the values across the systems as a cleansing task and retest the join to ensure consistent cardinalities.

Note that well-implemented Data Profiling solutions will automatically support join tests between attributes even if their names and datatypes do not match exactly.

Issue Management

Managing issue progress can be a challenge in itself. In one single view of customer data warehouse project being run at a leading telecoms provider in the UK, a team of data analysts were identifying some 300 issues per week. With most projects like this, issues are managed separately to the analysis process using tools like Excel or “home-grown” databases. In either case, tracking issues back to source data is practically impossible and managing the issue list becomes yet another overhead.

No Data Profiling technology is complete without integrated support for documenting and managing the lifecycle of data quality issues unearthed. Although it will often be the case that cleansing and transformation tasks are carried out using other cherry-picked tools, it makes sense to have a central shared repository of all open and closed issues. It also makes sense to provide this functionality where the issues are identified so that they can be linked in the repository to the source structures and records that exhibited the problems.



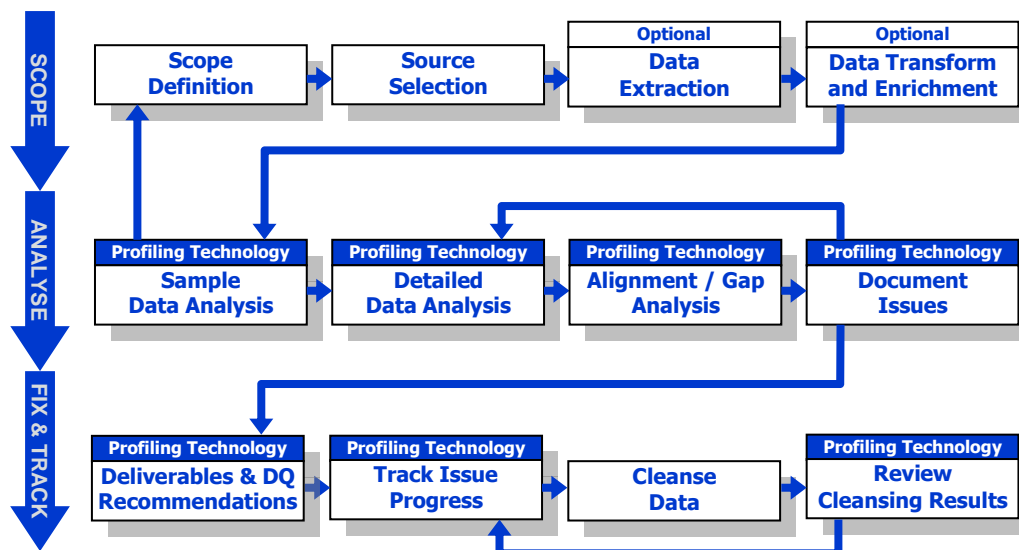
Few Data Profiling offerings provide more than the option to attach simple notes to data structures and values. Fewer still support issue resolution workflow where issues can be attributed with details such as ownership, assignment, priority, status, categorisation, and their histories tracked for management purposes. Add to this the ability to report on issues by any combination of these features and one can provide complete visibility for issue management.

Deploying Data Profiling Technology

Data Profiling technology can readily be integrated within existing data management functions. It supports, accelerates and enhances most activities related to data analysis. Analysts can achieve their goals to a greater depth in a fraction of the time normally required with such tasks.

In the context of implementing new systems and solutions, Data Profiling becomes part of the project toolkit along with other potentially necessary technologies such as Cleansing and ETL (Extract Transform Load) capabilities. But, due to the enhanced ability for auditing data sources relevant to the project, Data Profiling can provide a much clearer view of any existing quality issues at a time when decisions about downstream project activities are still being finalised. Hence, more informed decisions can be made about the project timescales, approach and which other technologies are required for successful project completion. In many cases, being able to perform a complete, rather than partial, data assessment can form a crucial element in qualifying the feasibility of the entire project itself.

The following chart illustrates how Data Profiling technology can be applied within the data analysis and preparation phase of a typical project:



Beyond specific projects, Data Profiling also delivers significant benefits to any user engaged in assessing data on even just an ad-hoc basis. Whether dealing with large or small volumes of data, profiling can accelerate analytical work and provide a bridge for communication between technical, business and management members of a team.

Furthermore, as companies increasingly look to implement ongoing strategic data quality initiatives, Data Profiling technology offers existing corporate analysis and data management resources the tools required to analyse data within meaningful timeframes and the means by which to monitor progress and performance. Investing in data assets is only as valuable as the ability to sustain quality going forwards.

Conclusion

Who does Data Profiling benefit, and how?

As a mature technology, it has been amply demonstrated that a Data Profiling led approach can deliver tangible value across the business when applied to the challenges of data analysis and quality management. Adoption is straightforward and the potential returns on investment very significant. At the enterprise level, its ability to raise and maintain the quality of corporate information promotes competitive advantage and cuts costs.

The following summarises the direct benefits that can be expected:

Data Analysts and Data Managers

- Step improvements in analysis performance through automation – do more in less time
- Significant increase to achievable breadth and depth of analysis scope
- Far clearer understanding of data content and business rules
- Facilitated communication between analysts, business users and quality managers

Project Managers

- Visibility of all data quality issues and their current statuses
- Condensed and achievable delivery timescales
- Reduced risk of project delays and budget over-runs due to unexpected data quality issues

System Owners

- Sustainable data quality
- Diminished operational costs
- Fewer disruptions and less manual intervention
- Ability to deliver a better value service

Data Owners/Stewards

- Framework for effective delivery of data quality strategy
- Ability to meet data quality responsibilities
- Greater confidence in data assets

Executive Management

- Effective business processes based on accurate, complete and trustworthy information
- Better guarantee of a return on investment from corporate systems and data assets
- Reliable information supports better strategic and tactical business decisions
- Increased profitability through improved efficiency and customer management

Business Data Quality Ltd (BDQ) is an independent software company specialising in next-generation solutions aimed at helping organisations realise the true potential of data as a strategic asset through the development of practical data analysis and management products.

BDQ Suite allows you to analyse, track and report on the quality of data stored within IT systems. Combining data profiling and monitoring technologies with a multi-user repository and issue management framework, BDQ Suite delivers a detailed and accurate view of corporate information from a data quality perspective. Aside from issue identification, it also addresses the second core ingredient of an effective data strategy – quality management. Administrative data activities and remedial cleansing tasks can be documented, assigned and tracked through BDQ's integrated issue management platform, giving visibility to improvements in quality and, thereby, raising confidence across the business.

For further information about BDQ products and services, please contact us at:

Business Data Quality Ltd
Block A
Southgate Office Village
284 Chase Road
London
N14 6HF
United Kingdom

Tel: +44 (0) 870 429 5236
Fax: +44 (0) 208 886 0814

enquiries@businessdataquality.com
www.businessdataquality.com

Copyright © 2007 Business Data Quality Ltd

All rights reserved. No part of this publication may be disclosed in whole or in part, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopied, recorded or otherwise, without the written permission of Business Data Quality Ltd. All other company and product names mentioned herein may be trade names or trade marks of their respective owners.