STRATEGY PARTNERS INTERNATIONAL

Chasa Ursa Major
CH-7550 Scuol
Switzerland

Data Quality – the Base for your Enterprise Applications

White Paper - Version 1.0

August 2007

Helmuth Gümbel - Managing Partner

Strategy Partners International

Chasa Ursa Major
CH-7550 Scuol
helmuth.guembel@strategypartners.com

Text and Illustrations were compiled with utmost of care. However, the author will not take responsibility or be held liable for any possible remaining incorrect information and the resulting consequences thereof.

Table of Contents

Why you need to read this

Data Quality always was an issue ever since electronic data processing started to happen. IT veterans like the author recall the once ubiquitous motto “garbage in – garbage out” which pretty much expresses in a nutshell what data quality is about.

While the trend recently was more on focusing on process quality, the data quality issue really was exacerbated despite the comparatively low level of attention given to it. Processes require correct data to perform – quality data is meaningless without correctly designed and implemented processes. We may compare the processes to the human organs and data with the blood – one just cannot exist without the other.

Recent developments have, however, carried the issue several steps further. Tight application integration requires sharing of data. Improved application infrastructure enables users to build composite solutions with applications interacting in a more flexible and powerful way than ever before. Higher processing speeds and strongly increased transaction rates cause more data changes per time unit than ever before. We now are able to collect a host of data from very different sources and create relationships. The Internet, mergers and acquisitions, fast changing market scenarios – all these factors have provided us with both more data and higher volatility.

At the same time, emphasis on transparency and traceability is strongly increasing and pressure to implement suitable underpinnings is increasing on various levels. Sarbanes Oxley, ISO 9000, Basel II, FDA certification procedures – they all point into the same direction: ensuring traceability, transparency, and ultimately, quality.

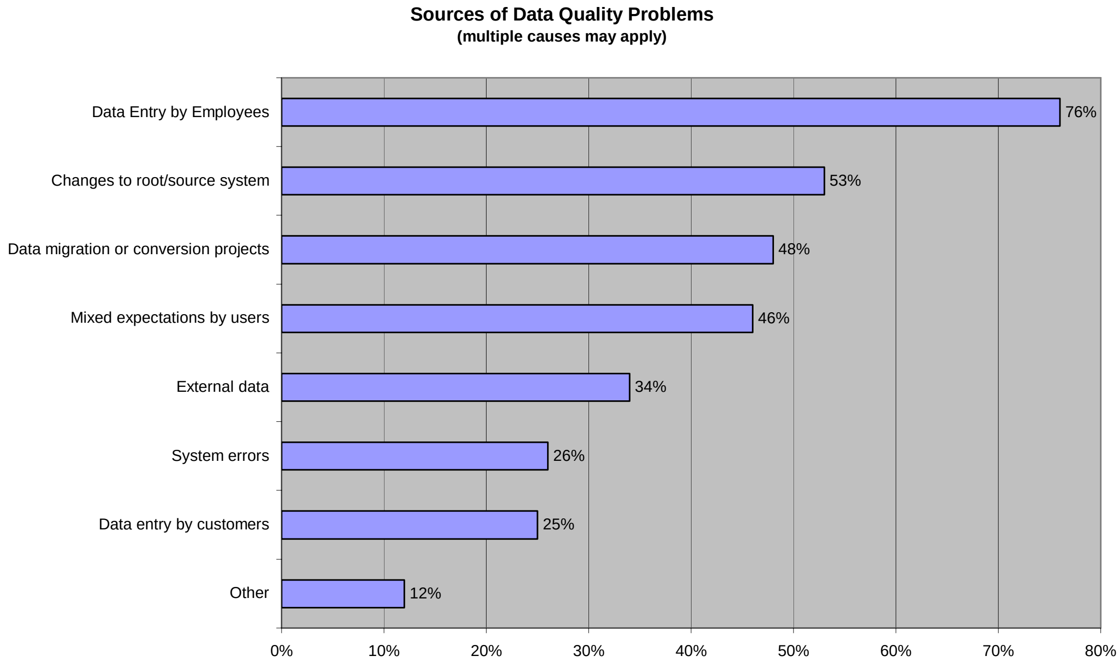
Data quality problems can spread like a disease in an application environment. The effects may be similarly devastating as in the case of modern global health threats like SARS. The reason is also very similar: improved infrastructure, higher traffic rates, and the tearing down of barriers contribute to the potential for epidemics.

Data quality is of such central importance that it cannot be left up to the individual applications as this would not allow for centrally enforced quality standards, policies and procedures. Rather, a carefully architected, centrally administered approach is required to ensure ongoing data quality. Any new insights gained from newly arising data quality issues can be quickly and universally adopted.

Data migration and data population offer opportunities to achieve quantum leaps in data quality as most packaged applications to not offer easy ways to correct high volumes of data once it has been entered into the systems. Many SAP installations will upgrade from R/3 versions to mySAP ERP shortly due to SAP having terminated standard maintenance. Most of these will perform merely “technical migrations” – upgrades that port the preexisting functionality and data to the new software environment and higher capacity hardware. Such migrations do not add significant visible benefits – they are only meant to help escape the added costs of extended maintenance. In our mind, this constitutes an opportunity to establish the base for better data quality as data could be cleansed prior to and following the upgrade.

Corporate Data – a Fundamental Asset

Corporate data is the foundation of any enterprise. It is generated from a variety of sources that never have the same level of quality and consistency. The Data Warehouse Institute has ranked the different sources in terms of their likelihood to impact data quality negatively:



- Figure 1 Data Quality Problems mostly have several causes (Source: TDWI)

Corporate data today is understood in a very holistic way. Data that used to be separately kept is now increasingly viewed in context. This includes all varieties of data – formatted data and unformatted data, engineering data and commercial records, mail as well as transactional data.

Data volume is increasing rapidly due to a number of factors. Some are very intrinsically connected with the nature of the underlying activity (many engineering activities today have become tremendously data rich because there are many simulation possibilities that reduce the requirements for models and prototypes). There are, however, a number of common factors that apply to most businesses globally:

Many industries face price and margin decays that can only be compensated by higher volume resulting in more transactions that increase data volume.

Increased competition causes enterprises to accelerated business model changes. Many of the new strategies are based on hybrid models. Again, more data is produced and recombining the data in various ways becomes essential when controlling results and developing strategies.

Macroeconomics and demographic changes play also into the dramatic data volume increase. Today, enterprises can address more customers than ever – many of them globally spread. New markets (like India or China) have become very attractive. As purchasing power increases in these people rich countries, business activity on all levels will increase. Inevitably, more data from a variety of sources will be required and more data will be produced as a mirror image of the products and services delivered.

The need for compliance and transparency becomes more obvious on all fronts. This need can be imposed by regulatory measures such as the Sarbanes-Oxley act or Basel II, it may be the result of an already established procedure like FDA drug approvals or ISO 9000 certification. Corruption scams such as recently experienced by Siemens[1] cause lengthy and expensive investigations involving very deep data analysis. Today, it may look desirable to have routine access to high quality data and to be able to track any changes, tomorrow it certainly will become de rigeur.

Having accessible data is one thing, turning it into useful information another – it requires reliable data quality. Data quality is influenced by a number of factors. Some of these are of a more technical nature while others are based on organizational and business aspects. Some of these factors have positive influences while others have a negative impact.

Negative technical factors are interestingly enough rooted in technological progress: cheap storage, high-speed communications, and the ability to collect a large variety of data easily from very different sources (all usually touted as blessings) help to amass enormous amounts of data. It increases every day and the growth even still accelerates. This data is very different in structure (if there is any) and semantics. Since corporate data design (once a strongly hyped noble endeavor) proved to be unfeasible, data in enterprises is, at best, a patchwork with small islands of consistency. Most organizations have not yet found out

[1] A series of corruption allegations caused the resignation of the CEO, the chairman of the supervisory board and trigerred public prosecution. Siemens expects to be subject to huge fines imposed by US authorities. Millions of Dollars have been spent so far attempting to get all relevant data on the allegations. Estimates on costs to achieve compliance have not been divulged so far.

that it is better to have less, but consistent and reliable data than having huge amounts of low quality data.

When implementing new applications, focus is usually on the functional and process side. Data is “shoe-horned” from pre-existing systems using technical conversion processes that perform only minimal and often more cosmetic data cleansing. It is fascinating to see the massive disappointment that sets in when the new application fails on poor data as badly as the old did.

Application migrations (usually version upgrades) similarly help to proliferate the data quality issue: rather than quality assuring the data before new software functions are exploited, the old data is kept and fed into applications at face value.

Here, the deep application integration that we have learned to appreciate because it allows us to do everything from everywhere if you are authorized shows its Achilles heel: it can only function correctly, if the data is right. Usually, this is not even tested.

Negative organizational factors usually originate from the ever increasing business dynamics. Mergers and acquisitions bring many data assets together that are never the same in design and quality[2]. New practices such as computer aided customer relations management or fraud detection attempt to exploit synergies that may be gained by putting all suitable data sources into a new perspective. These data sources, however, have not been designed to be used together and they differ in many aspects. Some experts even doubt that applications like ERP and CRM, which share many important data assets, can use the same data model.

Even data warehouses, specifically designed to enable analytical applications to process data from a potentially large variety of sources, cannot eliminate semantic data problems.

Organizations depend increasingly on correct data. Increasingly, good data is not an option, but a must that is enforced by standards and government regulations. As data volume grows, the issue is becoming of vital importance. Successful corporate strategies require quality data. Compliance, on the other side, is impossible without good and comprehensive data as a reliable base. Quality data is a key asset that cannot be substituted.

[2] Exceptions to this rule exist in the few cases where data has been acquired from the same source AND has been quality assured using the same standards.

Why Packaged Applications cannot guarantee Data Quality

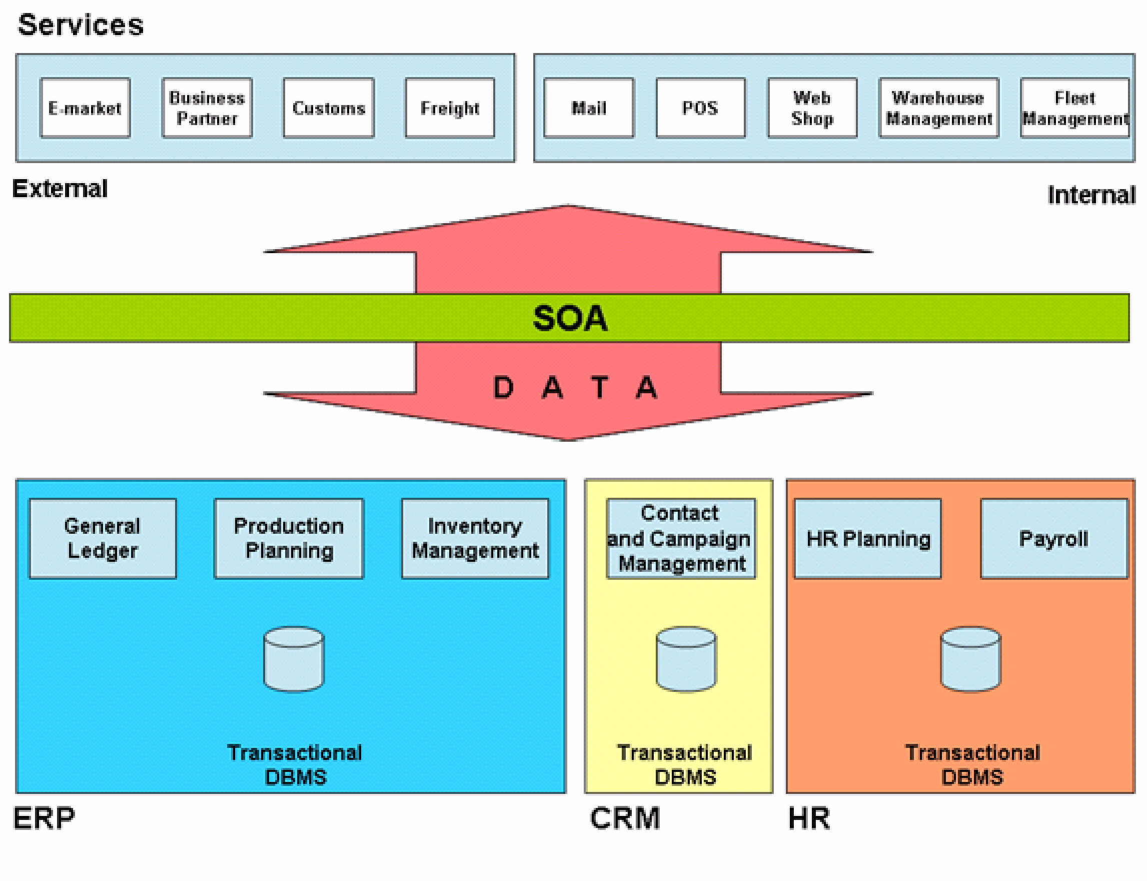
Packaged applications such as ERP or CRM software do a great deal to ensure data integrity on a transactional level using middleware such as database management systems. They also ensure a certain level of consistency when new data is entered. The level of consistency and the degree to which this can be judged, however, depends on various factors:

- Plausibility checks vary by product. There are no hard rules on what is being checked and how the checks are being performed. There is no consistency even between products of the same vendor as history and legacy play a strong role.
- There is rarely (if ever) full documentation on what is checked and how. Hence, you often can only guess what the application does to ensure data quality.
- Vendors change applications aiming to improve them. Again, data quality impacts are not reliably documented.
- Data is shared among several applications that may update the data using a variety of semantic integrity standards. There is no central device that allows organizations to set and control these standards across the applications sharing the data.
- Data is sometimes entered into the system bypassing checks that may exist in the application. This happens frequently when volume batch updates are applied. Another cause is the entry of data by users that are not or poorly assisted by their system, as it is the case when sales persons key in “raw” and often incomplete data into their CRM-systems.

In short, the usage of packaged applications is no guarantee at all to arrive at a high and transparent data quality level.

SOA and Data Quality: Be Prepared!

SOA is positioned as a key enabling technology allowing users to freely combine and recombine data and services. Rather than using proprietary point-to-point interconnectivity, SOA uses a set of standards that supposedly allow for "plug-and-play" interconnectivity regardless of the location of the interconnected functions. These functions are called services in the context of SOA and they have technically well defined interfaces that allow for a high degree of implementation independence. Data is owned by the services and data formats matter only between services, not within.



- Figure 2 SOA interconnects applications and data (Source: Strategy Partners International)

While SOA certainly helps with interconnecting services and designing cross platform processes, it does not solve all problems associated with application integration. SOA defines a format for data interchange (XML), but it does not define any measures that ensure data integrity or data quality. In making it easier to interconnect, SOA also makes it easier to infect IT-environments with bad data. Similar to computer virus infections, the effects can be disastrous and very difficult to cure. To that extend, SOA is not only an enabler of new integration scenarios helping to address business challenges more flexibly but it also creates a much higher possibility for chaos. Very much like using the Internet, where we have learned to guard ourselves from the perils with firewalls and anti-virus software, SOA can pose serious threats if data quality measures are not taken. It can spread over the whole application structure and cause massive problems that may even be propagated into other SOA structures. SOA fundamentally is a network architecture and it shares all the possibilities for good and bad synergies that networks have by design.

SAP Users: Prepare for Quality Migrations

When new applications are installed, data frequently is migrated from the preexisting application. Usually, the format is different and, apart from the necessary technical conversion, some data cleansing is performed to prevent the most obvious failures. Very rarely systematic data quality assessment and improvement is practiced.

Migrating from one application version to the next usually is not accompanied by data quality activities at all. It is mostly seen as a routine technically dictated upgrade that is a necessity and that should not consume more resources than absolutely required.

The situation of most of the SAP installations is a little different. While most of these installations have upgraded to a mySAP license, their installations are actually running versions whose regular maintenance is about to be terminated by SAP. SAP wants its customers to migrate believing that this can enable them to use more functions.

The first step of such a migration is purely technical. Customers install the new software and perform only those adjustments that are absolutely required. Usually, very little effort is put into removal of dead wood (customizations no longer required) or data quality analysis. Most companies allocate budget strictly on a bottom line impact base. Version migrations do not have an obvious and positive business impact.

SAP positions mySAP ERP as the base for “SOA by evolution”. Customers are meant to exploit the benefits of SOA step-by-step. As SAP aims to broaden its presence in customer accounts attempting a higher share of wallet, SAP will try, and often succeed, to sell more vertical applications and additional functions. It is no surprise that one of the most successful products in this context is SAP the compliance suite that was acquired from Virsa. Another popular addition in SAP accounts is Enterprise Content Management.

Typically, SAP applications are positioned as corporate hubs and back office solutions acknowledging the central and pivotal role of these products. The original concept of tightly integrating applications around a single database (the single instance of truth as it was often put) contributed strongly to this view.

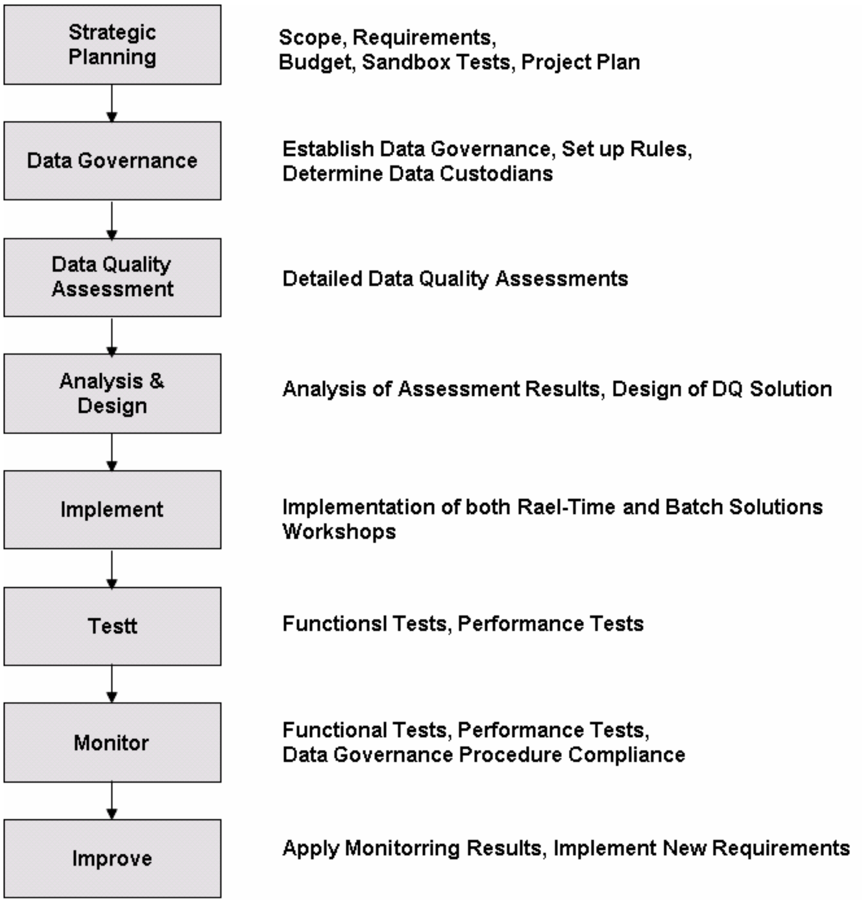
Over time, most SAP installations witness a certain deterioration of data quality that becomes obvious when new ways of using data are practiced. No such extension should be ventured without a profound analysis of the data quality requirements and impacts. The best time for a start into systematic data quality assurance is before or right after the technical migration to mySAP ERP and, definitely before any functional or architectural extensions happen.

A Plan for Consistent and Persistent Data Quality

As data quality is both a vital and strategic issue for most organizations, it takes more than just a search for duplicate entries in a database to ensure that comprehensive, consistent, and persistent results are achieved.

Project Plan

Establishing data quality is not a trivial and quick endeavor at all. It requires systematic and persistent action to ensure results. Hence, a carefully designed project plan is required. As with all plans that cut through whole organizations, individual adaptation is critical. Expectations have to be managed carefully as it is not possible to solve all data quality issues in one step.


Strategic Planning
Scope, Requirements, Budget, Sandbox Tests, Project Plan
Data Governance
Establish Data Governance, Set up Rules, Determine Data Custodians
Data Quality Assessment
Detailed Data Quality Assessments
Analysis & Design
Analysis of Assessment Results, Design of DQ Solution
Implement
Implementation of both Rael-Time and Batch Solutions Workshops
Testt
Functionsl Tests, Performance Tests
Monitor
Functional Tests, Performance Tests, Data Governance Procedure Compliance
Improve
Apply Monitorring Results, Implement New Requirements

- Figure 3 Sample high level project plan (Source: Strategy Partners International)

Data Governance

Data quality is an organizational issue augmented by IT. Fixing quality problems requires definitions on what the organization needs to function properly. Data governance requires:

- Organizational Awareness. Both lines of business and IT-departments have to understand the importance of the issue and the impacts. Data ownership needs to be defined.
- Rules defining correct data.
- Policies and procedures on security, data privacy, and compliance
- Data quality improvement measures
- Data architecture
- Lifecycle management
- Audits and reports

Data governance is an ongoing task that will develop as data quality matures. There is no universal blueprint for data governance as it requires measures that are very specific for each enterprise. It is the very foundation for technically assisted quality measures.

Technology Evaluation

When evaluating technology that may help with the implementation of data quality, it is important to go beyond point solutions that only address problems like finding duplicates or postal address validation. Data quality is a very comprehensive issue that involves all data that an organization receives, produces, and stores.

Hence, data quality products must cover these three aspects. Solutions help when

- Loading the data
- Processing data in real tine (like duplicate entry search and prevention)
- Tracking changes
- Publishing and distributing data

Different from applications, which are frequently licensed from a vendor, most data in an organization is its property and will persist in some form and fashion even if the applications are changed. Hence, it is important to view data quality and the related technology independently of applications.

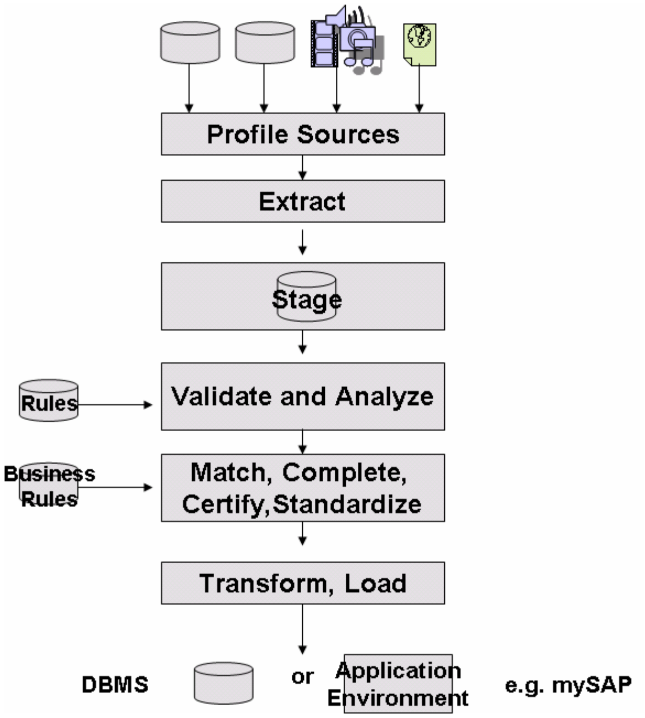
Key criteria for technology evaluation are[3]:

- Completeness – coverage on all levels, easy installation and good connectivity with applications
- Enterprise Design – coverage for all relevant data types
- Ease of use
- Sophisticated rules engine
- Best record capability – the software can successfully select between duplicates on a rule base
- Business users can participate in the data quality process
- Coverage for all data points of entry
- Processing power – very much needed as data volume grows
- Vendor support – you will be entering a long term relationship with a vendor that helps you to guard the "crown jewels"

The Architecture

As the initial creation of a quality data environment is an I/O and compute intensive step, it cannot be done in real time while applications using the data are running. Over several stages, data is extracted, profiled, validated, enriched/completed where applicable and possible, matched against additional sources and, finally, after re-formatting and transformation either loaded into a DBMS or routed to an application environment such as mySAP.

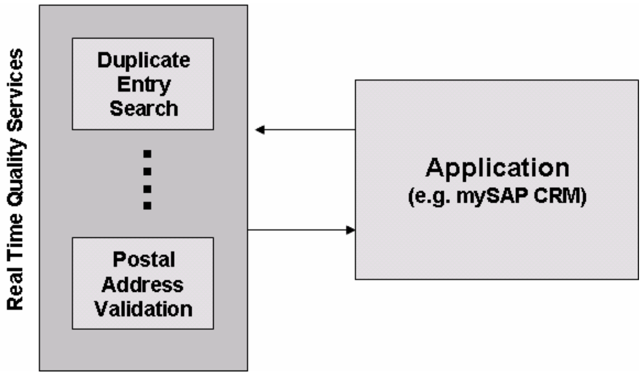
[3] More on this in „10 Questions to Ask When Evaluating a Data Quality Solution“ by Thomas Brennan and Steven Kleinmann, Stalworth Inc. to.brennan@stalworth.com, steve.kleinmann@stalworth.com



• Figure 4 Batch data quality environment (Source: Strategy Partners International)

Some data quality services can, and should be, performed in real time calling from the transactional system. Increasingly, we expect such external invocations to be based on SOA standards.

This should lead to more real time quality services to appear on the market. Interface standardization creates a better technical environment for vendors while demand, as explained before, is building up. In an SOA-environment, data quality services can use centrally created definitions and rules. These can be shared across the enterprise contributing greatly to corporate data quality and governance.



• Figure 5 Real time data quality services (Source: Strategy Partners International)

Bottom Line

Data quality has been neglected over most of the history of IT. If addressed at all, it has been worked on with various point solutions.

We feel that the days of this somewhat lax attitude are over – the volume of data, the degree of dependence on its quality and, last, but not the least, regulatory demands require a much more determined and consistent approach.

IT environments will remain heterogeneous notwithstanding attempts to the contrary. Application portfolios will change and develop as the needs require. Data quality cannot be guaranteed by a single application or by blindly sourcing middleware from application vendors. Data quality is the responsibility of the data owner – hence, best of breed is not an option but rather a must.

Data quality is also independent of middleware stack architectures[4]. Regardless of the SOA middleware stack selected, users should select and insist on using the best solution available for their data.

[4] IBM actually offers its data quality products as part of the IBM Information Server module WebSphere Information Analyzer and QualityStage. This is purely a branding matter – they can be used in total independence of WebSphere.

Table of Figures