Address Data Quality

The Foundation of Operational Effectiveness

What is Data Quality?

Simply stated, data quality involves ensuring the accuracy, timeliness, completeness, and consistency of the data used by an organisation while also making sure that all parties utilising the data have a common understanding of what the data represents. For example, does product sales data include or exclude internal sales? Is it measured in units or Euros, RON or, perhaps even Dollars? The scope of a data quality initiative is not limited to the data generated by an organisation's own operations; it must include data obtained from external sources. While some definitions would also include accessibility, this is not strictly part of the data quality definition, although it is certainly a desirable and pragmatic characteristic for the data to be of practical use.

Given the current emphasis on the need to maintain a "360 degree view of the customer", some people naively assume that this involves the creation of one massive record for each customer. Rather, data quality involves being able to link all of a given customer's records together – a task that can only be accomplished with identifiers for the records associated with each customer.

Municipalities looking to collect their fair share of tax revenues need to ensure that neighbouring municipalities are not incorrectly collecting their revenues. Data quality software provides for accuracy in the tax jurisdiction assignment process, thereby helping telecommunication companies remain within the law and helping ensure that customers and local municipalities are treated fairly in terms of taxation and revenue apportionment.

The Importance of Quality Data

Organisations make decisions based on the data available at that time. If an organisation can improve the overall quality of this data, it can improve the quality of the resultant decisions and increase both its effectiveness and efficiency. This will enable the organisation to better, and more profitably, serve its constituents, be they customers, employees, business partners, or stockholders.

Much attention has been focused on the data cleansing aspects of data quality. However, this is only one part of the equation. Other aspects include data integration and consistent business rules.

While data quality may once have been considered a nice-to-have initiative, organisations now realise that it is an absolute necessity especially for mission-critical applications or those that are required in order to meet governmental reporting and disclosure requirements. In fact, when applied to applications related to Homeland Security or .eg. Solvency II, data quality quickly becomes a serious corporate concern.

What is the Magnitude of the Problem?

According to a report published by The Data Warehousing Institute (TDWI) in 2002, "poor quality customer data costs the United States a staggering $611 billion a year in postage, printing, and staff overhead." TDWI cited several examples including a telecommunications company whose data entry errors incorrectly coded accounts and lost $8 million a month when it couldn't send out bills. The true cost is undoubtedly much higher, as the $611 billion dollar amount was limited to customer name and addresses data and even then did not include secondary effects such as those associated with alienating and losing customers.

We also know that the problem is persistent. In a more recent 2009 study conducted by Gartner, participants estimated that poor data quality cost their organisations an average of $8.2 million a year as a result of data quality issues. However, twenty-two percent of the respondents calculated their annual losses at $20 million or more and four percent indicated annual losses as high as $100 million.

While losses of millions of dollars are significant, Gartner analysts believe these figures understate the true financial impact on most organisations.

Much attention has been focused on resolving name and address issues or consolidating multiple

records in order to provide a single customer view, or eliminating duplicate mailings to the same person or family. However, these are just a few examples of where data quality can be applied to benefit an organisation.

Data quality applies to more than just customer name and address data. It applies to product numbers and associated descriptions, part numbers and units of measure, medical procedure codes and patient identification numbers, telephone numbers, e-mail addresses, commodity codes, vendor numbers, and vehicle identification numbers, to name just a few. For example, if an insurance company sells some of its products in RON, others in Euro and some in dollars, then the potential for error is considerable.

Remember the Mars Climate Orbiter (1998) that crashed because of an incorrect conversion from metric units to feet and inches!

Additional Effects

Poor data quality can negatively influence how a company is perceived in the marketplace. The first impression should be one of quality. Just as visitors to a company's offices can be strongly influenced by their initial impression of the lobby and reception area, the way a company is perceived by others can be influenced by the quality of its data, especially if it results in miss-addressed mail, incorrect invoices, or erroneous shipments.

The inability to eliminate redundant name and address records results in unnecessary postage costs. Recipients of duplicate mailings can become frustrated and question the firm's overall operating efficiency.

If these redundant mailings each inconsistently misspell the individual's name or address, the frustration level may approach alienation or even a legal concern.

Add to this the cost of catalogues or even merchandise delivered to the wrong address (some of which will not be returned or, if returned, may now be considered as used), and the real magnitude of the problem only just begins to surface. Furthermore, if a single customer is included in a company's database multiple times, each time with a different value for the customer identifier, the company will be unable to determine the true volume of this customer's purchases. It could even be placed in the embarrassing situation of attempting to sell the customer an item that he/she has already purchased from the company!

Techniques and Solutions

Data quality is a multi-phase process involving data capture, data integration, data profiling, data cleansing, and data augmentation.

Capturing and Collecting the Data

Data capture involves the capturing and collection of source data. It can include a wide variety of sources and input mechanisms and can be conducted in both real-time and batch modes. Customers entering orders over the web, problem and resolution codes in call centre logs, and point of-sale data collected by cash registers are all examples of primary data collection sources.

While data can be cleansed at any point, it is optimal to ensure its accuracy from the start. Once an error is introduced into a system, it is usually more expensive to correct it after the fact than it would have been to take the proper steps to avoid it in the first place.

It is interesting to note that, with the increased reliance upon web-based order forms completed by consumers, these individuals have now become data entry clerks. However, unlike data entry professionals who are trained on the meaning of every field, many consumers are satisfied merely to accurately enter only the data that is needed to ensure the receipt of their order.

Catching and correcting errors at the source can significantly reduce the magnitude of data collection errors and reduce problems further downstream, for example, by utilising software that can verify an address to see if it actually exists.

An invalid delivery address strongly suggests a data entry problem that in a real-time environment could be corrected on the spot. Other software exists that can determine appropriate county and local tax jurisdictions and associated tax rates and apply these to customer purchases to determine the correct tax amounts. These are especially effective if implemented at the data entry source, integrated with the transaction system, and then be available in real-time. Even if professional data entry clerks entered the data, as is often the case when placing an order via the telephone, it is important that the data be collected in a format suitable for both its immediate use (e.g. order shipment and payment) and long-term use (e.g. analytical purposes such as identifying future up-selling and cross-selling opportunities).

A healthcare organisation that acquired several other smaller healthcare organisations discovered that a newly acquired company was quite careless in its data entry procedures. Social Security numbers, used as the patient identifier, were routinely entered inside of a freeform patient name field. The placement was seemingly random as the Social Security number sometimes appeared at the beginning of the name field, sometimes between first, middle, last names, and sometimes even within one of these names. Only after the company deployed data quality software capable of recognising the pattern for a Social Security number was it finally able to merge the patient records of its newly acquired company with the patient records of its other companies.

Data Integration

Organisations collect data from multiple sources. In many cases these sources each have their own data format. Data integration involves combining the data from these multiple, disparate sources. These may include data contained in packaged enterprise application software systems, subject-oriented databases such as a parts master file, or call logs from a customer support help desk. While many organisations have tried to write their own data integration software, even those that have succeeded quickly realise that this becomes a never-ending task as the application-specific extract modules must be continually maintained and synchronised with each new release of the application software. Most organisations have found that it is more efficient to utilise third-party data integration software rather than trying to develop and maintain custom program code.

Data Profiling

It is not uncommon that the file specification documentation does not always accurately reflect the actual file contents. This is especially true in legacy environments where programmers have continually modified home-grown applications and may have added new data fields, or used existing data fields to store new data elements without documenting these changes.

By examining individual record fields or columns, data profiling tools can determine whether these fields conform to their assumed content, including data type and allowed values or value ranges. In addition, they can validate intra-record and inter-file dependencies. Checking that a person's gender does not contradict the salutation associated with his or her name is an example of intra-record dependency; checking that a part number in a customer order matches a valid part number on the part master file is an example of an inter-file dependency.

An example of both is checking that an employee's salary falls within the range specified by his or her job code. In order to determine if the job code and salary on the employee's record are consistent with each other, the salary range associated with the job code must be determined by checking the appropriate entry on the job code file.

Data profiling frequently accompanies the data cleansing process or precedes it to identify areas of concern or to validate assumptions. The earlier in the data quality process that data profiling is conducted, the sooner problems can be discovered and the work required to correct them is minimised.

Data Cleansing

Once the data is collected and merged from these multiple sources, it must be cleansed to correct inconsistencies and errors. While verifying and

Has this ever occurred in your organization?

Issue: After receiving the latest quarterly company metrics and statistics report, the head of sales is somewhat alarmed to see that the total number of customers has declined from 700 last quarter to 600 this quarter. Yet after asking her sales people to inform her of the number of new customers and the number of lost customers, the collective numbers indicate that the company gained 60 new customers while losing only 10, for a net gain of 50. Is the report wrong?

Cause: The report is correct. The apparent customer count inconsistency was due to the fact that the company discovered some customers were represented more than once in its database. During the quarter, the company discovered 150 'duplicate customers'. After cleaning the database to consolidate and remove these duplicate entries, the actual number of unique customers was 550, rather than the 700 that had been reported at the end of the previous quarter. Adding the net gain of 50 customers that occurred during the quarter to the true 550 number at the end of the prior quarter, demonstrates that the correct number is indeed 600.

correcting data as it is entered into the system can minimise errors, this won't necessarily catch inconsistencies across systems.

For example, if each division creates its own vendor file and uses a different identifier for the same vendor, it will be necessary to reconcile these files in order to recognise that what appears to be several different vendors can in fact be one. Data cleansing should also ensure that the data values are timely.

When comparing two records to determine if they contain the same data values, it is beneficial to have the data values in the same format. Telephone numbers, dates, names, addresses, identification numbers (e.g., customer, employee, vendor, product, procedure codes, etc.) can all be better compared if the data values for a given field are in a common, standardised format. While many data integration tools can convert source data from disparate files to the desired target format by, for instance reformatting telephone numbers to eliminate non-numeric characters such as parentheses, decimal points, or dashes, their ability to further cleanse the data is limited. They can't check that a home phone number area code is consistent with the postcode for the home address, for example.

Most pure data integration tools also cannot directly parse addresses into standardised values (e.g., St for Street) or convert a free-form entry into the appropriate individual fields (e.g., separating an unstructured name field into last name, first name, middle name, and title entries), or perform the necessary matching in order to recognise that. 'Popescu at 2/3B/16 Libertatii' is probably the same person as 'Nelutu Valentin Popescu at Nr.2, Bl.3, Sc.B, Ap.16 Bvd Libertății'. And only a country-specific data cleansing tool can recognise that in Bucharest, Splai is another name for Splaiul Independenţei .

Taking this one step further, the ability to link individual family members to the same household is clearly a task for the data cleansing, not the data integration, tool.

A good solution should accurately resolve most data cleansing problems and identify those that it can't and place them in a "suspense file" for follow -on resolution by a knowledgeable individual. Some products allow users to specify a threshold level as to the degree of ambiguity required before human involvement is required.

The power of data quality software is clearly enhanced by being integrated with robust data movement technology and the ability to share tasks as appropriate. This is usually accomplished either through strong technical partnerships between data quality and data integration vendors, or by having a single vendor provide both. In general, when a single vendor provides both, the integration is likely to be more transparent and tighter.

In summary, data cleansing involves:

- Converting data fields to a common format (a process often shared with the data integration phase).
- Parsing entries to convert unstructured group fields into their individual components
- Identifying and correcting errors
- Eliminating inconsistencies
- Matching records to eliminate duplicates
- Filling in missing values

Data Augmentation and Enhancement

In addition to the data a company collects from its own systems, it frequently needs to augment and enrich this with other data from external sources. Demographics, tax jurisdiction information, geocode data such as longitude / latitude, census tracts, credit information, and lifestyle information are some of the broad classes of data that can enrich consumer records.

Customer data is frequently enhanced with geocode-related data. This intelligence can be used to determine if an insurance applicant is in close proximity to a river or flood zone, or if a telephone customer desiring ADSL service is within the requisite distance of the central switching office.

While much attention has been focused on the value of data augmentation to enrich customer records for marketing purposes such as identifying prospect targets, its applicability extends to many other areas. These can include medical records used to research and discover possible common characteristics in a group of patients with the same disease or even security applications involving the identification of suspected terrorists.

Today, Data Quality is more important than ever

In the past, many organisations didn't truly recognise the need for data quality until it became apparent after an expected high-return company initiative, usually related to customer relationship

management, or an enterprise data warehouse, failed due to poor data quality. Even if the organisation took the additional time and effort now required to successfully resolve its data dilemma, momentum was lost, and confidence in the initiative, as well as the reputation of its sponsors, were often severely impaired.

A forward-thinking organisation should include data quality as a part of its everyday operations. While, clearly, this cannot be accomplished overnight, recent regulatory and security initiatives such as compliance with Solvency II.

In countries other than Romania, there are many strictures related to the accuracy and currency (up-to-date) of data: for example, the U.S. Department of Treasury's Office of Foreign Assets Control (OFAC), Sarbanes-Oxley, the USA Patriot Act, and the Health Insurance Portability and Accountability Act (HIPAA) [Solvency II in Europe], all require a solid data foundation.

Concerns for public safety and individual confidentiality will cause even lagging organisations to recognise that an effective data quality program is rapidly becoming close to a mandatory requirement.

While many think that Solvency II mainly addresses privacy issues and health insurance portability, it also includes reporting requirements that are targeted at catching and eliminating fraud, in part by recognising inappropriate and inconsistent claims. The ability to accurately link medical provider and the supplied procedures and medications will rely on consistent, high-quality data.

The U.S. Office of Foreign Assets Control (OFAC) of the Department of the Treasury administers and enforces economic sanctions against countries and groups of individuals, such as terrorists and narcotics traffickers. OFAC publishes a master list of Specially Designated Nationals (SDNs) and Blocked Persons. U.S. persons and companies are generally prohibited from dealing with any person or organisation on the list and are subject to substantial penalties for violations, if they do.

Sarbanes-Oxley requires that CEOs and CFOs to certify the accuracy of their company's financial statements. However, if the underlying internal controls system relies on poor data, no matter how well-designed the software is, how comfortable are these executives going to feel about certifying the resultant reports?

The Patriot Act requires financial institutions to verify customers' identities and link transactions together to identify, for example, possible money-laundering activities.

These might seem to encompass a suite of rules and regulations which are of no direct concern to Romania, but any organisation with international links is, and will be under increasing pressure to maintain clean and auditable data.

An online marketer that required users to first register in order to view its content and place orders thought that it had 18 million users. The company was about to make a major investment in a high-end operational CRM system when it discovered that many of the registered users had names of cartoon characters or well-known politicians; one of the most common was George Bush of 1600 Pennsylvania Avenue. After eliminating the obvious fictitious entries, the number of valid entries was reduced to approximately 6.5 million. Using de-duplication

Has this ever occurred in your organisation?

Issue: A customer places a call to a company's support centre. While inquiring about her problem, the support centre specialist accesses her records to view her previous support interactions with the company. As the support specialist inquires about the nature of her current problem, she yells into the phone that "this is the same problem I complained about in last week's email and why I do need to explain it to you again?!"

Cause: Although the company thought it had implemented a CRM solution that permitted it to track all of its interactions with a customer or prospect, the customer's email interaction was initiated from her work computer, not from her home computer. While the customer records in the company's database were able to link together her correct name, address, phone numbers, and home e-mail address, her work email address, was not linked to these entries. There were, however, another set of records, containing her work e-mail address, that were assumed to be for another customer with the same name! Simply asking the customer if she had previously reported the problem might have resulted in prompting the support specialist to link her two sets of records together and gone a long way towards appeasing her.

software to eliminate multiple entries for valid users who simply registered again when they forgot their passwords, the number of unique users was further reduced to less than 2.5 million. The existing CRM software could easily handle this volume and still provide sufficient headroom for reasonable future growth. The purchase of the new, and expensive, software was postponed indefinitely.

Data Quality is a Continuous Process and a Way of Corporate Life

While data quality may be imperative, it is definitely not a onetime process. Although in a perfect world, all data would be verified and cleansed as it enters an operational system, we live in a world where things change. If a woman marries and changes her last name, a customer moves to a new address, or one vendor merges with or acquires another, a company should be able to account for these changes rather than assuming any new transactions originate from a new customer.

Data quality should be built into application system design. For example, in the USA, most order entry and shipping systems were originally concerned with making sure that collected addresses were in a format acceptable to the U.S. Postal Service or the appropriate foreign postal services; they were not necessarily designed from the perspective of being able to integrate all customer activity.

If not already established, a strong data management program including well-defined data definitions and associated business rules should be implemented. Internally developed applications should adhere to these rules and packaged software evaluated relative to how well it conforms, or how easily it could be modified to do so. In addition to avoiding the name and address issues mentioned above, a data management program would have prevented the metrics issue associated with the Mars Orbiter problem cited earlier. As the data definitions would include attributes such as that unit-of-measure, it would establish consistency across the enterprise so that when consolidating numbers the company wasn't comparing "apples to oranges" or more likely 'euros to dollars', 'feet to inches', or even 'feet to metres'.

Audits should be conducted to check on data quality. Adding a 'date last updated' field to each master file entry (e.g. employee, customer, product, vendor, organisational structure) helps determine its 'freshness' and could also be used to select the records to be audited. In addition, an audit trail of data transformations should be maintained so that the source data is traceable. One simple way of accomplishing this is to retain the original data values and create new fields for the resultant standardised values.

Components of a Data Quality Solution

Begin at the beginning: A strong data quality solution should begin at the source where data entry errors can be minimised and information content standardised and verified. While standardising name and address fields is somewhat obvious, additional steps should be taken to prevent the same entity (e.g., customer, part, employee, vendor, medical procedure, etc.) from being represented several times in the database, each time with a different key identifier, and thus considered as several different entities.

Verify Assumptions: An investment in data profiling is worth a thousand assumptions.

A good data quality product will provide a wide range of tools and optional extensions. In addition to performing the traditional validation, standardisation, matching, householding, and de-duplication functions, it should be able to run on a variety of hardware platforms and integrate with optional, but nonetheless important, data integration, data profiling and data augmentation components. The suite should be able to operate in both real-time and batch modes. Real-time is frequently appropriate for data collection, while batch may be more suitable for periodic data verifications and audits. It should present a common look-and-feel for all capabilities and allow any address to be verified from a single interface.

Provide an Audit Trail: Even the best software or subject matter expert can produce a reasonable but incorrect result that will need additional investigation. A simple way to accomplish this is to retain the original values while storing "standardised values" in additional fields rather than replacing the original values.

Maintain Data Currency: Recognise that people will change their names and addresses while vendors will also move and merge with other organisations. Additionally, street names and postcodes can change. Consequently, it is necessary to monitor this during subsequent interactions and periodic audits, perhaps based on periodic updates

from third-party data augmentation and enrichment suppliers.

Deploy Good Data Management Techniques: An enterprise-wide data management program to establish consistent data definition formats, units-of-measures, and business rules will go a long way towards preventing data inconsistencies and identifying those that do occur. Data is a corporate asset and should be treated as such.

Data management also involves the selection of key identifiers. For example, while some organisations use a telephone number as the primary key for identifying customers, valuable customer history is often lost when the customer moves and acquires a new phone number.

Be Able to Integrate a Wide Variety of Sources: As most organisations utilise packaged software from multiple vendors, the ability to "capture data at its source" must be extended to include a variety of potential sources. Even if an organisation desired to write its own data integration software, it would have to keep up with the steady stream of modifications and enhancements that the application vendor released each year, in particular those that modify the underlying file structure. Fortunately, data integration software is available to address this issue both at the database and application-specific levels.

Consider Third-Party Data Augmentation Sources: It is highly unlikely that your own systems will yield all the information you would like to obtain. Third-party data augmentation and enrichment vendors can be a valuable source for obtaining additional customer or vendor-centric data. This can include telephone numbers, geographic overlays, consumer demographics and lifestyle information.

Provide a Variety of Ways to Access Data Quality Software: Ideally, a well designed data quality software offering should be accessible as needed. The technology should be offered both as a hosted service and a licensed offering. Additionally, it should be executable in both batch and real-time implementations. While most tools offer their own application programming interfaces in order to be callable from custom applications, there should be a current ability or short-term plan to also make the functionality available as a web service.

Work with Experienced Data Quality Vendors: Just as callers to a vendor's support centre don't want to be told that the cause of the problem rests with some other vendor and they must now contact the other vendor themselves for problem resolution; companies seeking to implement a data quality solution want to deal with an experienced company. Proven vendors are able to take responsibility for the complete solution as they offer a comprehensive product line augmented by strong partnerships to fill in the missing pieces.

Summary

Data quality is at the foundation of almost all organisational processes, both operational and analytical. The quality of the data used by an organisation will have a major impact on the organisation's overall effectiveness and efficiency. While data quality has sometimes been ignored until a major, and usually highly visible initiative gets underway, many organisations now recognise that data quality should be a part of its everyday operations and is certainly not limited to customer-centric initiatives. These organisations also realise that data quality is not a one-time exercise, rather, it is an on-going process that must be continually maintained.

Fortunately, as high data quality becomes an established organisational best practice, maintaining quality becomes relatively easier. Effective data quality not only includes data cleansing but also data integration, consistent business rules, and a strong dose of common sense. Success involves not only software tools and solid data management practices, but, more importantly, a commitment to make data quality a number one priority.

SMARTaddress

Address Cleansing &/or Geocoding

SMARTaddress® has been specifically designed to cleanse and geocode Romanian addresses. In line with best practice, it is based on data inputs from multiple sources which also include Poşta Română.

It is based on a sophisticated parser which can separate unstructured data into the separate components (counties, localities, street type, street name, street type, street name, and numbers).

The address data is then compared with a most comprehensive, standardised database for Romania and the output is available with or without diacritics and with postcodes.

Additionally, SMARTaddress incorporates a geocoder which provides coordinates for all addresses in Romania.

Pitney Bowes
Business Insight

Data Integration

Sagent® Dataflow from Pitney Bowes is a powerful and flexible integration engine that brings together data from heterogeneous sources into a single view where you use a comprehensive set transformation tools to merge and cleanse the non -address components of data for optimal value.

Once your data is transformed, apply Sagent Data Flow Solution analysis tools to create meaningful information and useful reports. You gain better understanding of the critical data in your business and your users are empowered to make smart decisions based on real data.

Its interface reduces the learning curve so that end users can master the solution more quickly than other BI tools. A powerful visual development environment helps you create sophisticated data transformations quickly and simply.

About Geo Strategies

Geo Strategies was founded in 1993 and has become the leading supplier of geo-spatial information and tools for Romania.

Their core expertise is built around geo-spatial data products, analytical and modelling tools, consumer segmentation, bespoke data services, consultancy, training, and project management.

This expertise has been developed according to internationally recognised best practices and methodologies, to which significant innovation / value is being added to make it relevant and actionable in the local context.

Geo Strategies partner with Pitney Bowes, Experian and Navteq (part of the Nokia group) to provide the best-in-class products for data quality management (data cleansing, manipulation and integration), socio-demographic profiling and targeting and world-standard mapping for GIS and other spatial applications.

From services in data cleaning and enhancement, data integration to Mosaic consumer segmentation, Micromarketer area analysis and profiling, to the application of customer driven insight and targeting - we can assist.

Contact us at info@geo-strategies.com or call us (+44 1223 205080 or +40 722 940) to discuss the steps for your data cleansing requirements.

To find out more about what Geo Strategies could do for your business in Romania go to www.geo-address.com

Geo Strategies