Measuring and Monitoring the Quality of Master Data

By Thomas Ravn and Martin Høedholt, November 2008

Introduction

We've all heard about the importance of data quality in our IT-systems and how the data that flows through our applications is the fuel of the business processes. Yet, surprisingly few organizations have a structured approach to measuring the quality of their data. Some may have a few custom reports or manually compiled Excel sheets that show a few aspects of data quality, but if information is truly an enterprise asset, shouldn't we be measuring and monitoring it like we do with all the other assets of the organization? Like most other things, data quality can only be managed properly if it is measured and monitored.

The main reasons for implementing a data quality monitoring concept are to ensure that you identify:

- Trends in data quality,
- Data quality issues before they impact critical business processes
- Areas where process improvements are needed

In this article we present a structured and methodological approach to measuring and monitoring the quality of data. Doing this should be part of a larger master data management or information management strategy, but in this article the focus is specifically on data quality monitoring.

We start off with a description of the dimensions of data quality, provide some insight in how to define good data quality KPIs (Key Performance Indicators) and finish with some thoughts on the process of monitoring data quality.

The Dimensions of Data Quality

Looking at the offerings of many data profiling and data quality tool vendors, one could be led to think that data quality is just about checking if all fields have a value and if it's valid compared to the domain that is defined for that specific field. While these are certainly relevant aspects of data quality, they only represent two out of the six common dimensions of data quality.

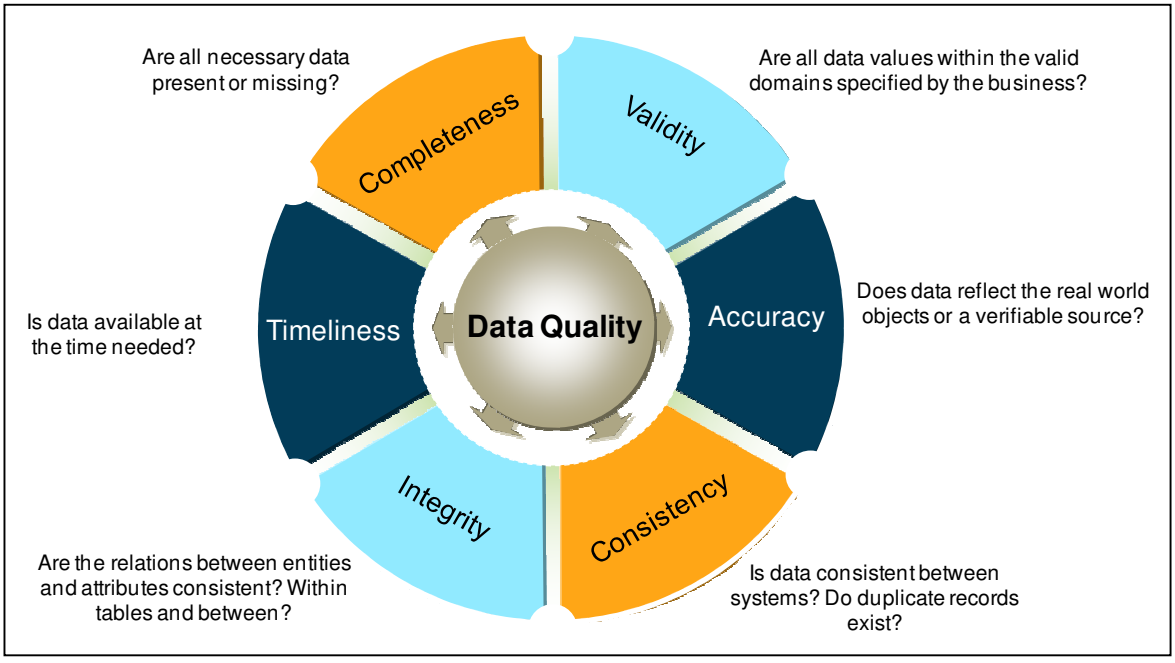

Figure 1 - Data Quality dimensions

When defining a data quality measuring concept, you should aim to focus on the dimensions that are meaningful and relevant for the business without spending too many resources, i.e. you need to justify the business relevance of what you measure. On the other hand measuring all the different dimensions of data quality gives you the most complete picture. What is it for instance worth to have a valid address for a customer if it is not correct because the customer has moved?

A challenge that organizations face as they attempt to define data quality KPIs is that completeness, validity and integrity may be relatively easy to measure, while measuring consistency, accuracy and timeliness is a whole other story. In general we would argue that the below illustrated relationship exists between the difficulty of measurement and the business impact.

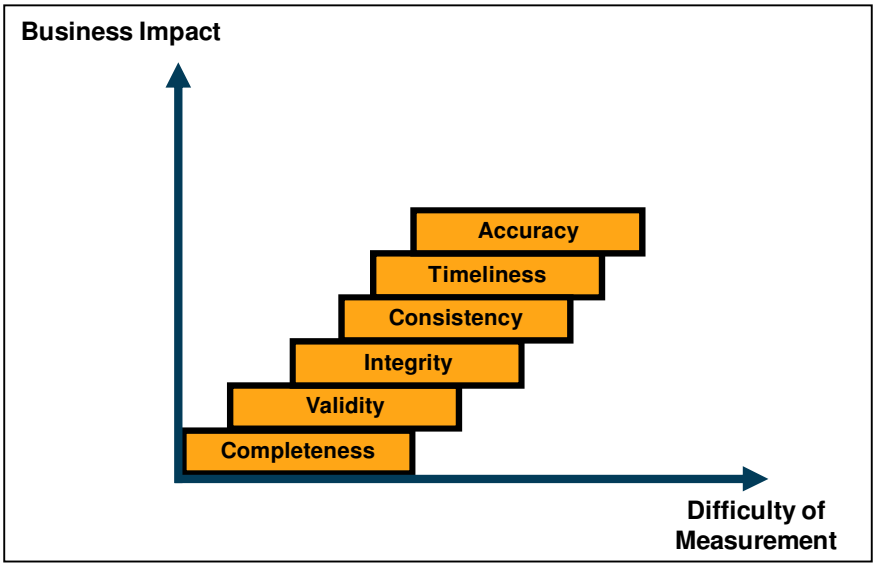

Figure 2 - The Difficulty of Measurement

Since accuracy, timeliness and consistency are the more complex dimensions to measure, here are a few hints on how to address them.

Accuracy:

The options to check accuracy are either manual sampling or verification against a trusted external source. For name and address data there are several external sources available such as D&B (Dun & Bradstreet). Another common form of external data verification is bank account verification that is offered by several companies. This is a good way to ensure data as critical as bank account information from a vendor is correct before you make any payments.

As part of the patriot act, US companies are restricted or prohibited from exporting or providing services of any kind to any party listed on any of a number of restricted party lists maintained by the U.S. Department of Commerce, Department of Treasury, U.S. Customs Service, Defense Department and others. This goes beyond the classic view of ensuring accuracy of data, but it's really about ensuring that the organizations you do business with are really who you think they are. There are several companies (e.g. JP Morgan Vastera and Management Dynamics) that can help you screen your new business partners.

Timeliness:

A good approach here is to measure the process from the request of a new master data element to it is available for use. This is especially relevant for large complex master data objects such as a new customer or product, where multiple different people in different departments are typically involved in the data creation process. Depending on your information management architecture it could also be relevant to measure the time from data is entered in an operational system to it is available for reporting in your business intelligence environment.

Consistency:

The aspect of consistency can overall be split into duplicate records and cross system consis-

tency. Duplicate records are a very common problem and have two typical sources:

a) They are created by mistake, simply because the user was not aware that the record existed already or
b) Duplication due to system limitations. Common problems include systems that cannot store different payment terms or different currencies for the same vendor, and then the same vendor is created multiple times as a work around

In both cases the duplicate records for the same customer of vendor will result in incorrect reporting and affect the business directly. You should ideally try to catch duplicate records at the point of data entry, but as it is difficult to completely avoid the creation of duplicate records, you should also define KPIs that capture and present potential duplication issues. This way you can handle and at least map duplicates to the master record in order to deliver trustable reporting.

wise the business will experience inconsistency between the systems. A cross system consistency KPI could either represent entire records or individual fields.

Defining good KPIs

Building a data quality monitoring concept involves the following four basic steps:

1. Define master data objects of importance (e.g. customer data)
2. For each master data object, define a set of data quality KPIs
3. For each KPI, define measure details
4. Define procedures for follow-up on data quality issues

Basic column analysis (number of blanks, max, min, uniqueness, etc.) and integrity analysis is a great place to start when defining data quality KPIs, but in order to give a more complete picture you should also collect input from the sources such as business intelligence and ask about data quality pain points. We recommend gathering candidate measures liked illustrated here in figure 3:

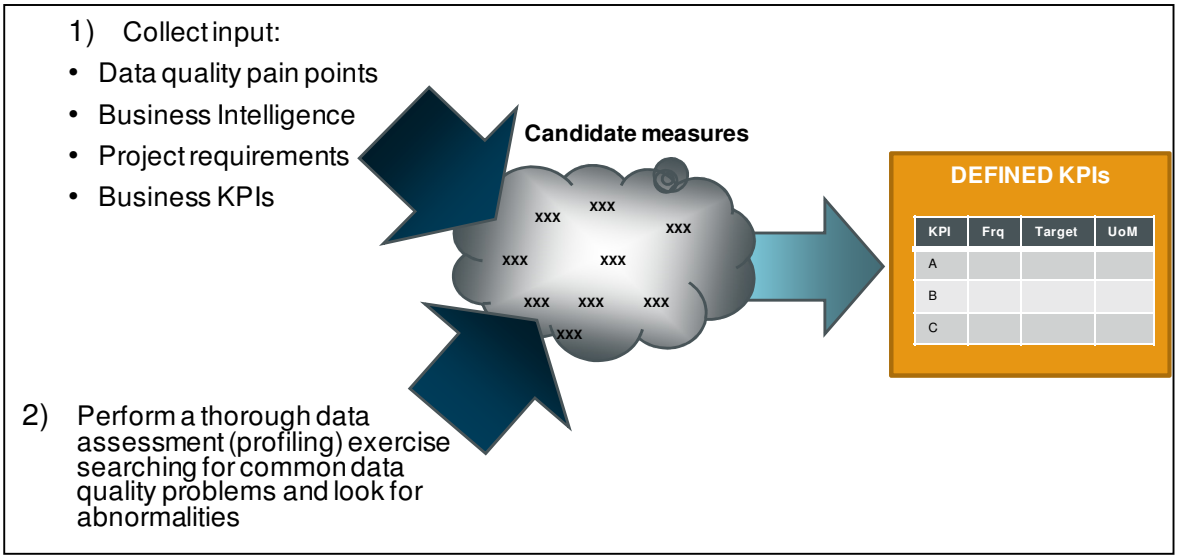

Figure 3 – Identifying KPIs

Consistency is typically focused on ensuring that data that is synchronizing across systems is consistent. If somehow the synchronization fails between two systems it's very important to get the error identified and correct the error, other-

Every output from the data assessment exercise is NOT a KPI. But there are likely a couple of measures that you analyze during the assessment that you would like to monitor on a continuous basis. These are selected to be KPIs.

Keep in mind that data quality KPIs should express the important characteristics of data quality of a particular area of a data object.

Measures are typically percentages, ratios or number of occurrences, and here is an important point: For consistency reasons, aim to harmonize the measures. If for instance one measure is "Number of customers without a postal code" while another is "Percentage of customers with a valid VAT-number" a list of KPIs will look confusing, since the first KPI should be as low as possible, and the other as close to 100 as possible.

A good approach is to have all data quality measures as percentages, with a 100% indicating the highest level of quality. This way a data quality report can easily be reviewed by a manager who wants to quickly review the quality of a particular data object covered by several KPIs.

Be careful not to define too many KPIs, as this will just make the organizational implementation more difficult. If someone is presented with a report containing 50 or 100 data quality KPIs about customer data they are very unlikely to look at each one. Choose a few good and relevant KPIs to start with and then you can always add more later on.

Also pay attention to controlling fields as they are very important in order to reflect the correct measure. Controlling fields are fields used to segment records into smaller groups. To avoid KPI's looking at all records in a table it's necessary to take the controlling fields into consideration as they will for instance decide if a field should be filled or not. Common SAP examples include the country code field that defines if a value in the state field is required, or a material type that defines if a weight is required for a material record.

Indirect measures

In some cases there might be critical fields (e.g. MRP type or weight for a product) where the correct value is of utmost importance, but at the same time, it's close to impossible to define the rules to check if a new value entered is correct. In these cases, one approach is to measure indirectly by for instance reporting what users have changed these values for which products over the last 24 hours, week or whatever is appropriate in your organization.

Prevention

An interesting point is that in a packaged solution like SAP ECC, a lot of the things like completeness and validity should be enforced by the system at the time of data entry, and if it isn't you should consider implementing a data input validation rule rather than allowing bad data to be entered and then measure it!

However, there are cases, where the business logic of a field is too ambiguous to be enforced by a simple input validation rule. In addition data may enter you SAP system through interfaces where some input validation could be ignored. Another common source of data quality issues is data migration efforts, where data is migrated without sufficient validation and verification prior to loading.

Documenting KPIs:

It is essential that KPIs are properly documented, as this necessary both to implement them and to be able to follow up. So we recommend that you for each defined KPI document the following (see figure 4):

KPI Name:	A meaningful name of the KPI that express what is being measured.
Objective:	Why do you measure this? What business processes are impacted if there data is not ok?
Dimensions:	What data quality dimensions (integrity, validity, etc.) are this KPI related to?
Frequency of measure:	How often do you wish to report on this KPI? Daily, weekly or monthly?
Unit of measure:	What is the unit of the KPI? Number of records, pct of records, number of bad values, etc.?
Lowest acceptable measure:	Threshold that indicates if the data quality aspect the KPI represents is at a minimal acceptable level. The value here must be in the unit of measure of the KPI.
Target value:	At what value is the KPI considered to represent data quality at a high level?
Responsible:	The person responsible for the particular KPI.
Formula:	The tables and fields that are used to analyze and calculate the KPI. This is the functional design formula that forms the basis for the technical implementation.
Hierarchies:	When reporting on a KPI it is very useful to be able to slice and dice the measure according to different dimensions or hierarchies. For a customer data KPI for instance, good hierarchies would be regions, country, company code and account group. Being able to view the KPI through a hierarchy also makes it easier to follow up with specific groups of business users.
Notes and assumptions:	If certain assumptions are made about the KPI make sure to document it.

Building and Presenting KPIs

Building the technical architecture to analyze and present KPIs can be done in various ways. Essentially you need to extract your data from your SAP ERP system (or whatever system you data are in), analyze the data and calculate KPIs and finally present it to the business community. The extract and analysis can be done using an ETL tool, a data quality tool and/or a data profiling tool. But if some cases you can also reach quite far using simple SQL. Below in figure 4 is a simplified illustration of a typical architecture.

Presenting the KPIs to your organization in the right way is critical. The reports need to be easy to get to and understand for the business users,

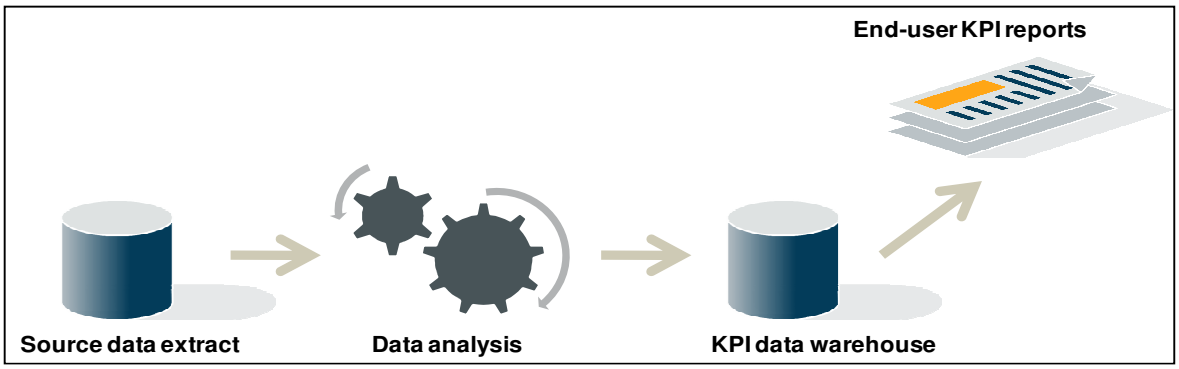

Figure 4 – KPI Concept Architecture

and therefore it's very often a help if you can leverage already existing reporting tools that business users are accustomed to. In an SAP shop, end users are probably used to getting reports from SAP BI, so building on this familiarity is strongly recommended. By presenting data quality reports in the same framework as other business reports, you also send the message, that data quality is an important part of managing the business. We have had great success building complete data quality monitoring solutions this way.

Another approach is to use a "Packaged data quality monitoring solution" such as Data Insight XI from Business Objects or DataDialysis ® from Back Office Associates. Both of these tools offer a rapid implementation of a data quality monitoring solution, yet for more complex KPIs you'll probably need some custom ETL/SQL.

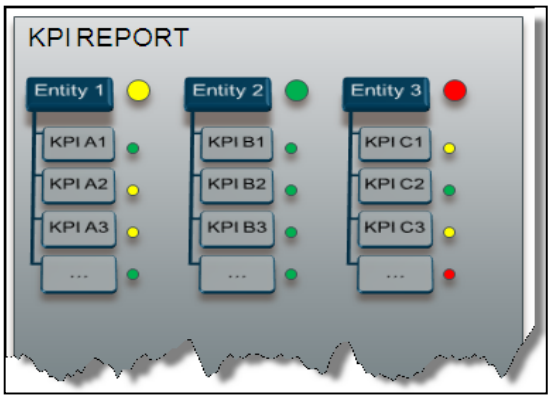

Figure 5 – Sample KPI Report

A word about process and governance

For a data quality monitoring effort to be effective, clearly defined targets for data quality and follow up mechanisms are required. Making sure there is a robust process for monitoring data quality and following up on issues is important. Having the business feel a sense of ownership and responsibility is critical.

As this article is not about how to define and implement a leading practice data governance organization, we'll stick to the process for now.

An example of a simple but effective process for monitoring and addressing issues related to data quality is an iterative process is illustrated in figure 6.

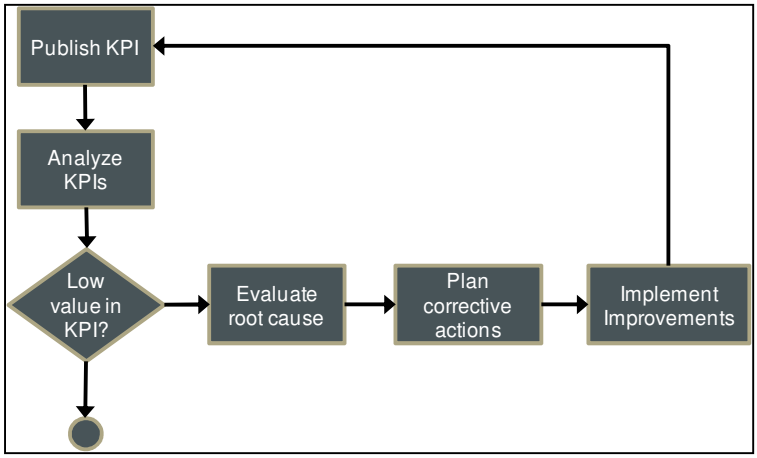

Figure 6 - KPI process

Starting with Publish KPI, the first three boxes are rather straight forward, but we want add some comments to *Evaluate root cause* and *Plan corrective actions.*

Evaluate root Cause

It is of key importance that the root cause of the problem is identified. We need to identify the originating process and find out where in the process is the problem created and why? In this step the focus should be on the process rather than on the person. You don't want to create a blame culture!

Common root causes include:

- Unclear definitions for how to enter data
- Lack of training for data maintenance employees
- Data is entered by the wrong person (someone without the required knowledge or someone without a sense of ownership)
- Interface problems (causing inconsistent data)
- Workload too big for some individuals
- Speed of data entry is prioritized over quality of data entered

Plan corrective Actions

Prepare the steps required to fix the data quality issue. And remember there are two aspects to this:

a) Fix the wrong data, and
b) Fix the bad process to prevent the problem from reoccurring

In some cases changes to security setup can also be part of the solution.

Common improvements include: more training for data maintenance employees, more automated validations, additional approval steps, and change of person that enters the data.

In planning of corrective actions, the impact of the data quality issue must be assessed as this is a key parameter in deciding the type of preventive actions that are warranted.

About Platon

Platon is a leading global independent consulting company, specializing exclusively in Information Management, which covers: Business Intelligence, Performance Management, Data Warehousing, Master Data Management, Information Life Cycle Management and Data Integration.

Platon does not sell hardware or software, but concentrates solely on consulting, project management, solution implementation, support and training. Over the last 10 years Platon have successfully solved the toughest Information Management related challenges in over 300 organizations worldwide.

The Platon Group has offices in Denmark, Sweden, Norway, Finland, Iceland, United Kingdom, USA and Australia, and currently has over 200 competent and highly qualified employees. For more information please visit www.platon.net.