

Data Quality Initiatives

Ivan Schotsmans

Global Director Events IAIDQ





Agenda

- Set the scene
- How to start?
- Formatting & Standards
- Data Corections
- Data Governance
- Lessons learned
- Supporting websites



Quality is never an accident; it is always the result of intelligent effort.
John Ruskin (19th century writer, art and social critic)

SET THE SCENE



A few stories

- 1999, Mars Orbiter
 - NASA lost a \$125 million Mars orbiter because a Lockheed Martin engineering team used English units of measurement while the agency's team used the more conventional metric system for a key spacecraft operation
- 2006, Tax blunder undermines Belgian federal budget
 - "calculation error" of 883M Euro
- 2009, Belgacom & WWF
 - a dispute for hacking telephone system (7173 Euro)





A continuous program is necessary to improve data quality.

- Data Quality, What?
 - Not understood
 - Either underestimated or overestimated
 - How to measure?
 - Who is responsible?





The added value of data quality initiatives...

- ***Overly ambitious data-quality professionals get shot down because trusted data will not directly increase revenue, reduce costs or improve operational efficiencies, key criteria for getting funding approved, Forrester Research says.***





... requires an excellent business case.

- Difficult to convince senior management
- Too much downstream efforts
- Lack of data governance
- Maturity of the application



“Information quality is a journey, not a destination.” - Larry P. English

HOW TO START?



Budget & Sponsorship

- Set-up a data quality initiative:
 - Clearly define scope
 - Limit to one (small) project or a few data objects
 - Use critical data objects for the organization
 - Make some noise
 - Go for buy in from business and IT
 - Deliver a result in maximum one month
 - Find a budget/sponsor





Define the cost of non-quality

- Immediate cost of non-quality data
 - Scrap and rework
 - Manual corrections
- Cost of Information quality assessment
 - No trust in data results into efforts/costs necessary to invest in data errors
- Cost of process improvement to fix broken business processes
 - Each broken business process needs to be fixed to avoid unnecessary costs





Use a workshop approach

- Workshop with focus groups
 - Assessment
 - Rate the quality of information you receive from other departments
 - Rate the quality of information you provide to other departments
- Result
 - Should be equal
 - You perceive that information which is correct for you, is also correct for others





Documentation Review

- Review
 - Data definitions
 - Data standards
 - Database schemas
 - Written policies and procedures
 - Written business rules
 - Validation criteria





Measure, Measure, Measure, ...

- Make use of metadata
 - Create a metadata repository
- Develop statistics on data distribution
- Measure data quality in your ETL process
- Share your data quality metrics



What to measure?

- You need a method to measure data quality
 - Keep it simple, but do it systematically
- Use 'CACTUS'
 - Completeness
 - Whether or not all the data necessary meets current and future business information demand
 - Accuracy
 - the quality of being near to the true value
 - Correctness
 - Timeliness
 - Timely convenience
 - Uniqueness
 - Securitiveness
- Other measures exist, use those important for your industry





2 techniques

- Pareto Diagram
- Fishbone





Pareto Diagram

Pareto analysis is often referred to as the 80/20 Rule

- 80% of the total problems incurred are caused by 20% of the problem cause types
- It's a simple tool for helping us to prioritise the entire root-cause analysis process by focusing work on the most critical areas with higher impact.





Fishbone

The fishbone diagram identifies many possible causes for an effect or problem.

- Identify all causes and the root causes for a specific defect
- Analyze and relate some of the interactions among the factors affecting a specific process or effect
- Enable corrective action



Quality is free. What's expensive is finding out how to do it right the first time.”
Philip Crosby

FORMATTING & STANDARDS



What are we looking for?

- Metadata compliance
 - Is our data aligned with our metadata definitions?
 - Is our data aligned with expected patterns/formats?
- Content discovery
 - Accurate data
 - Complete data
 - Correct Data
- Look for relationships
 - Check for redundant data





Data Profiling

- Identify unique records
- Look for duplicates/inconsistencies
- Look for data patterns
- Frequency counts
- Blanks/Nulls/Low Values/High Values





Audit your profile result

- Establish Table Statistics
 - Total Size in bytes including Indexes
 - When was it last refreshed
 - Is referential Integrity applied
- Establish Row Statistics
 - How many rows
 - How many Columns / Table
- Establish Column Statistics
 - If a Key value, how many duplicates
 - How Many Unique Values
 - How many Null Values
 - How Many Values outside defined domain





Data Sampling

- A sample should have the same characteristics as the population it is representing
- Look for accurate data samples
 - Several techniques
 - Simple random sampling
 - Systematic sampling
 - Sampling with probability proportional to size
 - Stratified sampling
 - Cluster sampling
 - Multi-stage sampling
 - Multi-phase sampling
 - Use the internet as support to define the right sample





Definition of standards

- What is the data is
 - not HOW, WHERE or WHEN used or WHO uses
- Add meaning to name
- One interpretation
 - No multiple purpose phrases
 - No unfamiliar technical program
 - No abbreviations
 - No acronyms





Name Standards

- Comply with data element format
- Single concept, clear, accurate and self explanatory
- According to functional requirements not physical considerations
- Upper and lower case alphabetic characters, hyphens (-) and spaces ()
- No abbreviations or acronyms





Introduce a data standardization process

- Proposal Package – Data Model, Descriptive Information, Organization Information, Integration Information, Tool Specific Information
- Technical Review – Model Compliance, Metadata Complete and accurate
- Functional Review – Integration with Enterprise Model
- Issue Resolution – Maintenance and Management
- Total Process < 30 days
- All based on an integrated web accessible application. Results integrated to the Enterprise Metadata Repository.



DATA CORRECTIONS



It's Time To Invest In Upstream Data Quality!

- **A recent Forrester paper titled *It's Time To Invest In Upstream Data Quality* suggests that when companies *realize short-term data cleanup ROI immediately, it's hard to justify front-end investments that may take years.***





Downstream vs Upstream data

- Downstream data, data flow in line with the data process
 - Corrections at the end of the data flow are common use
 - Fixes in different applications
 - Put's the pressure on IT
- Upstream data, the direction opposite the data flow
 - Prevention at the source is one of the founding principles of good data quality management
 - Embed data quality in your business process
 - Responsibility at the business site



Eliminate the short-term practice of data cleansing

- Scrap and Rework
 - One of the biggest cost
 - Delivers quick fixes
 - No prevention of issues in the end-to-end information chain
 - Ignore the root-cause analysis of a particular issue



DATA GOVERNANCE



Data Governance Definition

- Data Governance means "the exercise of decision-making and authority for data-related matters."
 - A system of decision rights and accountabilities for information-related processes
 - Execution according to agreed-upon models which describe
 - Who can take action
 - What actions with what information
 - When
 - Under what circumstances
 - Using what methods
- When people refer to Data Governance, they might be talking about
 - organizational bodies
 - Rules
 - decision rights (how we "decide how to decide")
 - accountabilities, or monitoring, controls, and other enforcement methods





The challenge to set-up a data governance organisation

- Where to start?
- Who should be involved?
- Set the right scope to start
- Regulatory drivers





Data Migration and Data Governance

- Chicken and the egg situation
 - Data governance is needed for a successful Data Migration program / project
 - Data governance is long term view
 - Data Migration initiatives mostly have a short/midterm deliverable





Develop a Data Quality Culture

- Set-up a key-team with business and IT
- Set-up regularly meetings
- Try to link data ownership to processes



Chapter 2

LESSONS LEARNED



Bridge the gap between business and IT

- IT can start a Data Quality initiative but:
 - IT must be temporary owner, business should take over on the short term
 - Data definitions are owned/defined by the business, IT can be facilitator
 - Business should be owner
- A formal process to evaluate the data quality
 - Involvement business and IT is necessary





Link data objects to processes

- A data





Think big, start small

- Don't try to boil to ocean:
 - Use step-by-step approach
 - Start with 5-10 data objects
 - What do you want to achieve?

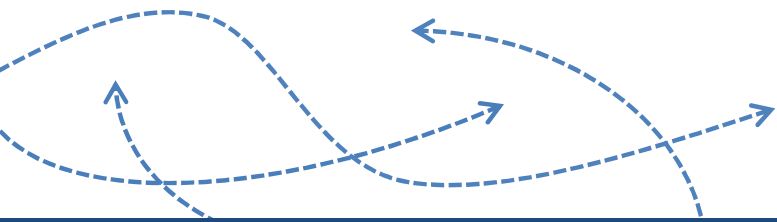




Metadata

- End-to-end view on data elements
 - Business metadata
 - Technical metadata
 - Operational metadata
- Impact analysis
 - Data lineage is a good first step (ICT initiative)



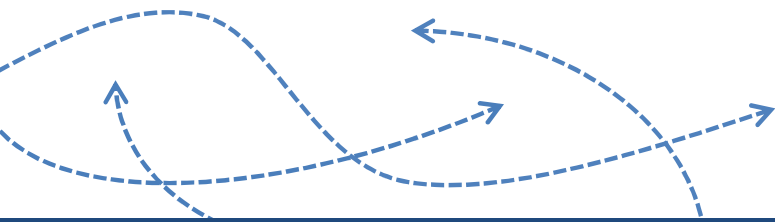


- Company's strategy
- Data Quality objectives
- Measure
- This should be obvious, but it is not:
- Make an effort to **know exactly what you want to say**
- Write down your story as an “elevator pitch”:
- Something you can get across in 1 minute or less
- **KISS – Keep it Simple, Stupid**
- Make it as simple as possible
- Make sure it **makes sense by testing it with colleagues**
- Organize



Chapter 2

REFERENCES

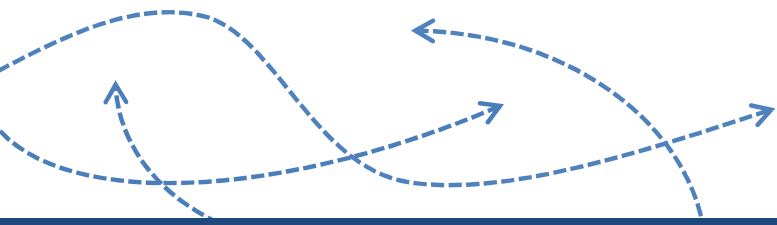


- **Mistakes to Avoid in Data Quality**

Similar to best practices, there are many mistakes to avoid. Among the common mistakes are:

- Saying that data quality is "everybody's responsibility" (it's the same as saying nobody is responsible)
- Treating data quality as an IT problem
- Lack of a business case for data quality efforts
- Seeking the easy fix or the "silver bullet"
- Reactive data quality management using a repair-only approach
- Lack of data quality standards--especially for master data and shared reference data
- Absence of measurement or lack of targets that give the measures context
- Lack of expertise, both business and technical
- Believing that data knowledge or data modeling skills are an acceptable proxy for data quality skills
- Insufficient measures--counting defects but failing to quantify cost or impact





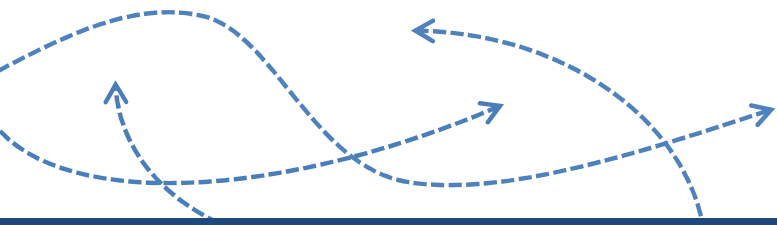
IAIDQ



Ivan Schotsmans

Managing Director of the BI-Community.

BI-community.org stands for a network of professionals bridging the gap between business, technology and university.



- Make sure your data are controlled in a proper manner
- **Step 1: Acknowledge that there is a problem**

