White Paper

Why is Data Quality so Hard to Achieve?

By Robert Grant Beauchamp

Contents

White Paper

Data Quality:

By Robert Grant Beauchamp

Introduction

Was there ever a golden age when data was always accurate and reliable? Or has the pace of change in the world of data always outpaced the ability to manage the innovations?

There is constant innovation in the world of data, making it difficult to keep up with the pace of change. Data, however, is also a very mature science. Business computing came of age in the 1960's. Three, if not four generations of data professionals have labored in this field, leveraging their intelligence and contributing vast amounts of energy. Yet, of all the problems solved, and challenges overcome, there is one thorn that has yet to be removed. That perennial irritant is data quality. As long as there has been data, it seems, the quality of the data being acquired, stored and utilized has been a source of complaint.

Why is this? In this titanic struggle between order and chaos, the tools available to manage data are numerous and accessible. The talent available to use the tools is legion. The field of methodologies and philosophies to guide the talent is rich and fertile.

So why is there a constant lament about the quality of data? Why is getting data quality right so hard for most organizations? With all the resources available it shouldn't be that hard, right?

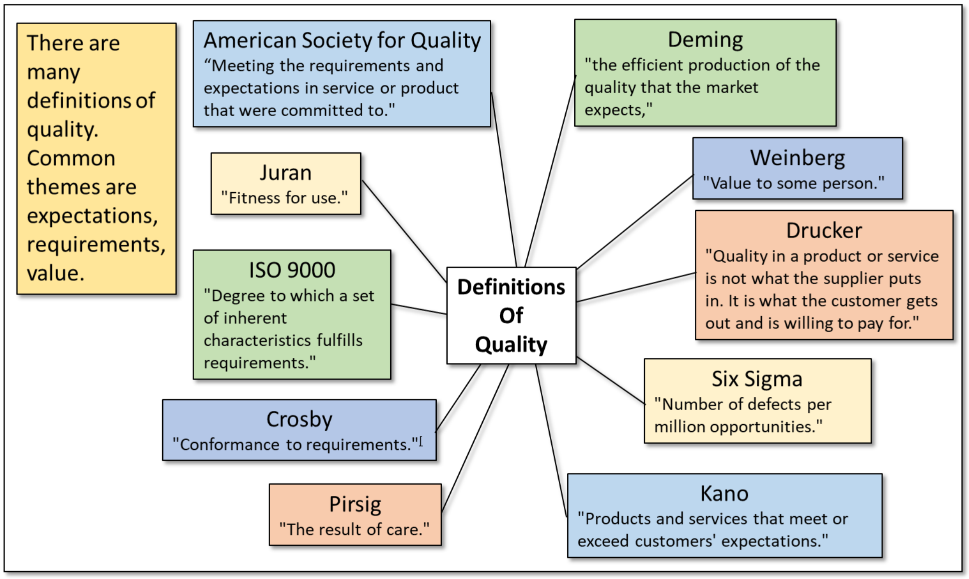
Purpose

The purpose of this document is to provide:

- A brief overview of the concepts of quality and data quality.
- An understanding of the forces working against data quality.
- A different way of looking at data quality that may help improve the quality of an organization's data.

What is Quality? What is Data Quality?

The great names of the philosophy of quality have provided any number of definitions to help answer the question ‘What is quality.’ From Deming, to Drucker, to Crosby, the definitions vary but they usually include some combination of expectations, requirements, consistency and value.



There are many definitions of quality. Common themes are expectations, requirements, value.

Definitions Of Quality

American Society for Quality
“Meeting the requirements and expectations in service or product that were committed to."

Deming
"the efficient production of the quality that the market expects,"

Juran
"Fitness for use."

Weinberg
"Value to some person."

Drucker
"Quality in a product or service is not what the supplier puts in. It is what the customer gets out and is willing to pay for."

ISO 9000
"Degree to which a set of inherent characteristics fulfills requirements."

Six Sigma
"Number of defects per million opportunities."

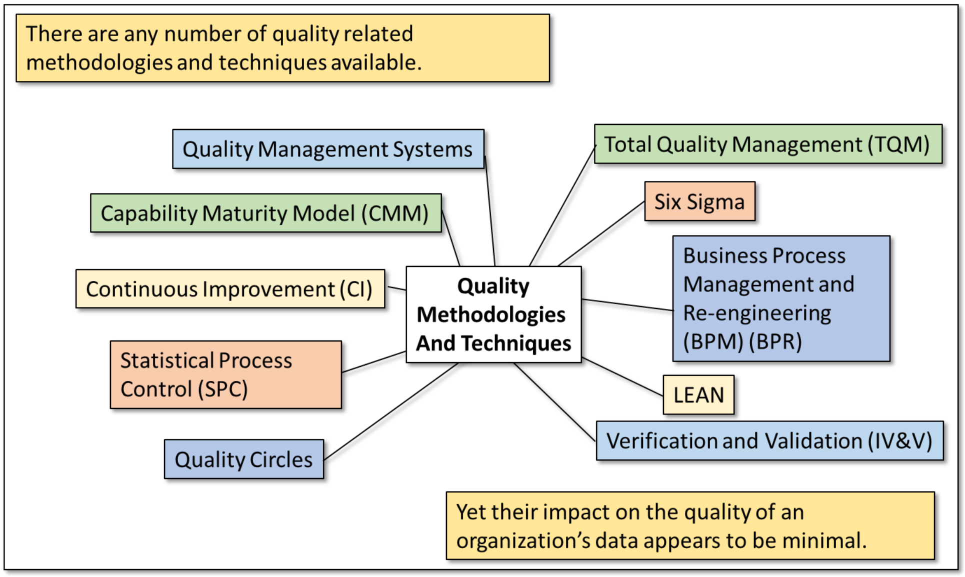
Crosby
"Conformance to requirements."[i]

Pirsig
"The result of care."

Kano
"Products and services that meet or exceed customers' expectations."

Quality methodologies

The more there is at stake, the more important quality becomes. An entire industry has grown up around improving quality. From TQM to Six Sigma and Lean, these methodologies are well documented. Excellent training is also available. In addition, many professionals have had training or experience with these methodologies even if they are not currently using them.



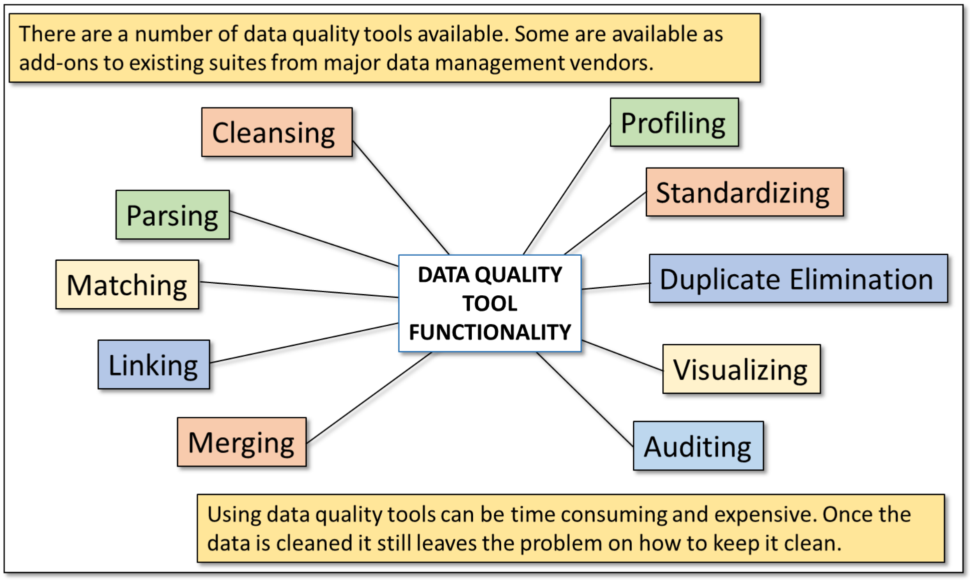
There are any number of quality related methodologies and techniques available.

Quality Methodologies And Techniques
Quality Management Systems
Total Quality Management (TQM)
Capability Maturity Model (CMM)
Six Sigma
Business Process Management and Re-engineering (BPM) (BPR)
Continuous Improvement (CI)
Statistical Process Control (SPC)
LEAN
Verification and Validation (IV&V)
Quality Circles

Yet their impact on the quality of an organization’s data appears to be minimal.

Data Quality Tools

Tools specifically designed to improve data quality are numerous. Many of these tools are included with the major databases, organizations already own. The capabilities of these tools range from sophisticated statistical analysis to simply identifying redundant records.

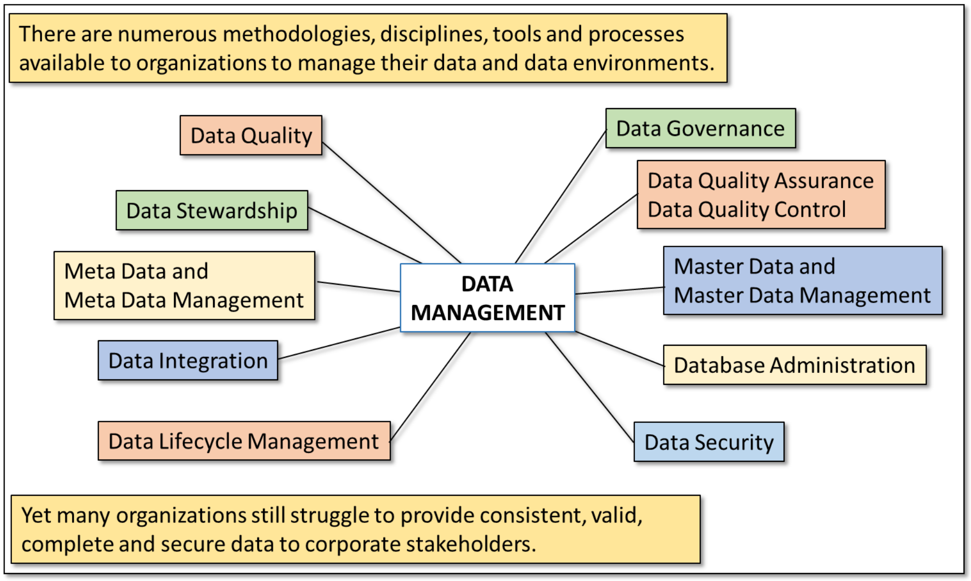


There are a number of data quality tools available. Some are available as add-ons to existing suites from major data management vendors.

Using data quality tools can be time consuming and expensive. Once the data is cleaned it still leaves the problem on how to keep it clean.

Data Management

It doesn't take long to understand that managing data requires discipline. As data-related functionality increased, new methodologies were developed. As with quality methodologies, data management methodologies, disciplines and tools are mature and readily available.



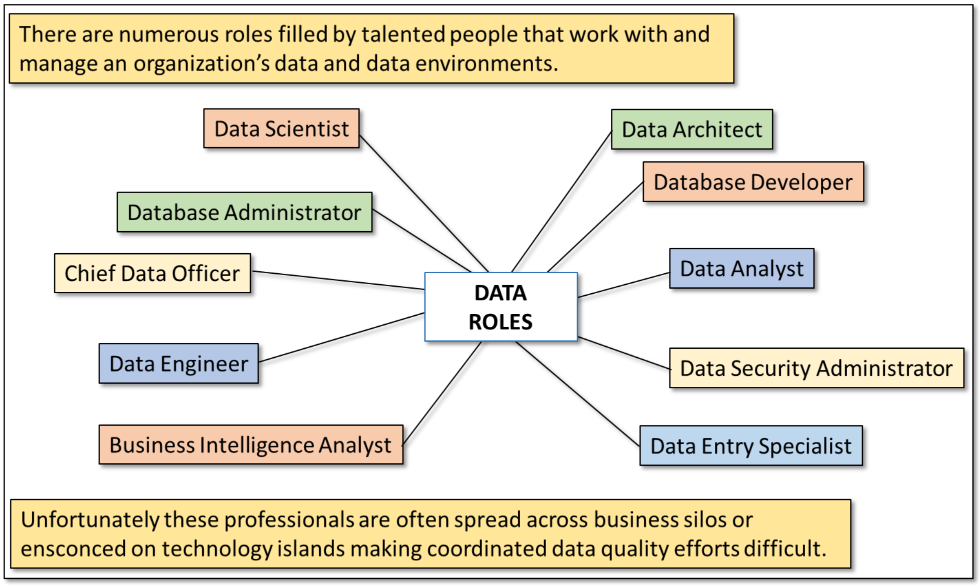
There are numerous methodologies, disciplines, tools and processes available to organizations to manage their data and data environments.

Data Quality
Data Stewardship
Meta Data and Meta Data Management
Data Integration
Data Lifecycle Management
DATA MANAGEMENT
Data Governance
Data Quality Assurance Data Quality Control
Master Data and Master Data Management
Database Administration
Data Security

Yet many organizations still struggle to provide consistent, valid, complete and secure data to corporate stakeholders.

Data Professionals

New and complex data technologies require increased specialization. Data professionals, while often in short supply, still possess a wide range of talent and skills that can be applied to the problem of data quality.



There are numerous roles filled by talented people that work with and manage an organization's data and data environments.

Unfortunately these professionals are often spread across business silos or ensconced on technology islands making coordinated data quality efforts difficult.

Despite the availability of all these resources, the data quality at most organizations is nowhere near the desired level.

The only conclusion to be drawn is that there must be powerful forces at work that frustrate the efforts to generate, maintain and use quality data.

Business Structure makes Data Quality Hard

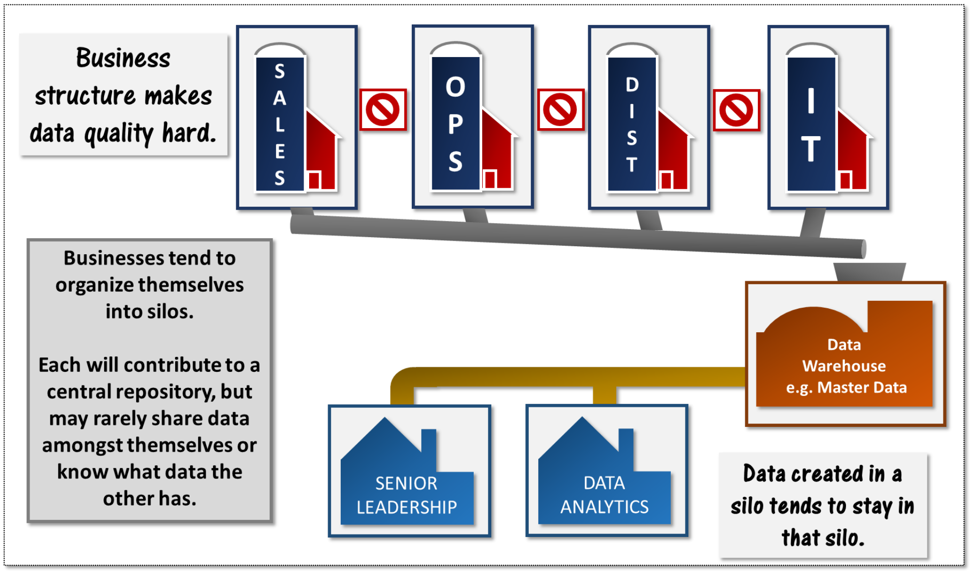
Despite the dependence of business on data, organizations seem to go out of their way to make data quality difficult to achieve. The major causes are familiar to most and notorious for interfering with the success of all kinds of business initiatives, not just data quality. They are: business silos and competing priorities.

Business Silos

Organizations tend to structure themselves around business functions or silos. Business silos are universally reviled and sighted as the root cause of many problems that make it difficult to run an organization efficiently. Yet they persist.

Business silos persist for one very important reason. They work. Silos reduce dependency on others and increase control over resources and processes.

The executive at the top of the silo naturally wants to control as many variables that impact their success as possible. Managers want those resources reporting to them, so they can direct those resources most efficiently. In addition, managers don't want to be dependent on another entity's ability to manage a hand-off. And they certainly don't want to be responsible to another entity for something that distracts attention and diverts resources away from whatever success criteria by which they are measured.


Business structure makes data quality hard.
SALES
OPS
DIST
IT
Data Warehouse e.g. Master Data
SENIOR LEADERSHIP
DATA ANALYTICS

Businesses tend to organize themselves into silos.

Each will contribute to a central repository, but may rarely share data amongst themselves or know what data the other has.

Data created in a silo tends to stay in that silo.

Business Silos (continued)

Whatever efficiencies are gained by silos for business processes, resource management or political survival, they are bad news for the overall data quality of an enterprise.

Data created in a silo tends to stay within the silo. A silo may be responsible for contributing data to an enterprise data warehouse, but this still leaves great quantities of data behind locked doors.

Data management across business silos will necessarily be inconsistent. The leadership of one silo may be, more-or-less, invested in data quality than the leadership of another silo.

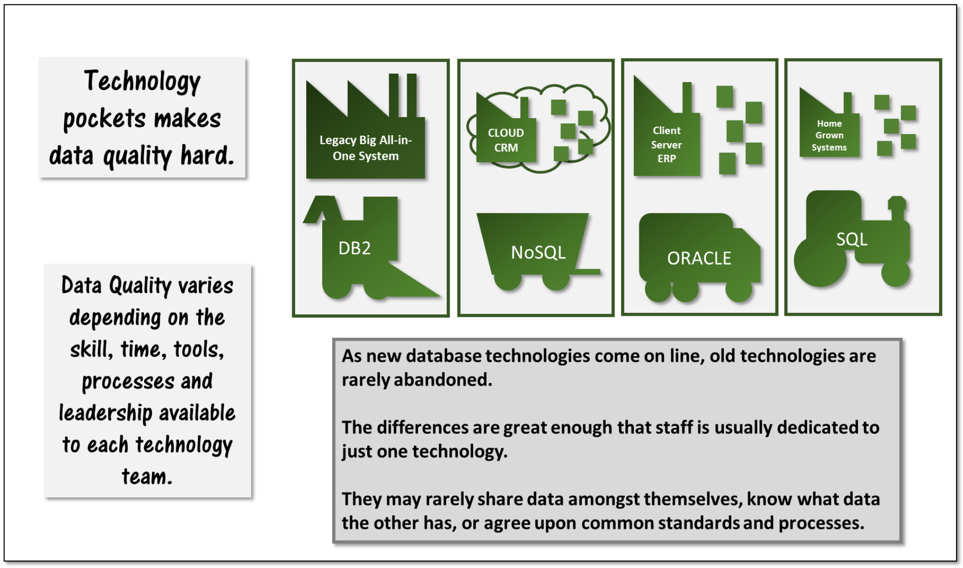
Data professionals are isolated from one another in silos. This makes coordinating efforts to establish and maintain data quality standards difficult.

Technology Pockets

Organizations tend to structure themselves not only by business function but by technology as well.

As new systems are added to an organization, old technologies are rarely abandoned, creating the need for data professionals with distinct knowledge and skills.

As with business silos, data created by a technology tends to stay in that technology. Data management across technologies is inconsistent and data professionals primarily concern themselves with the technology for which they are responsible.



Technology pockets makes data quality hard.

Data Quality varies depending on the skill, time, tools, processes and leadership available to each technology team.

As new database technologies come on line, old technologies are rarely abandoned.

The differences are great enough that staff is usually dedicated to just one technology.

They may rarely share data amongst themselves, know what data the other has, or agree upon common standards and processes.

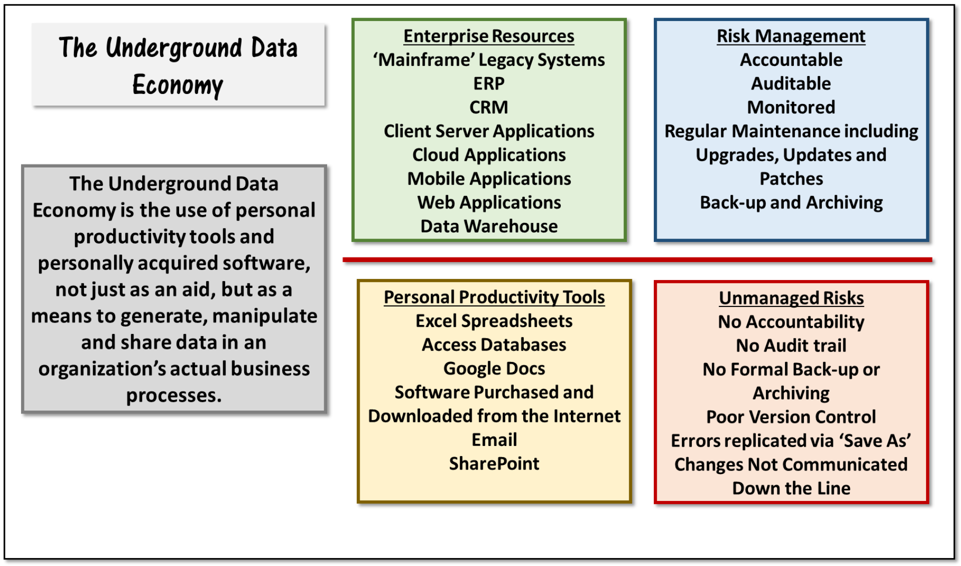
The Underground Data Economy makes Data Quality Hard

When discussing data quality, the attention is most often focused on areas of concern such as major business systems, databases, warehouses and ETL jobs.

However, the leadership of most organizations simply does not comprehend the vast number of Access Databases, Excel Spreadsheets, Word documents and unsupported and unapproved software that is used not just for personal productivity or support tasks, but to carry out actual business functions.

Data is input manually, cut and pasted frequently. Files are saved and renamed regularly. Attached files are worked on directly within email and sent on. Files are stored off prem via Google docs and Google drive. Files not properly archived or deleted. Without version control, multiple copies with slightly different data are in use simultaneously.

The data generated in the underground data economy is relied upon for business performance, financial transactions and for decision making. Yet it exists in a world without oversight as to how, where, or when it is used. This vast amount of unregistered, unregulated and unfortunately, once deleted, unknowable and untraceable data creates tremendous risk for the organization.



The Underground Data Economy

The Underground Data Economy is the use of personal productivity tools and personally acquired software, not just as an aid, but as a means to generate, manipulate and share data in an organization's actual business processes.

Enterprise Resources
'Mainframe' Legacy Systems
ERP
CRM
Client Server Applications
Cloud Applications
Mobile Applications
Web Applications
Data Warehouse

Risk Management
Accountable
Auditable
Monitored
Regular Maintenance including Upgrades, Updates and Patches
Back-up and Archiving

Personal Productivity Tools
Excel Spreadsheets
Access Databases
Google Docs
Software Purchased and Downloaded from the Internet
Email
SharePoint

Unmanaged Risks
No Accountability
No Audit trail
No Formal Back-up or Archiving
Poor Version Control
Errors replicated via 'Save As'
Changes Not Communicated Down the Line

The drivers of the underground data economy

Normally one would applaud an employee's innovative methods for making their job more efficient and getting better and faster results. But what is good for the individual employee is not always good for the organization. There are strong incentives that motivate participation in the underground data economy.

Accessibility

Personal Productivity tools like Excel are already on everyone's desktop, are easy to master and don't require any expensive training or license fees.

Control

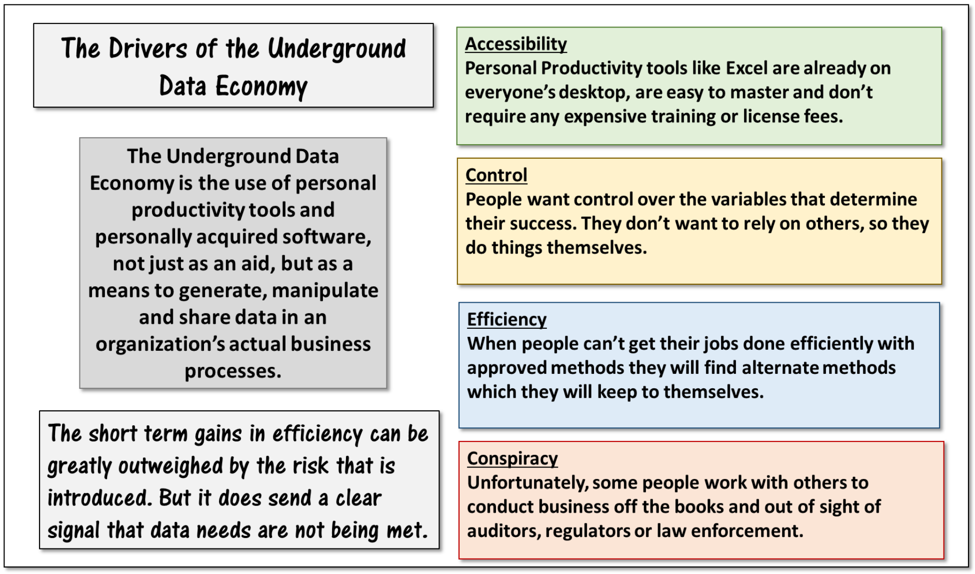
People want control over the variables that determine their success. They don't want to rely on others, so they do things themselves.

Efficiency

When people can't get their jobs done efficiently with approved methods they will find alternate methods which they will keep to themselves.

Conspiracy

Unfortunately, some people work with others to conduct business off the books and out of sight of auditors, regulators or law enforcement.



The Drivers of the Underground Data Economy

The Underground Data Economy is the use of personal productivity tools and personally acquired software, not just as an aid, but as a means to generate, manipulate and share data in an organization's actual business processes.

The short term gains in efficiency can be greatly outweighed by the risk that is introduced. But it does send a clear signal that data needs are not being met.

Accessibility
Personal Productivity tools like Excel are already on everyone's desktop, are easy to master and don't require any expensive training or license fees.

Control
People want control over the variables that determine their success. They don't want to rely on others, so they do things themselves.

Efficiency
When people can't get their jobs done efficiently with approved methods they will find alternate methods which they will keep to themselves.

Conspiracy
Unfortunately, some people work with others to conduct business off the books and out of sight of auditors, regulators or law enforcement.

Sheer Numbers make Data Quality Hard

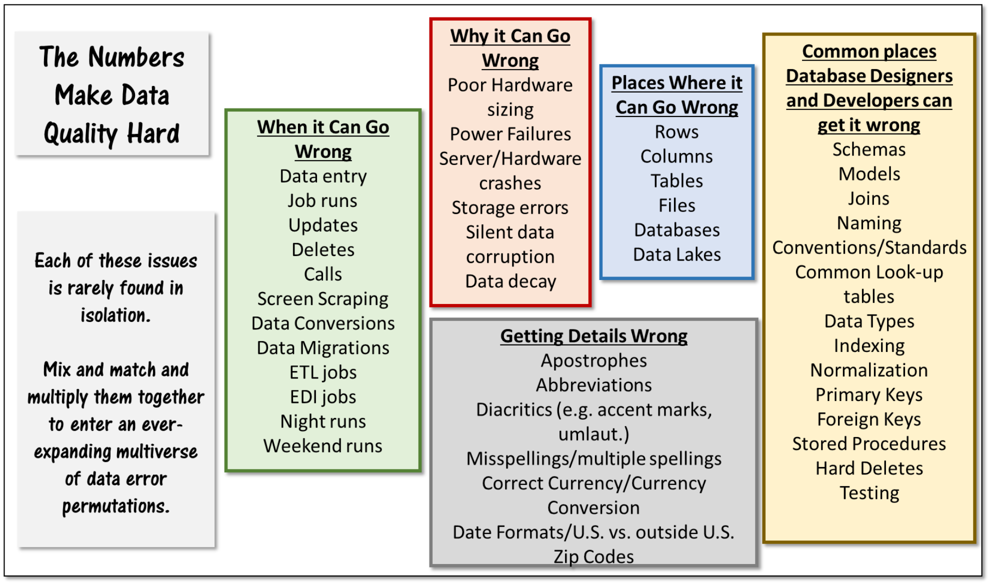
There is no way all your data can be right all the time, yet there are literally millions of ways it can be wrong.

What is wrong with the data is one question, where is it wrong, when is it wrong, how is it wrong and why is it wrong are equally good questions that need to be answered before hoping to improve data quality.

A data issue is rarely found in isolation. Sometimes one issue causes another. A single issue can replicate itself multiple times over. Sometimes completely unrelated issues exist, creating difficulty diagnosing cause and effect. Frustratingly, fixing one issue can cause other issues.

And worse, these things can be happening all at the same time. Multiply this by a host of rows, columns, tables, files and databases and you enter an ever-expanding multiverse of data error permutations.

The math is simply against data quality. To paraphrase Tolstoy "All good data is always good for the same reasons. Bad data is always bad for many unique, varied and creative reasons."



The Numbers Make Data Quality Hard

Each of these issues is rarely found in isolation.

Mix and match and multiply them together to enter an ever-expanding multiverse of data error permutations.

When it Can Go Wrong
Data entry
Job runs
Updates
Deletes
Calls
Screen Scraping
Data Conversions
Data Migrations
ETL jobs
EDI jobs
Night runs
Weekend runs

Why it Can Go Wrong
Poor Hardware sizing
Power Failures
Server/Hardware crashes
Storage errors
Silent data corruption
Data decay

Places Where it Can Go Wrong
Rows
Columns
Tables
Files
Databases
Data Lakes

Getting Details Wrong
Apostrophes
Abbreviations
Diacritics (e.g. accent marks, umlaut.)
Misspellings/multiple spellings
Correct Currency/Currency Conversion
Date Formats/U.S. vs. outside U.S.
Zip Codes

Common places Database Designers and Developers can get it wrong
Schemas
Models
Joins
Naming Conventions/Standards
Common Look-up tables
Data Types
Indexing
Normalization
Primary Keys
Foreign Keys
Stored Procedures
Hard Deletes
Testing

Quality, by its very Definition, is Hard

Often overlooked in the data quality discussion is that the very notion, or idea, of quality sets a very high bar. To achieve a standard that meets or exceeds expectations is one thing. To consistently meet that standard, day in and day out, is another.

To consistently meet or exceed a standard requires a level of resources, attention and presence to which many organizations are unwilling, or unable, to commit.

Also overlooked is that the term quality means many things to many people and those meanings can shift over time. Hitting the quality target is difficult not only because it moves, but because there are multiple targets. Hitting one target inevitably leads to dissatisfaction from not hitting the others.

When faced with this dilemma it is natural to rush to define ‘quality’ for the organization. However, a better approach might be to understand ***not what quality means*** to stakeholders, but ***how the term quality is being used***.

The word ‘quality’ in respect to data, is often used as a placeholder descriptor for any issues that makes the data ‘bad’ in the eyes of a stakeholder. It is much less often used to denote the failure to meet an understood objective standard.

The difference is important because it drives the response to an organization’s data quality ‘problem.’ This is a trap in which IT leaders often find themselves. No matter how hard IT leaders try to put objective measures to the quality of ***any*** IT service, the standard is always subject to the perception of the individuals using the service.

Before starting any data quality initiative, it is therefore important to truly understand the nature of an organization’s data quality from the perspective of the people trying to use the data***.*** The stakeholder’s perception is rarely an assessment of what the data is, but a reaction to what the data is not. ***‘Poor Quality’ is often shorthand for ‘I’m not getting what I need.’***

Rather than ask the question of stakeholders ‘What does data quality mean to you?’ it is better to ask, ‘What do you need that you are not getting from your data?” The answer will have a significant impact on the level of effort an organization should put into its data quality efforts.

It may turn out that a comprehensive, enterprise-wide, inter-disciplinary approach isn't what's needed at all. It would be unfortunate if the face that launched a thousand data quality projects could have been resolved, and satisfied a frustrated stakeholder, by simply refreshing a database several times a day rather than just once overnight.

No Silver Bullet makes Data Quality Hard.

Data Quality would not be much of an issue to fix if:

- It was just a matter of purchasing a tool, adopting a methodology or adding staff.
- All data and data environments looked pretty much the same and solutions could be applied universally.
- There were only one or two dominant database vendors who could roll out updates and new features to all their customers simultaneously.
- There was a single class, seminar, bootcamp or degree program to which staff could be sent.
- Human beings didn't have the unquenchable desire to improve, modify and customize their data, databases and data environments to meet their 'unique' needs.

While there are many commonalities between businesses and industries, the many and varied differences of data, data environments and data capabilities ultimately negate simple universal solutions.

Unfortunately, there is no silver bullet. Each organization is required to fully understand the breadth, depth and scope of its data and data environment. It then needs to do the hard work necessary to assemble its own customized combination of talent, tools, methods, disciplines and training to improve its data-related capabilities.

Conclusion

Assembling the right combination of tools, methodologies, disciplines and skill will go a long way to improving data quality Unfortunately data management methodologies, data quality tools and data governance, while necessary, are inevitably inadequate.

The forces arrayed against data quality are too formidable. No amount of committee meetings, data stewards, or the best intentions of individuals will overcome these forces.

If clear heads, stout hearts and high-minded purpose is not enough to improve data quality, what is?

What is required is another way to look at data quality. ***What is most often overlooked in the quest for data quality is the organization's data-related activities that are creating the data in the first place.***

If the corporate adventure with quality over the last fifty years has shown anything, it is that quality cannot be imposed from above, from the outside or at the back end. Data Quality***, like any other quality in any other endeavor***, must be built from the bottom up, from the inside, and from the start.

About the Author

Robert Grant Beauchamp is a consultant, architect, and former CIO with a proven record of helping organizations understand and improve their data quality, data environments and data-based capabilities. As a systems integrator, Robert has successfully introduced and implemented data-related technologies such as BI, EDI, ETL, data warehousing and three-tier architectures.

If you would like to learn more about data quality or would like help improving data quality within your organization, connect with him on LinkedIn or at datahust.com.

Robert has filled the roles of computer journalist, tech writer, business analyst, marketing communications manager, business architect, project manager, program integrator, program manager, account manager, data security consultant, solutions architect and trusted advisor, including:

- Thirty years of business and information technology experience including over five years as Chief Information Officer of a rapidly growing health plan.
- Proven track record of successfully strategizing, developing and implementing enterprise-level business and technology initiatives in the health care and financial services industries.
- An experienced and practiced consultant with the ability to work with C-level executives to develop strategy, assess capabilities, manage risk, and offer solutions that can be successfully implemented in an organization's unique environment.
- A proven communicator well versed in public speaking, meeting facilitation, webinars, journalism and video.
- A proven history of developing and implementing successful service offerings for a major IT consulting firms including Y2K, HIPAA Security and Privacy and HIPAA Electronic Transactions.

Currently Mr. Beauchamp is championing a capability-based approach to data quality. He is a leader in the adoption, education, and implementation of Data Sourcing as a corporate capability. He is in the process of writing a book on Data Sourcing.