Data Quality Fundamentals

David Loshin
Knowledge Integrity, Inc.
www.knowledge-integrity.com

Agenda

- The Data Quality Program
- Data Quality Assessment
- Using Data Quality Tools
- Data Quality Inspection, Monitoring, and Control

THE DATA QUALITY PROGRAM

Data Quality Challenges

- Consumer data validation of supplied data provides little value unless supplier has an incentive to improve its product
- Data errors introduced within the enterprise drain resources for scrap and rework, yet the remediation process seldom results in long-term improvements
- Reacting to data integrity issues by cleansing the data does not improve productivity or operational efficiency
- Ambiguous data definitions and lack of data standards prevents most effective use of centralized "source of truth" and limits automation of workflow
- Proper data and application techniques must be employed to ensure ability to respond to business opportunities
- Centralization of integrated reference data opens up possibilities for reuse, both of the *data* and the *process*

Addressing the Problem

- To effectively ultimately address data quality, we must be able to manage the
  - Identification of customer data quality expectations
  - Definition of contextual metrics
  - Assessment of levels of data quality
  - Track issues for process management
  - Determination of best opportunities for improvement
  - Elimination of the sources of problems
  - Continuous measurement of improvement against baseline

Data Quality Framework


Data quality expectations


Measurement


Policies


Procedures


Governance


Standards


Monitor Performance


Training

Data Quality Policies

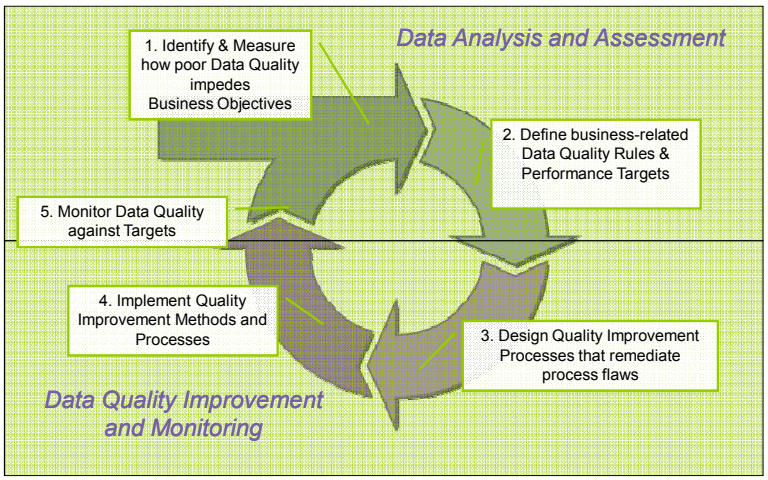
- Direct data management activities towards managing aspects of compliance with business directives, such as:
  - Data certification
  - Privacy management
  - Data lineage
  - Limitation of Use
  - Unified source of reference

Data Quality Procedures

- Data quality management processes support the observance of the data quality policies; examples include:
  - Standardized data inspection templates
  - Operational data quality
  - Issues tracking and remediation
  - Manual intervention when necessary
  - Integrity of data exchange
  - Contingency planning
  - Data validation

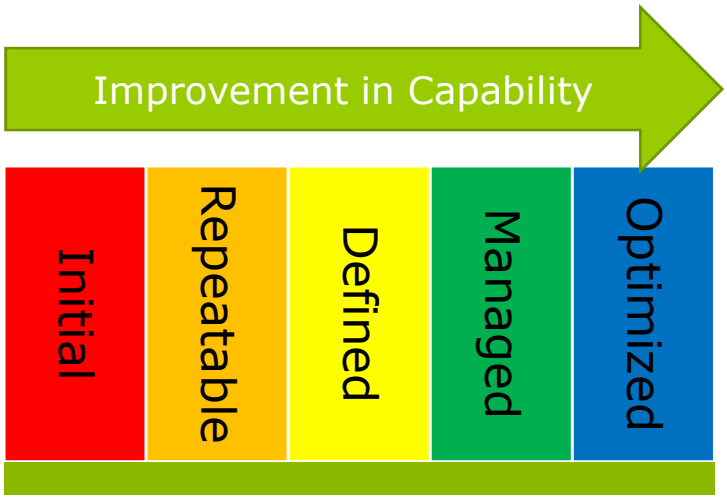
Data Quality Processes

Measurement, Discovery, Continuous Monitoring



Source: Informatica

Capability/Maturity Model



Data Quality Expectations

Level	Characterization
Initial	• Data quality activity is reactive • No capability for identifying data quality expectations • No data quality expectations have been documented
Repeatable	• Limited anticipation of certain data issues • Expectations associated with intrinsic dimensions of data quality can be articulated • Simple errors are identified and reported
Defined	• Dimensions of data quality are identified and documented • Expectations associated with dimensions of data quality associated with data values, formats, and semantics can be articulated using data quality rules • Capability for validation of data using defined data quality rules • Methods for assessing business impact explored
Managed	• Data validity is inspected and monitored in process • Business impact analysis of data flaws is common • Results of impact analysis factored into prioritization of managing expectation conformance • Data quality assessments of data sets performed on cyclic schedule
Optimized	• Data quality benchmarks defined • Observance of data quality expectations tied to individual performance targets • Industry proficiency levels are used for anticipating and setting improvement goals • Controls for data validation integrated into business processes

Dimensions of Data Quality

Level	Characterization
Initial	• No recognition of ability to measure data quality • Data quality issues not connected in any way • Data quality issues are not characterized within any kind of management taxonomy
Repeatable	• Recognition of common dimensions for measuring quality of data values • Capability to measure conformance with data quality rules associated with data values
Defined	• Expectations associated with dimensions of data quality associated with data values, formats, and semantics can be articulated • Capability for validation of data values, models, and exchanges using defined data quality rules • Basic reporting for simple data quality measurements
Managed	• Dimensions of data quality mapped to a business impact taxonomy • Composite metric scores reported • Data stewards notified of emerging data flaws
Optimized	• Data quality service level agreements defined • Data quality service level agreements observed • Newly researched dimensions enable the integration of proactive methods for ensuring the quality of data as part of the system development life cycle.

Policies

Level	Characterization
Initial	• Policies are informal • Policies are undocumented • Repetitive actions taken by many staff members with no coordination
Repeatable	• Organization attempts to consolidate “single source of truth” data sets • Privacy and Limitations of Use policies are hard-coded • Initial policies defined for reacting to data issues
Defined	• Tailored guidelines for establishing management objectives are established at line of business • Certification process for qualifying data sources is in place • Best practices captured by data quality practitioners • Data quality service level agreements defined for managing observance of policies
Managed	• Policies established and coordinated across the enterprise • Provenance management details the history of data exchanges • Policy-based data quality management • Performance management driven by data quality policies • Data quality service level agreements used for managing observance of policies
Optimized	• Automated notification of noncompliance to data quality policies • Self governing system in place

Procedures

Level	Characterization
Initial	• Discovered failures are reacted to in an acute manner • Data values are corrected with no coordination with business processes • Root causes are not identified • Same errors corrected multiple times
Repeatable	• Ability to track down errors due to incompleteness • Ability to track down error due to invalid syntax/structure • Root cause analysis enabled using simple data quality rules and data validation
Defined	• Procedures defined and documented for data inspection for determination of validity • Data quality management is deployed at line of business level as well as at enterprise level • Data validation is performed automatically and only flaws are manually inspected • Data contingency procedures in place
Managed	• Data quality rules are proactively monitored • Data controls are designed for incorporation into distinct business applications • Data flaws are recognized early in information flow • Remediation is governed by well-defined processes • Validation of exchanged data in place • Validity of data is auditable
Optimized	• Data controls deployed across the enterprise • Participants publish data quality measurements • Data quality management practices are transparent

Governance

Level	Characterization
Initial	• Little or no communication regarding data quality management • Information Technology is default for all enterprise data quality issues • No data stewardship • Responsibility for data corrections assigned in an ad hoc manner
Repeatable	• Best practices are collected and shared among participants. • Key individuals from community form workgroup to devise and recommend Data Governance program and policies • Guiding principles and data quality charter are in development
Defined	• Organizational structure for data governance oversight defined • Guiding principles, charter, and Data Governance Management Policies are documented • Standardized view of data stewardship across the enterprise • Operational data governance procedures defined
Managed	• Data Governance Board consisting of representatives from across the enterprise is in place. • Collaborative Data Quality Governance Board meets on a regular basis • Operational data governance driven by data quality service level agreements • Teams within each division or group employ similar governance framework internally • Reporting and remediation frameworks collaborate in applying statistical process control to maintain control within defined bounds
Optimized	• DQ performance metrics for processes are reviewed for opportunities for improvement • Staff members rewarded for meeting data governance performance goals

Standards

Level	Characterization
Initial	• No data standards defined • Similar data values represented in variant structures • No data definitions
Repeatable	• Data element definitions for commonly used business terms • Reference data sets identified • Data elements used as identifying information specified • Certification process for trusted data sources being defined • Data standards metadata managed within participant enterprises • Definition of guidelines for standardized exchange formats (e.g., XML)
Defined	• Enterprise data standards and metadata management • Structure and format standards defined for all data elements • Exchange schemas are defined
Managed	• Certification of trusted data sources in place • Master reference data sets identified • Exchange standards managed through data standards oversight process • Data standards oversight board oversees ongoing maintenance of internal standards and conformance to externally-defined standards
Optimized	• Master data concepts managed within a master data environment • Taxonomies for data standards are defined and endorsed • Conformance with defined standards is integrated via a policy-oriented technical structure • Straight-through processing is enabled for standard data

Technology

Level	Characterization
Initial	• Internally developed ad hoc routines employed • "Not invented here" mentality
Repeatable	• Tools for assessing objective data quality are available • Data parsing, standardization, and cleansing are available • Data quality technology used for locate, match, and linkage.
Defined	• Standardized procedures for using data quality tools for data quality assessment and improvement in place • Business rule-based techniques are employed for validation • Technology components for implementing data validation, certification, assurance, and reporting are in place • Technology components are standardized across the federated community at the service and at the implementation layers
Managed	• Automatic data correction guided by governance policies and defined business rules • Impact analysis and what-if scenarios supported by dashboard and reporting tools
Optimized	• Non-technical users can define and modify data quality rules and dimensions dynamically

Performance Management

Level	Characterization
Initial	• Impacts are manifested and recognized long after failure events take place
Repeatable	• Characterization of areas of impact of poor data quality • Data profiling used to identify data failures in process
Defined	• Impact analysis framework in place • Data quality service components identify flaws early in process • Data quality service components defined • Issues tracking system in place to capture issues and their resolutions.
Managed	• Data quality metrics fed into performance management reporting • Auditing based on conformance to rules associated with data quality dimensions • Consistent reporting of data quality management for necessary participants • Performance dashboards are in place • Role-based access to performance information • Well-defined visualization of data quality component contribution to business impacts
Optimized	• Enterprise-wide performance can be improved through policy modification via rules environment

Data Quality Management Lifecycle

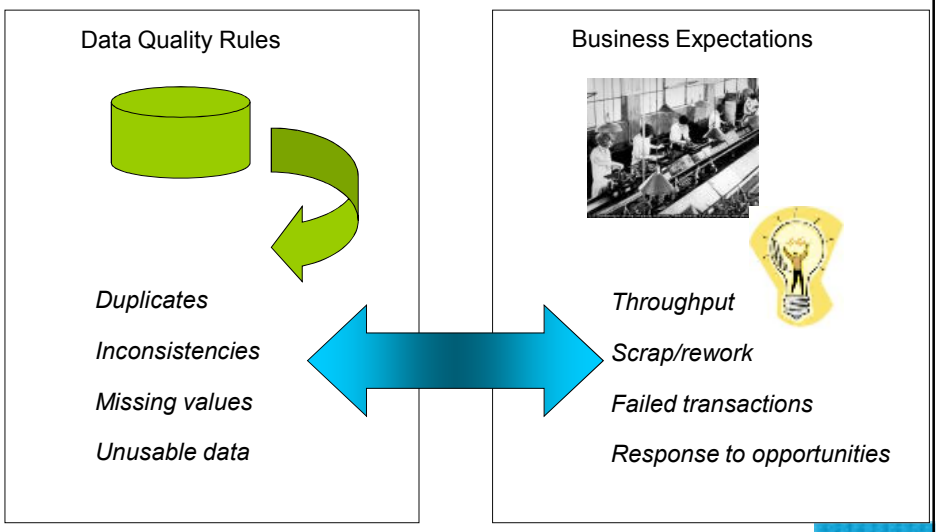
Initial Stage	*Early life cycle*	*Continuous Process Improvement*	*Mature*
Initial investment in: •Assessment •Consulting •Tools Time investment in: •Impact analysis •Building a business justification •Architecting governance model •Understanding data ownership paradigms •Education and socialization	Early life cycle resource investment in: •Root cause analysis •Reaction to, and evaluation to underlying issues •Resolution techniques •Identify dimensions of data quality •Assess improvement opportunities and target measures of data quality •Identify key pilot projects •Evaluate business needs for tools acquisition •Tools acquisition	Maturation phase: • Continuous improvement •Identify and solidify best practices •Transition from reactive to proactive •Pareto Principle: 80% of benefit achieved through 20% of work •Diminishing returns	Later life cycle: •Maturity models •Proactive management •Transition of governance to all staff •Reduction in staff dedicated to "reaction" •Data stewardship managed by lines of business

DATA QUALITY ASSESSMENT

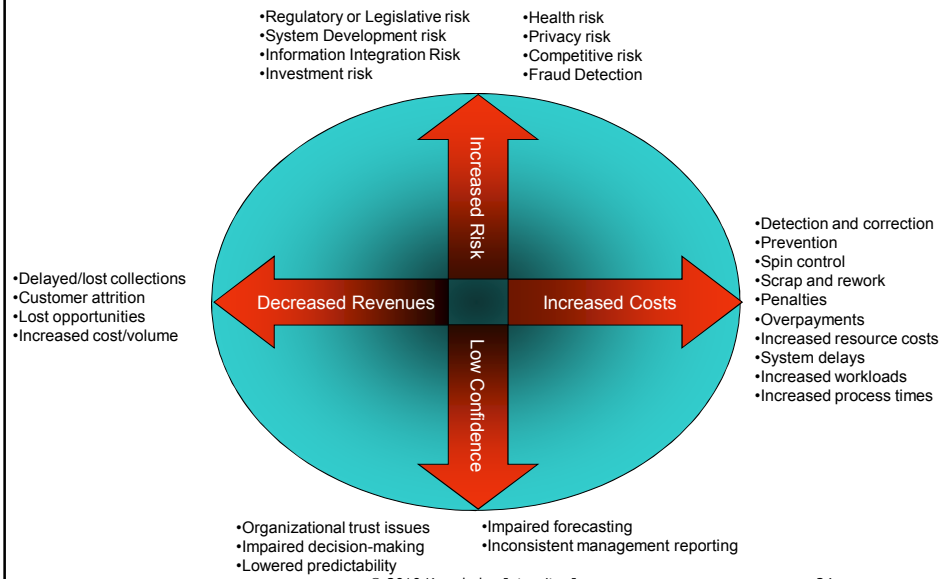
Building the Business Case

- Identify key business performance criteria related to information quality assurance
- Review how data problems contribute to each business impact
- Determine the frequency that each impact occurs
- Sum the measurable costs associated with each impact incurred by a data quality issue
- Assign an average cost to each occurrence of the problem
- Validate the evaluation within a data governance forum

Business Expectations and Data Quality



Business Impacts



Examples – Increased Costs

- Large Office Supply Company:
  - Redundant data and unused data accounted to large percentage of their storage usage
  - Elimination of unused or redundant data would result in significant (20%) reduction in DASD (and corresponding data management) costs
- DoD Guidelines on Data Quality:
  - "… the inability to match payroll records to the official employment record can cost millions in payroll overpayments to deserters, prisoners, and "ghost" soldiers."
  - "… the inability to correlate purchase orders to invoices is a major problem in unmatched disbursements."
- Manufacturing Company:
  - Inability to determine that similar components had already been designed and built incurred duplicated design and development costs exceeding $70,000 per item

Examples – Decreased Revenues

- Telecommunications company:
  - Applied revenue assurance to detect underbilling indicated revenue leakage of just over 3 percent of total revenue due to poor data quality
  - Identified 49 misconfigured (but assumed to be unusable) high-bandwidth circuits that could be returned to productive use
- Federal Agency:
  - "55 percent of the records in a building database were wrong, resulting in the underbilling of $12 million in rent."
- Another Federal Agency:
  - Stale contact addresses slows process of collecting debt obligations

Examples – Decreased Confidence

- Pharmaceutical company:
  - Large investment made in creating front-end sales application fed by back-end database
  - Application clients refused to use new application due to mistrust of back-end database
- Agriculture company:
  - Multiple sales databases conflicted with accounting databases
  - Sales staff did not trust that their commissions were being properly calculated

Examples – Increased Risk

- Pharmaceutical/Medical Device company
  - Party database used to manage grantees
  - Grantees may also be providers
  - Inability to properly track grantees exposed company to risk of violating Federal Anti-Kickback statute
- Banking industry, credit risk:
  - "PWC estimates that 90% of the top 100 world banks are deficient in credit risk data management in…maintenance of clean counterparty static data repositories, … common counterparty identifiers, …, staff dedicated to data quality, consistent data standards."

Business-Driven Information Requirements

Driver	Benefit	Data Quality Requirement
Supplier Insight	*Complete supplier view, consistency across applications*	Unified supplier data, matching/linkage, eliminate redundancy
Data Enrichment	*Enhanced views, improved analytics*	Matching/linkage, vendor management, 3rd party data integration
Operational Efficiency	*Lowered costs, streamlined processes, increased volumes, increased throughput*	STP, eliminate redundant data, functionality, licenses, rules/policy-driven
Organizational Performance	*Optimized staff value, opportunities to improve or change business processes*	Centralized analytics, unified employee data, inspection, monitoring, control
Risk/Compliance	*Compliance, privacy, risk management, accurate response to audits*	Data quality, semantic consistency across business processes, consistency, availability
Product Performance	*Product design, time to market, product performance, better manufacturing processes*	Unified product data, matching/linkage, centralized analytics
Marketing Intelligence	*Cross-sell/up-sell, segmentation and targeting, improved product development*	Unified master data, matching/linkage, centralized analytics
Customer Service	*Satisfaction, retention, ease of doing business*	Multi-channel data provision, embedded analytics, unified customer data

Assessing and Addressing the Quality of Data

Unified supplier data, matching/linkage, eliminate redundancy

Matching/linkage, vendor management, 3rd party data integration

STP, eliminate redundant data, functionality, licenses, rules/policy-driven

Centralized analytics, unified employee data, inspection, monitoring, control

Data quality, semantic consistency across business processes, consistency, availability

Unified product data, matching/linkage, centralized analytics

Unified master data, matching/linkage, centralized analytics

Multi-channel data provision, embedded analytics, unified customer data

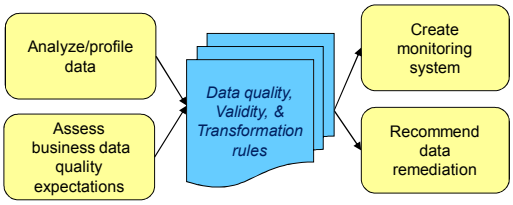
- Important questions:
  - What are the most critical business issues attributable to poor data quality?
  - What constitutes "poor" data quality?
  - How is data quality measured?
  - What are the levels of acceptability?
  - How are data issues managed?
  - What remediation and correction actions are feasible?
  - How can we know when the data has been improved?
  - How is data quality improvement related to business process performance?

Data Quality Assessment: Objectives

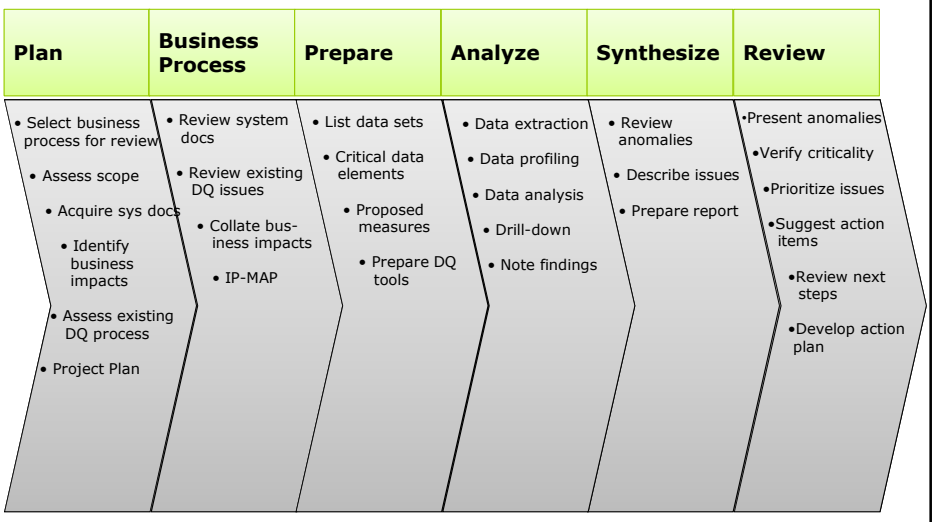
- Identify key business objectives and corresponding metrics
  - *Identify specific data issues related to known business impacts*
  - *Correlate discovered issues to business impacts*
- Data profiling and analysis
  - *Understand what you are working with, provide quantified metrics*
- Improve automated matching/linkage
  - *Reduce false positives, expand universe of identifying attributes, reduce need for manual intervention*
- Institute managed data quality
  - *Collect organizational data requirements, data inspection and control, incident management, data quality scorecards*

Analysis Process: Correlating Business and Data Issues

- *Business Impact Analysis*
  - Identify business data issues
  - Prioritize impacts
  - Identify critical data elements
  - Correlate data dependencies and business impacts
- Engage business subject matter experts
- *Empirical Analysis*
  - Statistical analysis of actual existing data
  - Identification of potential anomalies
  - Validation of known expectations
- Data profiling



Data Quality Assessment – Process



Business Interviews and Classifying Business Impacts

- Interview business subject matter experts to isolate and categorize business impacts

Impact Category	Examples of issues for review
Operational Efficiency	• Time and costs of cleansing data or processing corrections • Inaccurate performance measurements for employees • Inability to identify suppliers for spend analysis
Risk/Compliance	• Missing data leads to inaccurate credit risk • Regulatory compliance violations
Revenue	• Lost opportunity cost • Identification of high value opportunities
Productivity	• Decreased ability for straight-through processing via automated services
Procurement Efficiency	• Improved ease-of-use for staff (sales, call center, etc.) • Improved ease of interaction for requestors and approver • Reduced time from order to delivery
Performance	• Impaired decision-making

Using the Business Impact Template

Issue ID	Data Issue	Business Impact	Measure	Severity
Assigned identifier for the issue	*Description of the issue*	*Description of the business impact attributable to the data issue; there may be more than one impact for each data issue*	*A means for measuring the degree of impact*	*An estimate of the quantification of the cumulative impacts*

Empirical Data Observations and Synthesis

ID	Table and Column Name	Inspection	Reported items	Issues for Review	Fitness Assessment
Assigned identifier for issue	*Table name and column name(s)*	*What measure or dimensions were reviewed*	*Result of measurement*	*What needs to be reviewed, next steps*	*Characterized based on business impact and severity*

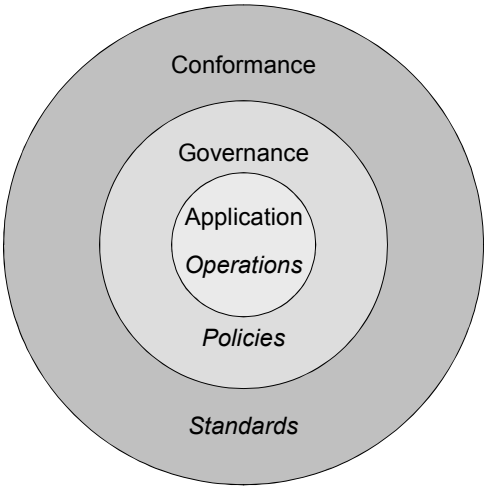
Review

- Document key business issues that are attributable to poor data quality
- Perform empirical assessment to identify potential anomalies
- Prioritize based on
  - Correlation to business impact(s)
  - Severity of impact
  - Opportunity for improvement
- Scope focus to areas that can feasibly provide tactical improvements and strategic value
- Specify data inspection rules to quantify levels of acceptability
- Institute inspection, monitoring, and reporting
- Provide continuous process for assessment, remediation, reporting of measurable improvement

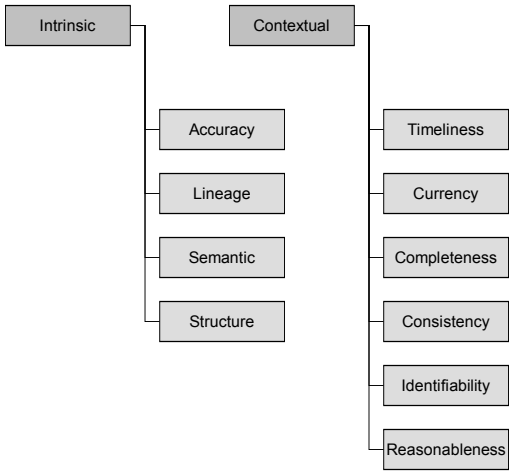
What are "Dimensions of Data Quality"?

- The concept of a dimension evokes thoughts of measurement
- Different dimensions are intended to represent different measurable aspects of data quality
  - Used in characterizing relevance across a set of application domains and to ensure an enterprise standard of data quality
  - Measurements are taken to review data quality performance at different levels of the operational hierarchy
  - Monitoring overall both line-of-business and enterprise performance
- Each group within the organization has the freedom to introduce its own dimensions with customized characteristics.

Categorization of Dimensions



Dimensions of Data Quality



Caveat

- This list is intended as guidance and as a starting point for defining the dimensions that are relevant within the organization
- The methods for measurement should be identified before agreeing to the selection of a metric
- The metrics should be correlated with the impacts identified during the impact analysis

Accuracy

- Integral to ensuring that the data values that are managed are accurate with respect to systems of record
  - System of Record
    - *A registry of data sets for accuracy corroboration exists*
  - Precision
    - *Data elements are defined with the proper level of precision*
  - Value Acceptability
    - *Acceptable values for each data element are defined*
  - Domain Definitions
    - *Commonly used data value sets have conceptual domains defined*
    - *Value domains for conceptual domains are enumerated once*
  - Value Accuracy
    - *Data values are accurate*

Critical *for validation of information against recognized sources of record*

Lineage

- Documentation of originating data source
  - Do data elements incorporate attribution of its original source and date
  - Are provenance audit trails archived?

Critical *to managing compliance with information policies (e.g., limitation of use), root cause analysis, and quality ratings*

Semantic

- Semantic consistency refers to:
  - consistency of definitions among attributes within a data model
  - similarly named attributes in different enterprise data sets
  - the degree to which similar data objects share consistent names and meanings

Semantic

- Data definitions
  - *A metadata repository with all data elements named and defined is available for all participants*
- Conformance to naming conventions
  - *An enterprise naming convention has been documented and all data element names conform to the convention*
- Name ambiguity
  - *No two elements share the same name*
- Semantic consistency
  - *Similarly named data attributes are assured to refer to the same business concept*

***Critical** for maintaining concept consistency across systems*

Structure

- Syntactic consistency
  - *Formats of shared data elements that share the same value set have the same size and data type*
- Documentation of common types
  - *Data element length and type are specified in the metadata repository*

***Critical** for organizational data standardization*

Timeliness

- Accessibility
  - *Newly posted records should be available to enterprise applications within a specified time period*
  - *Policies specifying acceptable time delays must be provided.*
- Response time
  - *Ensure that requested data is provided within the acceptable time period*
  - *Expectations for response time must be specified*

***Critical** for observing service levels for data availability and synchrony*

Currency

- Age/Freshness
  - *The acceptable time period lifetime between updates for each data element is defined - Expiry date*
- Time of release
  - *The date/time upon which the data becomes available is defined*
  - *If data is expected to be delivered to specified participants, the release date/time should be specified*
- Synchronization of replicas
  - *Data synchronizations and replication policies between systems must be specified*
- Correction/update promulgation
  - *Polices for promulgation of corrections and updates, must be specified.*
- Temporal
  - *Temporal Consistency rules are valid*

***Critical** for observing service levels for data availability and synchrony*

Completeness

- Population Density
  - *Specify the minimum degree of population for each data element*
- Optionality
  - *Mandatory attributes are expected to have assigned values in a data set*
  - *Optionality must be specified for all data elements*
- Null validation
  - *Null value rules for all data elements are defined*
  - *Null value rules are conformed to*

***Critical** for ensuring that required information has been acquired*

Consistency

- Presentation
  - *Common presentation formats for each data element are defined*
- Presentation completeness
  - *Each data presentation format can convey all information within the attributes*
- Null presentation
  - *Standards for the presentation of absent information for each data type are defined*
- Capture and collection
  - *Data entry edits and data importation rules should be defined for each data element*

***Critical** for standardizing representation and management of shared concepts*

Identifiability

- Entity uniqueness
  - *No entity exists more than once within the system*
- Search and match
  - *A probability of a successful or partial match for the identifying information associated with a specific record will be defined*
- Coverage
  - *The central repository is expected to identify the universe of unique entities across the enterprise*
  - *The potential total universe of entities by classification must be defined*
- Linkage
  - *Links between data records in different data sets is properly maintained*

***Critical** for reducing duplication and improving*

Reasonability

- Multi-value consistency
  - *The value of one set of attributes is consistent with the values of another set of attributes*
- Temporal reasonability
  - *New values are consistent with expectations based on previous values*
- Agreements
  - *Service level agreements (SLA) governing data provider performance will be defined*

***Critical** for business rule validation*

USING DATA QUALITY TOOLS

Data Quality Activities and Technologies

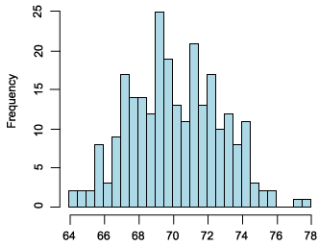
Analysis	*Cleansing*	*Enhancement*	*Monitoring*
•Data profiling	•Parsing and standardization •Matching •Transformation	•Matching •Enhancement	•Data profiling •Reporting

Data Analysis Using Data Profiling

- Empirical analysis of "ground truth"
  - Statistical analysis
  - Functional dependency analysis
  - Association rule analysis
- Rule validation can be used for monitoring
- Three activities:
  - Column
  - Cross-column
  - Cross-table

Column Profiling Techniques

- Range Analysis
- Sparseness
- Format Evaluation
- Cardinality and Uniqueness
- Frequency Distribution
- Value Absence
- Abstract Type Recognition
- Overloading



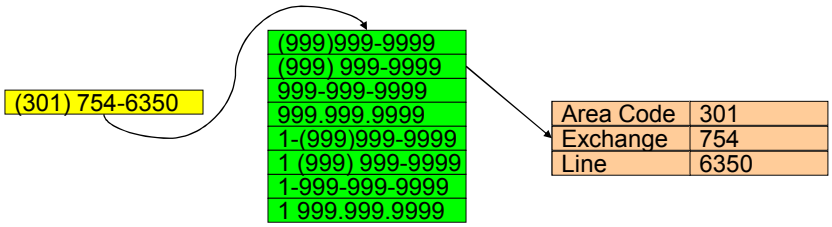
Cross-Column Analysis

- Key discovery
- Normalization & structure analysis
- Derived-value columns
- Business rule discovery

Ongoing Monitoring Using Data Profiling

- Rule validation can be used to assert data quality expectations throughout the processing flow
- Use profiling jobs as "probes" across the information flow graph to identify where flaws are introduced
- Correlate occurrences of errors to documented business impact for prioritization

Parsing and Standardization



Triggering Actions

- Whether the string matches a pattern or not, actions can be triggered, e.g.:
- If the string can be parsed:
  - The tokens can be extracted and forwarded into component data element attributes
  - Tokens can be transformed into a standard form
- If the string cannot be parsed into a success pattern:
  - There may be common error patterns than can trigger corrections
  - Uncorrectable errors can be forwarded back to the data owner

Data Correction

- If we can automatically recognize data as not conforming to a standard, can we automate its correction?
- If we have translation rules or mappings from incorrect values to correct values
- This is how many data cleansing applications work
  - example: Internatinal→International

Data Standardization

- Use standard form as a pivot for linkage and consolidation
- Example
  - Elizabeth R. Johnson, 123 Main St
  - Beth R. Johnson, 123 Main St
- It's a good hunch that these records represent the same person
- We can standardize components based on nicknames, abbreviations, etc.

Transformation Rules

- Standardization is a process of transforming nonconforming forms to conforming forms
- Use mappings/transformation rules
- Create a rule engine instance and integrate the rules
- Engine becomes a filter
- This capability is embedded in many data cleansing tools, as well as in many ETL/integration tools

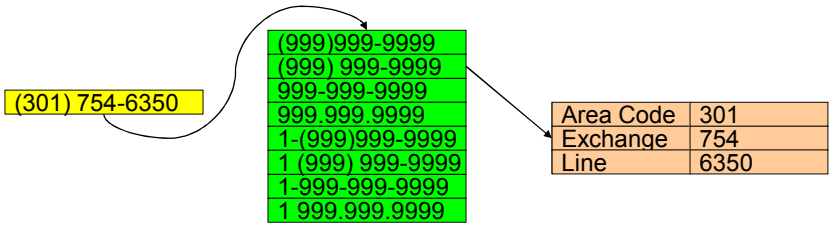
Parsing

- Similar data concepts may be represented using a collection of common patterns
- Parsing is a process of defining patterns and using those patterns to identify key tokens

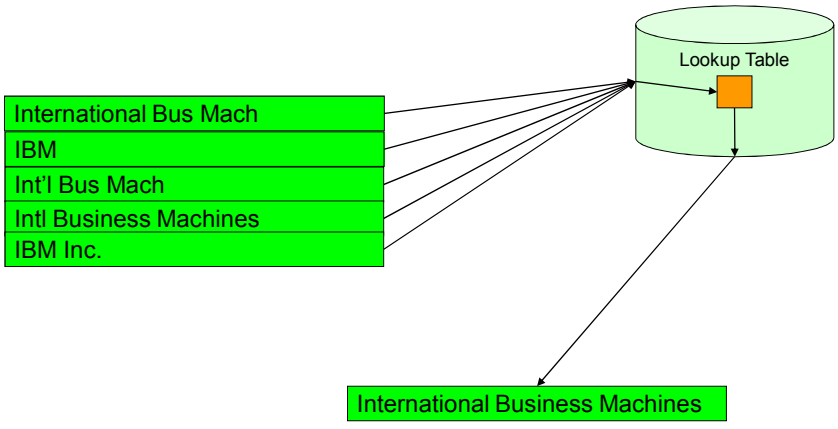
Area code 301
Exchange 754
Line 6350

Standardization

- Standardization uses the patterns to distinguish valid from invalid data values
- Valid values are parsed and their component tokens can be rearranged into a standard form



Data Cleansing



Matching/Record Linkage

- Identity recognition and harmonization
- Approaches used to evaluate "similarity" of records
- Use in:
  - Duplicate analysis and elimination
  - Merge/Purge
  - Householding
  - Data Enhancement
  - Data Cleansing
  - Customer Data Integration

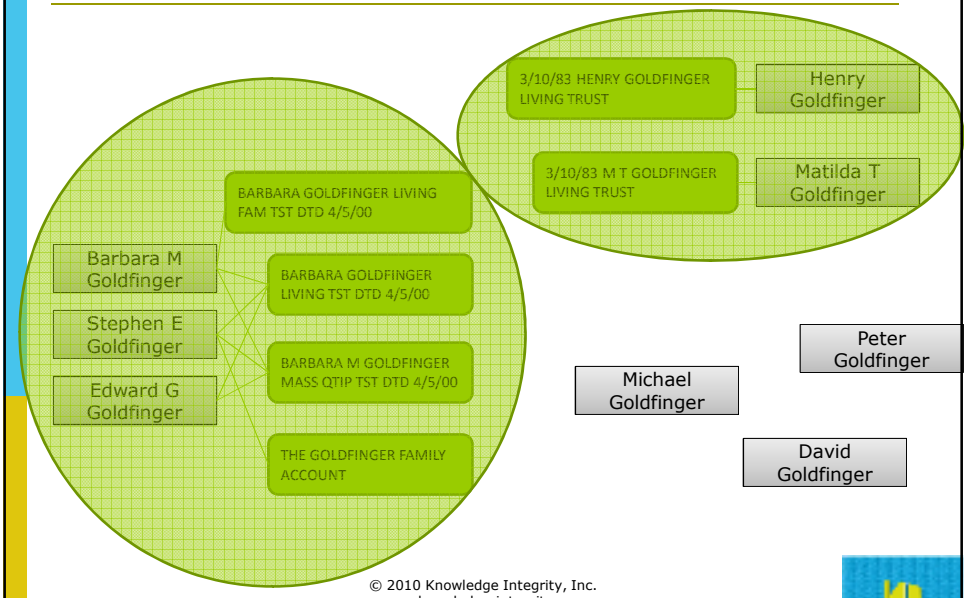
Example – Identity and Entities

BARBARA GOLDFINGER LIVING TST	DTD 4/5/00 BARBARA GOLDFINGER	STEPHEN GOLDFINGER TRUSTEES	33 BIRCH HILL ROAD	W NEWTON, MA 02165
BARBARA M GOLDFINGER FAM TST	DTD 4/5/00 STEPHEN GOLDFINGER	& EDWARD G GOLDFINGER TSTEES	33 BIRCH HILL ROAD	W NEWTON, MA 02165
BARBARA M GOLDFINGER MASS QTIP	TST DTD 4/5/00 STEPHEN E	& EDWARD G GOLDFINGER TTEES	33 BIRCH HILL ROAD	W NEWTON, MA 02165
DAVID GOLDFINGER	6 CHANDLER STREET	LEXINGTON, MA 02420		
EDWARD GOLDFINGER	950 FOUNTAIN STREET	ANN ARBOR, MI 48103		
HENRY GOLDFINGER	TTEE 3/10/83 HENRY GOLDFINGER	LIVING TRUST	P O BOX 320372	TAMPA, FL 33679
MATILDA T GOLDFINGER	TTEE 3/10/83 M T GOLDFINGER	LIVING TRUST	P O BOX 320372	TAMPA, FL 33679
MICHAEL GOLDFINGER	11 CRESCENT HILL AVE	LEXINGTON, MA 02420		
PETER GOLDFINGER	7506 HAMPTON AVE	LOS ANGELES, CA 90046		
STEPHEN GOLDFINGER	THE GOLDFINGER FAMILY ACCOUNT	33 BIRCH HILL ROAD	WEST NEWTON, MA 02165	

"Entities"

BARBARA GOLDFINGER
BARBARA M GOLDFINGER
BARBARA GOLDFINGER LIVING TST DTD 4/5/00
BARBARA GOLDFINGER LIVING FAM TST DTD 4/5/00
BARBARA M GOLDFINGER MASS QTIP TST DTD 4/5/00
BARBARA GOLDFINGER TRUSTEE
BARBARA GOLDFINGER TSTEE
STEPHEN GOLDFINGER
STEPHEN GOLDFINGER TRUSTEE
STEPHEN GOLDFINGER TSTEE
STEPHEN E GOLDFINGER
STEPHEN E GOLDFINGER TTEE
EDWARD GOLDFINGER
EDWARD G GOLDFINGER
EDWARD G GOLDFINGER TSTEE
EDWARD G GOLDFINGER TTEE
DAVID GOLDFINGER
HENRY GOLDFINGER
HENRY GOLDFINGER TTEE
MATILDA T GOLDFINGER
MATILDA T GOLDFINGER TTEE
MICHAEL GOLDFINGER
PETER GOLDFINGER
THE GOLDFINGER FAMILY
THE GOLDFINGER FAMILY ACCOUNT
3/10/83 HENRY GOLDFINGER LIVING TRUST
M T GOLDFINGER

Matching/Record Linkage



Data Correction

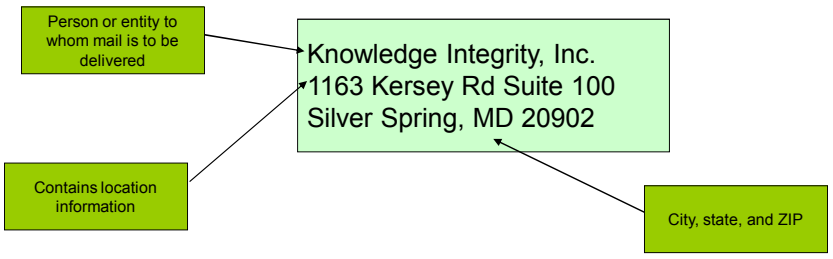
- Correction by consolidation
- Makes use of **record linkage**
  - Find a pivot attribute across which to link
  - The pivot should be unique (such as social security number)
  - Link records together and consolidate "correct" name based on other factors, such as data source, timestamp, etc.

Data Enhancement

- Data improvement process that relies on record linkage
- Value-added improvement from third-party data sets:
  - Address correction
  - Geo-Demographic/Psychographic imports
  - List append
- Typically partnered with data providers

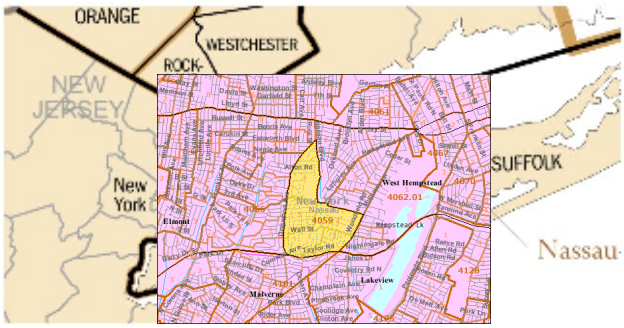
Enhancement: USPS Address Standardization

- Multiple address lines
- Recipient line
- Delivery Address line
- Last line



Address Standardization

- First: Is the address already in standard form?
  - Address special cases (East West Hwy)
  - Identify all addressing elements
  - Make sure placement is correct; if not, correct it
  - Are street and city names valid?
  - Is the address number valid within the street address ranges?
- Next: Correct if necessary
  - Identify all address elements
  - Look up proper city name
  - Look up correct ZIP+4
  - Move elements to proper location in address block
  - Transform elements into standard abbreviated form
  - Generate bar code (if needed)

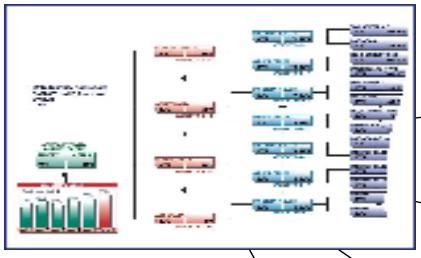


Subject	Number	Percent
OCCUPANCY STATUS		
Total housing units	**1,704**	**100.0**
Occupied housing units	**1,681**	**98.7**
Vacant housing units	**23**	**1.3**
Tenure		
Occupied housing units	**1,681**	**100.0**
Owner-occupied housing units	**1,566**	**93.2**
Renter-occupied housing units	**115**	**6.8**

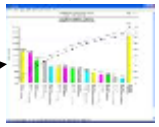
Subject	Number	Percent
SCHOOL ENROLLMENT		
Population 3 years and over enrolled in school	**1,571**	**100**
Nursery school, preschool	142	9
Kindergarten	102	6.5
Elementary school (grades 1-8)	662	42.1
High school (grades 9-12)	335	21.3
College or graduate school	330	21
EDUCATIONAL ATTAINMENT		
Population 25 years and over	**3,438**	**100**
Less than 9th grade	122	3.5
9th to 12th grade, no diploma	249	7.2
High school graduate (includes equivalency)	996	29
Some college, no degree	674	19.6
Associate degree	180	5.2
Bachelor's degree	676	19.7
Graduate or professional degree	541	15.7
Percent high school graduate or higher	89.2	(X)
Percent bachelor's degree or higher	35.4	(X)

Auditing and Monitoring Data Quality Performance

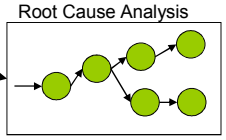
Data Quality Scorecard



Prioritizing Impacts



Root Cause Analysis



Productivity

Decision-making

Issues with Data Quality Tools

- Data quality is seen as a *technical* problem requiring a *technical* solution
- Disconnect between information value and achieving business objectives leads to ignoring DQ until it is too late
- Data is often corrected, instead of flawed processes
- There are opportunities for improvements in the data quality tools space…

Review and Thoughts

- 1st generation tools focus on "correction"
- 2nd generation tools look at analysis and discovery
- Possible next generation:
  - Standardized rules
  - Business impact correlation
  - Performance metrics
  - Semantic metadata
  - Generic metadata-based descriptive capability
  - Historical auditing and tracking

DATA QUALITY METRICS AND DATA QUALITY CONTROL

Expectations, Rules, Auditing, and Monitoring

- Data quality rules can be used to monitor conformance to data expectations as dictated by information policies
- Conformance can be measured, thresholded, and reported at each handoff location in the processing stream
- Specific failures can generate events as directed by Data Quality Service Level Agreements
- Static auditing: measurement applied to a "static" data set
  - Examples: SQL queries, data profiling tools
- Inlined monitoring: measurement performed within a process flow
  - Example: edit checks, dynamic monitors
- All measurements are compared against acceptability thresholds
- Acceptability threshold is related to the degree of impact

What Makes a Good Metric?

- Clarity of Definition
- Measurability
- Business Relevance
- Controllability
- Representation
- Reportability
- Trackability
- Drilldown Capability

Working with Subject Matter Experts

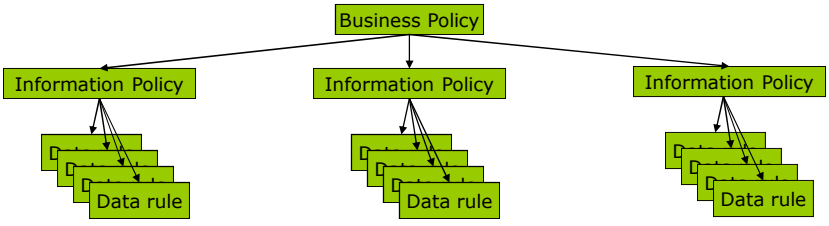
- Focus on information requirements, not functional or technical requirements.
- Directed Questions:
  - What are the functions of your business unit?
  - What are the Business Unit processes?
  - How is progress tracked and measured?
- Open-Ended Questions:
  - What are your information requirements? i.e. Performance reporting, operational management, financial management
  - Describe scenarios of how you currently use or envision using the data in the data warehouse

Metrics and Measurement

- Decompose information policies into specific measurable data rules
- Apply tools and techniques for measuring conformance to data rules (think: data profiling)
- Metrics can be "rolled up" from data rules defined as a by-product of analyzing the information policy
- Institute protocol for alerting key staff members when controls trigger data quality events
- Establish agreements for resolving issues within a reasonable time frame
- Monitor conformance to service level agreements

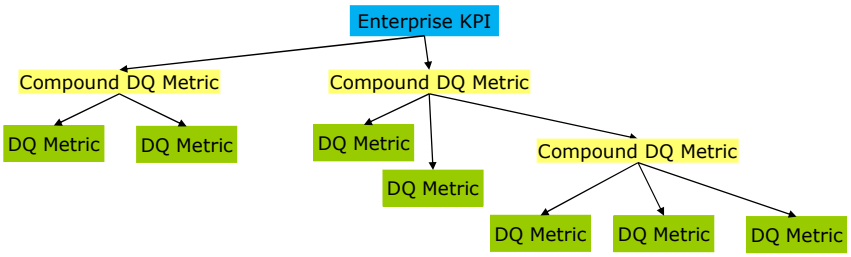
Monitoring and Evaluation

- One business policy can encompass multiple information policies
- Each information policy may encompass multiple data rules
- Each data rule, therefore, contributes to monitoring compliance with business policy!

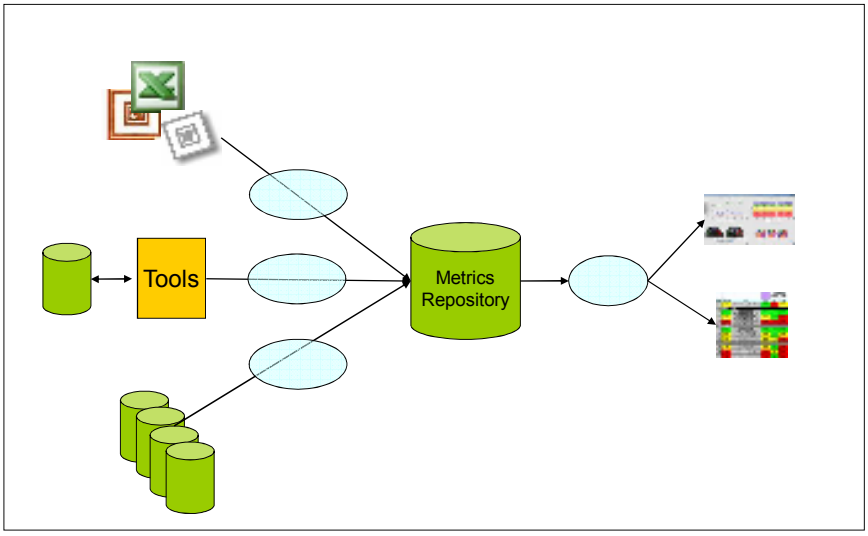


Hierarchies for Reporting

- Base-level metrics capture directly monitored scores
- Compound metrics have scores that are composed of rolled-up scores
- Collected scores can reflect conformance to business objectives



Monitoring and Reporting



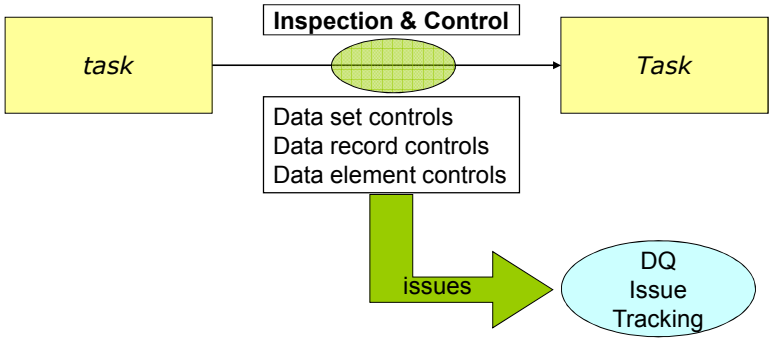
Gap Analysis Report

Requirement	Unmet Requirement	Gap Description	Recommendation
The name of the target data element whose criteria have not been met	*Description of the functional requirement not met*	*Description of the gap (e.g., no source data element met the data quality requirements based on the defined dimensions, or no source data element could be identified)*	*The suggested approach to be taken to resolve the gap*

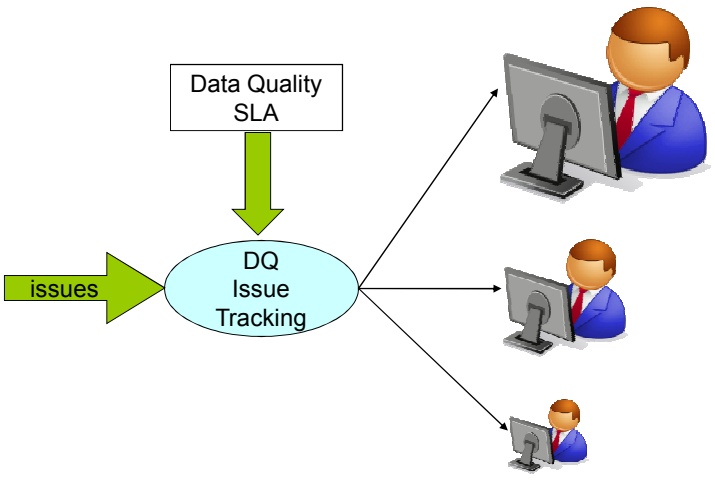
Reporting, Aggregating, Protocols for Metrics

- What is being measured?
- What are the roles of those reviewing the scorecard?
- What do the people in these roles need to know?
- How are the scores aggregated?
- What tasks are dependent on the scorecard?
- Reporting along different dimensions:
  - By *data set* (e.g., what is the overall quality score for "Customer_Address")
  - By *business process* (all impacts attributable to data issues within a business process)
  - By *issue* (all impacts across multiple business processes for the same issue)
  - By *impact* (all issues across multiple business processes with the same business impact)
  - By *organization* (by office, by region, etc.)
- How does one define acceptability thresholds
- How are data quality issues tracked?
- What actions are taken by the data stewards when issues are flagged?
- How are resolution workflows defined?
- How are Service Level Agreements agreed to?

Inspection & Control



Data Quality Control and SLAs



DQ Issue Tracking

- Alerts are generated when inspection shows that control indicates missed objectives
- Process to log and notify:
  - Description
  - Characterization
  - Prioritization
  - Routing
  - Start the clock…
- Measures added to metrics repository
- Resolution terms dictated by SLA

DQ Service Level Agreements

- An SLA is a "contract" between service providers and their customers detailing the specific services provided as well as levels of:
  - Availability
  - Performance
  - Operation
  - Cost
  - Duration/Time for resolution
  - How service levels are measured and tracked
  - Roles & Responsibilities
  - Metrics
  - Thresholds
  - When events are generated and notification strategies
  - Escalation strategy for identified issues

Questions?

- www.knowledge-integrity.com
- If you have questions, comments, or suggestions, please contact me

David Loshin
301-754-6350
loshin@knowledge-integrity.com