Data Quality Seminar
From business value to project management

Presented by:
Frank Dravis
VP Information Quality
Firstlogic, Who We Are...

- 20 years as an industry leader in data quality solutions
- Manufacturer of Information Quality and Postalsoft® suites
- 5,000+ customers worldwide
- Supplier of data profiling, cleansing, enhancing, matching, and solutions
- Sponsor of MIT’s TDQM program

Our solutions refine the fuel (information) that drives enterprise operations such as data warehousing, CRM, ERP, and Business Intelligence.
What We Will Cover Today...

• Building business value through Data Quality
  – Examples of doing just that
  – How to get started on a DQ project

• Data quality strategy development
  – Identifying the goals and then the requirements
  – Considerations of the six factors

• Data quality project management
  – Phases of a data quality project
  – Data stewardship
  – Data quality project risks
Information Quality (IQ) Curriculum

- IQ 102, Building Business Value through Data Quality
- IQ 201, Now That You Care
- IQ 202-6, Data Quality Technologies and Deployments
- IQ 301, Generations of Information Quality
- IQ 302, The Data Detective
- IQ 303, Data Quality Strategy
- IQ 304, Data Quality Project Management
- IQ 401, Householding for a Structured View
Building Business Value
Through Data Quality

IQ 102

Frank Dravis
VP Information Quality
Data Quality Defect Domain

- Duplicate Records
- Invalid Values
- Duplicate Attributes
- Business Rule Violations
- Mis-matched Data Between Multiple Source Systems
- Lack of Synchronization Between Source and Target
- Lack of Referential Integrity
- Defective addresses
- Invalid Date Dependencies
- Unreasonable Relationships
- Invalid Computations
- Invalid Ranges
- Non-standard Formats
- Duplicate Keys
- Poor Entity Construction
- Missing Data
- Dependencies
- Invalid Computations
- Invalid Ranges
- Non-standard Formats
The Importance of Data Quality
to CRM and Data warehouse deployments

Through 2005, over 50% of data warehouse and CRM deployments will suffer limited acceptance, if not outright failure, due to lack of attention to data quality issues (0.8 probability).

Gartner Inc.
T. Friedman
January 2002
Audience Survey Question

- After two years, are 50% of CRM and DW projects suffering limited acceptance, if not outright failure, due to lack of attention to data quality issues?

- How would you answer this question:
  - Yes
  - Maybe
  - No

Remember, Ted gave it a 80% probability.
Data Quality Builds Business Value By…

- Streamlines, shortens, and optimizes workflows
- Improves application user confidence and adoption
- Reduces risk of operational failures
- Enhances the output and results of BI, CRM, etc.
- Ensures customer interaction is uniform and trouble free
DQ Business Value - CRM

For those of you who voted “yes” or “maybe” to the survey question we get to the first case of DQ business value...

- Improved data quality has saved existing DW, CRM, ERP, BI and other applications from disuse, and worse: being scrapped.
DQ Business Value for Systems Design

Architectures and data modeling

• When:
  – Building a new data warehouse…
  – Migrating a legacy data set to a new data store…
  – Planning an ETL schema…

• How do you know what data is worth moving?

• Will legacy problems/defects be migrated?
  – Propagating the problems

• Does the data need restructuring, reformatting, cleansing, standardization?
  – How will these operations effect the final data model?
13 Different Patterns, 1 column

<table>
<thead>
<tr>
<th>Column name</th>
<th>Tested for</th>
<th>Query criteria</th>
<th># Rows read</th>
<th># matches</th>
<th>% matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>[F1]</td>
<td>PATTERNS RECognition</td>
<td>STRING (Matches)</td>
<td>40001</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>

**Data:**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ns2nas2n</td>
<td>4545</td>
<td>11.35</td>
</tr>
<tr>
<td>3na9n</td>
<td>2312</td>
<td>5.78</td>
</tr>
<tr>
<td>5n</td>
<td>1195</td>
<td>2.99</td>
</tr>
<tr>
<td>4n</td>
<td>855</td>
<td>2.14</td>
</tr>
<tr>
<td>7n</td>
<td>341</td>
<td>0.85</td>
</tr>
<tr>
<td>10n</td>
<td>296</td>
<td>0.74</td>
</tr>
<tr>
<td>3n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>an</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**F1 : 3na9n**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>IQI_DRP_FRQ_VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>212R286428375</td>
</tr>
<tr>
<td></td>
<td>212M900011080</td>
</tr>
<tr>
<td></td>
<td>201R202342506</td>
</tr>
<tr>
<td></td>
<td>617M161201784</td>
</tr>
<tr>
<td></td>
<td>202J360613138</td>
</tr>
</tbody>
</table>
DQ Business Value - Data Modeling

- Conducting a data profile (data quality assessment) provides invaluable information to the design process.
- Identifies data problems before they are migrated.
- Allows in the planning of cleansing activities prior to system builds.
- All of these can save many $$$ in rework of system architectures, delays in system deployments, and dissatisfaction of end-users.
While conducting research of our financial services customers we learned…

– A primary focus of DQ operations was to ensure:
  • Accurate, timely, and consistent customer communications
    • Prospectus’s, checks, securities info, monthly statements
    • Employ the same correct salutation, title, first name standard and address every time, no matter the contact
  • Some customer accounts totaled in the millions
  • To alienate these customers by demonstrating a careless attitude towards their identity or that of their family caused them to question the reliability and professionalism of the FSP.

– It costs much less to keep a customer than to acquire a new one.
Hard ROI and the Irony

• A manufacturer was shipping 3 million catalogs quarterly
  – Each catalog cost $2, including postage
  – The manufacturer then implemented our…
    • Cleansing, match and consolidation solution
    • Identified 50% of their catalog mailings were either duplicate, sent to the wrong addresses, undelivered, or sent to the wrong prospects, for an annual savings of $12 million.
  – Able to reduce catalog marketing expense by 50%
    • This was their chief means of advertising
  – Forbade us to mention the details of the ROI because…
    • It would expose a competitive differentiator to their competition
The Data Profiling ROI

• What is the value of knowing what data is bad and what is good?
  – Answer: Depends on what you do with the information
  – Halt a marketing campaign or know you can proceed

• However…
  – A medical products supplier began using our profiling tool
  – Analysis reporting to senior mgt increased
    • More customer data attributes to a greater depth was assessed
  – When management asked how could this be
    • The analyst reported what took them 2.5 days previously
    • Now only took 15 minutes.
    • An 80:1 productivity gain
Getting Started

• Once you have identified the potential business value, what then?
Start a Data Quality Project

First Steps on the Road

• Be aware that data quality is still a cultural issue
  – More people get it now, but not everyone

• Pick an information issue where pain is apparent

• Develop your strategy
  – Evaluate the six data quality factors
  – Plan a tiered approach
    • Point of capture, migration, maintenance, operational feeds
First Steps on the Road

• Pick a pilot project to start the initiative
  • A data profile/assessment is a good beginning
  • Followed by a root-cause analysis of the defects

• Find your sponsor

• Understand the perception gap

• Don’t assume you need a hard ROI
Data Quality Framework

Click arrows for more details

- **Consolidate**: Combines unique data elements from matched records into a single source.
- **Match**: Identifies duplicate records within multiple tables or databases.
- **Enhance**: Appends additional data increasing the value of the information.
- **Correct**: Verifies, scrubs, and appends data based on a set of sophisticated algorithms that work with secondary data source.

**Data Assessment**
- **Measure**: Quantifies the number and types of defects.

**Analyze**: Assesses the nature and cause of data defects.

**Parse**: Identifies and isolates data elements in data structures.

**Standardize**: Normalizes data values and formats according to your rules and that of third-party references.

**Data Cleansing**
Example of Data Cleansing

**Input record**

Maggie.kline@future ELECTRONICS.com Margaret Smith-Kline phd
FUTURE Electronics 5/23/99
101 6th ave
manhattan
ny
10012
001124367

**Output record**

- **Salutation:** Ms.
- **First name:** Margaret
- **Last name:** Smith-Kline
- **Postname:** Ph.D.
- **Match standards:** Maggie, Peg, Peggy
- **Company name:** Future Electronics
- **Address 1:** 101 Avenue of the Americas
- **City:** New York
- **State:** NY
- **ZIP+4:** 10013-1933
- **Email:** maggie.kline@future ELECTRONICS.com
- **SSN:** 001-12-4367
- **Date:** May 23, 1999
From here…

• From business value cases, and how to start a data quality project…
• We transition to data quality strategy development
Questions on Building Business Value…

Through Data Quality
Data Quality Strategy
IQ 303

Presented by:
Frank Dravis
VP Information Quality
Agenda

- Goals
- Stewardship
- Storage
- Data Flow
- Work Flow
- Continuous Monitoring
- Context

(firstLogic)
Reason 1: **Data Is Ignored**

“... CRM is about data -- customer, product, inventory and transaction.

…the vast majority of enterprises pay no attention to the data that will support investments and systems.

As such, enterprises must have a detailed understanding of the quality of their data -- how to clean it up, how to keep it clean...

Action Item: **Have a data quality strategy.** Devote one-half of the total timeline of the CRM project to data elements.

S. Nelson, J. Kirkby August 2001
Poor DQ Impedes Business Value

DM Review Readership Survey, September 2002

• What are the three biggest challenges of implementing a BI/DW project within your organization?
  – #1 Budget constraints 35%
  – #1 Data quality 35%
  – #3 Understanding and managing expectations 29%
  – #6 Time required to implement 22%
  – #11 Lack of personnel 17%

688 respondents, average annual revenue $5.26 billion.

Gantry Group, LLC
A Definition of Strategy…

• A cluster of decisions centered on goals that determine what actions to take and how to apply resources.

Therefore, data quality strategy is…

• A cluster of decisions centered on data quality goals that determine what processes to change, technologies to implement, and people to train.
Data Quality Goals

We need to know them

• All DQ Goals must support on-going functional operations, data management processes, or new information initiative like a:
  – New DW, CRM application, loan processing system

• Goal examples:
  – Reduce time to process quarterly customer updates
  – Cleanse & combine 295 source systems into 1 MCIF
  – Comply with Patriot act to identify customers
  – Determine vendor data fit for loading into ERP system
Characteristics of DQ Goals Can Be…

• The goals can be high or low level
  − Enterprise initiative, operational support, tactical cleansing

• Focused on workflow or process improvement
  − Half of all DQ projects are really for process improvement

• Focused on regulatory compliance
  − HIPPA, HMDA, OFAC, CASS, etc.

• Have little directly to do with DQ

• Be explicitly related to DQ

Goals shape requirements. DQ project requirements are covered in IQ 304, *DQ Project Management*. 
Inside a Data Quality Strategy

The Six Factors

1. Context: The type and usage of data
2. Storage: System architecture and platform
3. Data Flows: Opportunities to cleanse
4. Stewardship: People and organization
5. Work Flow: Process improvement
6. Continuous Monitoring: Quality must be sustained
Context Factor

The type and usage of data

- What type of data is being cleansed, and how will it be used?
  - Customer data
    - Names, addresses, phone #'s, Social Security #'s, etc.
  - Financial data
    - Dates, loan values, balances, titles, account 3’s
  - Supply chain data
    - Part numbers, descriptions, quantities, supplier codes, etc.
  - Telemetry data
    - Height, speed, direction, time, measurement type, etc.
Context Determines Type of Cleansing

• Specialized data demands specialized cleansing
  – Titles: VP, President, General Manager, Shoe Shine
    • What is the acceptable domain?
  – Dates: closing date versus first payment date
    • Which should be less?
  – Part numbers: Vendor A, B, C
    • What are the valid formats and range values
  – Addresses: 100 Cedar Haven NW CT
    • Is it correct? Does it exist?

• Determines:
  – The algorithms, coding, cleansing applications needed
Storage Factor

The data environment, system architecture and platforms

• Considers where the data physically resides
  – Platforms and operating systems
    • Mainframe, Unix, MS Windows
  – Data distribution: centralized or distributed
    • Locally, regionally, internationally
  – Source formats: homogenous or heterogeneous
    • Flat files, spreadsheets, RDBMS, mainframe non-relational
  – Applications: open or proprietary
    • Siebel CRM, Peoplesoft ERP, Oracle data mart, DB2 DW

• Determines connectivity options to the data
  – Identifies the type of DQ technology deployment
Deployment Options

- Batch applications
  - Auto or manual start
- GUI
  - interactive applications
- Web-enabled applications
  - Real-time
- Enterprise app. plug-ins
  - ERP/CRM/ETL integrations
- High level libraries
  - Custom applications
- Low level libraries
  - Custom applications
- ASP connections
  - Out-sourced functionality

- Internet Log Files
- RDBMS (all flavors)
- Mainframe Indexed Files
- ERP/CRM Systems
- Windows Flat Files
- AS/400 DB2
- Data Warehouse
- Data Mart
- Transactional data store
- Web Back-end database

Disparate or legacy data systems
Centralized Data warehouse
Operational data stores
Data Flows

Opportunities to cleanse

- Transactional updates
- Operational feeds
- Purchased data
- Legacy migration
- Regular maintenance
Data Flows and Lead Generation Work Flow

**Data Flows**

- **Obsolete, home-grown Call Center**
  - Legacy Migration
  - CRM System

- **Purchased Lists**
  - Operational Feeds
  - Transactional Updates

- **List of Attendees**
  - Raw Leads
    - Collect Leads
    - Store Leads
  - Contact Records
  - Qualify Leads
  - Distribute Leads
  - Engage Prospects

- **Qualified Prospect records**
- **Sales Prospect Lists**
- **Active Prospect Records**
- **Operational Feeds**
- **Maintenance**

**Lead Generation Work Flow**

- **Purchased Lists**
- **List of Attendees**
- **Collect Leads**
- **Store Leads**
- **Qualify Leads**
- **Distribute Leads**
- **Engage Prospects**

- **Operational Feeds**
- **Transaction Updates**

- **Legacy Migration**
- **CRM System**

- **Obsolete, home-grown Call Center**
A Definition of Information Quality is…

• … the integration of people, process, and technology in pursuit of deriving value from information assets

We must consider the next three factors in our strategy that are centered on people and processes.

Whereas the previous three factors were focused the data and physical systems.
Stewardship Factor

*The people*

- We need to answer these questions
  - Who are the stakeholders of the data?
    - How will they be notified of pending changes
  - Who are the current and new data stewards?
    - What education and training will they need?
    - How will pending changes impact organizational structure?
    - Will roles and responsibilities change?
    - Incentivization and accountability should change
  - Who are the current and new consumers of the data?
    - Beware that consumers and stewards can be the same people
    - How will the processes and technology impact them?
Stewardship Factor

We need to emphasize

• People are the most important part of strategy
  – Data is nothing but food for human consumption

• Deal with cultural and political issues with education
  – Diligently work the benefits of the project
  – Plan an internal PR campaign to advertise the benefits
    • Visit all stakeholders repeatedly. Educate them at every opportunity.
    • Do not under estimate the need for and value of repeatedly advertising the benefits of good data in your project.
Work Flow

- The sequence of physical (human) work tasks
- Catalog and diagram (flow chart) the work activity
  - Only the activities that affect or are affected by the data in question
  - Helps visualize how data and work flows interact
- You’ll know where and when you need:
  - Front-office transaction, real-time cleansing
  - Back-office transaction, delayed cleansing
  - Back-office batch cleansing
  - Cross-office enterprise application cleansing
  - Continuous monitoring and reporting
Example: Lead Generation Information Chain

Data Entry
Real-time

Point of Capture
Real-time

Billing, Shipping, Product reports
Manual Batch

Matching, Consolidation, and data appending.
Manual Batch

Enter Lead (data)

Qualify Lead

Qualification Center

CRM

Lead

Prospective Customer

Customer Data

Information Extraction
Enterprise App Plug-in

Contact Management
Custom Application

Sale

Matching (Leads to territories)
Automated Batch

Notify Sales of Lead

Sales learns lead information

Sales contacts lead and enters into sales process

Example: Lead Generation Information Chain

Work Flow Touch Points
DQ Deployment Option
Continuous Monitoring
Principles of Total Quality Management
Continuous Monitoring

• How will you monitor the data over time?
  – Data cleansing is not a one-time activity
  – Data ages constantly, and irrevocably
  – Corrupt data feeds are unexpected and insidious

• Need to identify DQ measurements and metrics
  – Define specific measurements that aggregate into divisional, business unit, or corporate metrics
  – Plan monitoring process to capture measurements
  – Plan DQ Intranet Web site to host DQ dashboard

• Continuous monitoring starts with a baseline
To Aid in Strategy Formulation…

**Conduct a baseline Assessment**

**Baseline Assessment**
Comprehensive analysis of the targeted data set(s).
Produces an inventory and magnitude of specific data defects and an overall assessment of the data quality condition.

**Cleansing Evaluation**
Focused analysis to evaluate the effectiveness of data cleansing on identified baseline defects and target further cleansing activities.

**Continuous Monitoring**
Measures improvement or degradation over time. Reuses baseline and cleansing evaluation tests.
Continuous Monitoring

The Benefits

- Provide periodic reports on data quality indicators
  - What gets measured gets done.
- Quantify the effectiveness of data improvement actions
- Identify which actions are/are not altering the data quality conditions
- Continually reinforce the end users’ confidence in the usability of the data
- Identify deterioration in data quality early in the trend

Leverages tests development in the baseline and runs them automatically in a production environment.
Tying the Six Factors Together

• Context tells you the type of cleansing algorithms you need
• Context + storage + data flow + work flow dictates technology implementations
• Stewardship + work flow determine near-term personnel impacts
• Stewardship + work flow + continuous monitoring determine long-term personnel impacts
• Data flow + work flow + continuous monitoring outlines changes to processes

“…integration of people, process, and technology in pursuit…”
A DQ Strategy Template

Contains these sections, and reflect consideration of the six factors

• A statement of the goals driving the project
• A list of data sets/elements that support the goal
• A list of the data types to be cleansed
• A system catalog of where the data resides
• A description of cleansing solutions needed per data type
  – Includes selected deployment options
• Data flow diagrams of
  – Applicable, existing data flows
  – The new improved data flows
• An explanation of why the data flow is changing
A DQ Strategy Template

Continued… Strategy is implemented as a series of linked tasks

- Work flow diagrams of
  - Applicable, existing work flows
  - The new or improved work flows with DQ touch points
- An explanation of why the work flows are changing
- A list of personnel and groups affected
- A section on educating and training the affected groups
- A plan for initial and regularly scheduled cleansing
- A continuous monitoring plan including
  - A baseline assessment
  - Measurements to capture, metrics to report
Where Strategy Ends…

Project management takes over.
- Responsible for implementation of strategy

In the next section we will cover

DQ Project Management
Questions

Frank Dravis
Frankd@firstlogic.com
www.firstlogic.com, 001.608.782.5000
Data Quality Goals

We need to know them

• All DQ Goals must support on-going functional operations, data management processes, or new information initiative like a:
  – New DW, CRM application, loan processing system

• Goal examples:
  – Reduce time to process quarterly customer updates
  – Cleanse & combine 295 source systems into 1 MCIF
  – Comply with Patriot act to identify customers
  – Determine vendor data fit for loading into ERP system

From IQ 303, Data Quality Strategy
The Six Factors of Data Quality Strategy

The Six Factors

1. Context: The type and usage of data
2. Storage: System architecture and platform
3. Data Flows: How data moves and changes
4. Stewardship: People and organization
5. Work Flow: How and where people access the data
6. Continuous Monitoring: Regression checking
Output of the Strategy Process: The Plan

• The program plan
  • Serves as the project charter document  
    – Facilitates understanding between IT, business, the client, and the project team
• Defines…
  – The problem, goals, objectives, scope, data flows, work flows, stakeholders, etc.
A DQ Strategy Template

Contains these sections, and reflect consideration of the six factors

• A statement of the goals driving the project
• A list of data sets/elements that support the goal
• A list of the data types to be cleansed
• A system catalog of where the data resides
• A description of cleansing solutions needed per data type
  – Includes selected deployment options
• Data flow diagrams of
  – Applicable, existing data flows
  – The new improved data flows
• An explanation of why the data flow is changing
A DQ Strategy Template

Continued… Strategy is implemented as a series of linked tasks

• Work flow diagrams of
  – Applicable, existing work flows
  – The new or improved work flows with DQ touch points

• An explanation of why the work flows are changing

• A list of personnel and groups affected

• A high level plan for initial and regularly scheduled cleansing

• A high level plan for continuous monitoring including
  – A baseline assessment
  – Measurements to capture, metrics to report
For Project Management Experts…

• All the standard elements in project management apply, such as costing, scheduling, risk assessment, contingency planning, reporting, etc.

• The purpose of this presentation is to identify specifically how a data quality project will vary from a generic project.
Building Blocks of a DQ Project

- Data
- Goals
- People
- Processes
- Requirements
- System resources
- Program plan (strategy document, program charter)
- Project plan (task list, estimates, schedule, resource assignments)
- Software (custom, COTS, in-house)
Phases of a DQ Project

1. Requirements Formulation
2. Requirements Review
3. Task/Resource Estimates
4. High-Level Schedule
5. Project Approval
6. Assess Data
7. Adjust per assessment findings
Phases of a DQ Project

1. **8 Data Preparation**
   - 8a Data model Redesign
   - 8b Data extract/load

2. **9 Tools preparation**
   - 9a Tool evaluation
   - 9b Tool procurement
   - 9c Tool Installation
   - 9d Personnel Training
   - 9e Custom programming

3. **10 Process redesign**
   - 10a Record parsing
   - 10b Standardization
   - 10c Correction
   - 10d Enhancement
   - 10e Matching
   - 10f Consolidation

4. **11 Cleansing Operations**
   - 11a Record parsing
   - 11b Standardization
   - 11c Correction
   - 11d Enhancement
   - 11e Matching
   - 11f Consolidation

5. **12 Cleansing evaluation**

6. **13 Data in production**

7. **14 Continuous monitoring**

---

Project monitoring
Applying the Wheel

- Ultimately, every DQ project must deal with specific DQ functions as represented by the IQ framework wheel.
- Your project should ultimately organize the DQ tasks and cleansing operations to flow in this sequence.
- These actions build one upon the other, making each step more effective.
- Your project should have elements that focus on the unique features of each of these 8 operations.
A data quality project will employ these operations.
Some Typical DQ Projects

- Data assessment (profile) on suspect data
- Data migration cleansing (legacy data)
- Regular operational data flow cleansing (ETL)
  - On back-end Web sites or data entry systems
- Transaction data cleansing (plugging the leaks)
- Maintenance cleansing
- Data warehouse/data mart build preparation
- Validation and reporting of purchased data
- Continuous monitoring
- Data analysis/correction of a failed business operation
IQ Pilot Map, Process Flow

Using a marketing campaign as an example

Decision: How to Improve Campaign result

Correct the data

Reporting: defects found in customer data

Data inspection 1st Pilot

Discovery: what happened?

DQ Assessment

Organizational process: Marketing Campaign

Action: Identify top 20% of Customers

Poor Result: Response rate too low

Data cleansing Operation: 2nd Pilot

Solution Install/Testing

Solution Build/Procurement

Data

Metadata

Metadata repository

DQ Solution Research

DQ project planning

Pilot projects divide and conquer enterprise behemoths
IQ Pilot Map

1. Requirements formulation
2. Requirements review
3. Task/Resource Estimates
4. High level schedule
5. Project approval
6. Assess data
7. Adjust per assessment
8. Data preparation
9. Tools preparation
10. Decision: How to Improve Campaign result
11. Correct the data
12. Action: Identify top 20% of Customers
13. Data in production
14. Continuous monitoring

Data Cleansing Operation: 2nd Pilot

Data Cleansing Operation: 1st Pilot

Metadata

Metadata repository

Poor Result: Response rate too low

Discovery: what happened?

Data inspection

DQ project planning

Organizational process: Marketing Campaign

Data in production

Solution Install/Testing

Solution Build/Procurement

Reporting: defects found in customer data

Decision: How to Improve Campaign result

Correct the data

First Logic
People: Stakeholders, Roles, and Data Stewardship
Stakeholders…

Who needs to be aware of the project?

• Managers of the data: suppliers and administrators
  – Do they know the data is defective and how to fix it?

• Consumers of the information: reports, queries, operations
  – Do they know the information is defective and the impact?

• Consumer (information) management
  – Inform them of impact, and gain preliminary buy-in for solution

• Subject matter experts (if not encompassed by the above)
  – Do they know the data is defective and what it should be?

• IT personnel responsible for the system maintenance
  – Crucial for solution planning, install, testing, and support

• Senior management
  – Inform them of results, and recommendations
DQ Project Team Roles

- Project sponsor (authorization)
- Client representative (the customer/consumer)
- Project manager (organization and tracking)
- Business analysts (assessors)
- Data analysts (cleansing operations)
- Data administrators (source system support)
- Systems analysts (network/hardware support)
- Data steward (primary responsibility for the data)
What is a Data Steward?

*Has one, some or all of the following attributes*

- Typically not a formal “job” but a concept
- Someone that knows and understands the data
- Manages the creation process of the data
- Manages the updates and extracts of the data
- Manages the maintenance processes on the data
- Accountable for quality of the data
- Can be either a IT or business function
Data Steward Responsibilities

- Resolving data integration issues
- Maintaining/updating business rules
- Analyzing and improving data quality
- Maintaining/updating the data model and schemas
- Documenting data definitions, calculations, summarizations, etc.
- Ensuring alignment of the business requirements, data contents, and IT support systems
- SME for the contents of the data and how it can be used
A DQ Project Plan Template is based on and references the DQ strategy planning document.

- List of requirements and sign-off mechanism
- List of deliverables
  - From status reports to cleansed data warehouses
- List of milestones
- Estimated task durations
  - Creates a high-level schedule
- Cost estimates for...
  - IT programming, outside consulting, software/hardware purchases or upgrades
- Proposed personnel assignments and responsibilities
A DQ Project Plan Template (cont.)

• Reference to data scope (from strategy planning)
• List of data preparation activities
• Tools procurement/custom coding plan
• Standard operational processes that need redesign
• A detailed plan for specific cleansing operations
• A refined plan for continuous monitoring
• List of project assumptions
• List of risks
• Critical success factors
  – Matched against the goals and requirements
Project Monitoring and Reporting

- Event triggers (E-mail, cell phone alerts)
- DQ dashboards (Intranet-based)
- Graphical reports of data issues
- Project status reports
- Cleansing tool output reports
- Assessment findings
- Measurements and metrics to collect
DQ Dashboard

Indicators

- Percent Incorrect or Missing Address Data
- Data Quality Confidence Score
  As of: Mar 28, 2003 2:36:20 PM
- Non-Correlated Account Numbers
  As of: Mar 28, 2003 2:06:52 PM

Monthly Data Error Rate

- Data Quality Error Types
  - 24.30%
  - 24.94%
  - 18.88%
  - 31.01%

Data Error Volumes by Fiscal Quarter

http://s1571647001aes/jsp/analyzeResults.jsp?id=221208
Assessment Findings: Exception Reports

Shows all records with a blank ID code

<table>
<thead>
<tr>
<th>FLF</th>
<th>MEMPHIS INTL</th>
<th>MEMPHIS</th>
<th>MEMPHIS</th>
<th>MEMPHIS</th>
<th>UNITED</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLF</td>
<td>LAWRENCE MUNI</td>
<td>LAWRENCE</td>
<td>LAWRENCE</td>
<td>LAWRENCE</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>PALO ALTO ARPT</td>
<td>PALO ALTO</td>
<td>PALO ALTO</td>
<td>PALO ALTO</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>MERRILL FIELD ARPT</td>
<td>ANCHORAGE</td>
<td>ANCHORAGE</td>
<td>ANCHORA</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>CRISTAL ARPT</td>
<td>CRISTAL</td>
<td>CRISTAL</td>
<td>CRISTAL</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>CHICAGO OHARE INTL</td>
<td>CHICAGO</td>
<td>CHICAGO</td>
<td>CHICAGO</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>TETERBORO ARPT</td>
<td>TETERBORO</td>
<td>TETERBO</td>
<td>TETERBO</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>WUTHAM FIELD ARPT</td>
<td>STUART</td>
<td>STUART</td>
<td>STUART</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>REDDING MUNI ARPT</td>
<td>REDDING</td>
<td>REDDING</td>
<td>REDDING</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>LAKELAND LINDER</td>
<td>LAKELAND</td>
<td>LAKELAND</td>
<td>LAKELAND</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>LAMBERT-ST. LOUIS ARPT</td>
<td>BRIDGETO</td>
<td>BRIDGETO</td>
<td>BRIDGETO</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>BEVERLY MUNI ARPT</td>
<td>BEVERLY</td>
<td>BEVERLY</td>
<td>BEVERLY</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>CLEVELAND-HOPKINS</td>
<td>CLEVELAND</td>
<td>CLEVELAN</td>
<td>CLEVELAN</td>
<td>UNITED</td>
</tr>
<tr>
<td>FLF</td>
<td>COLUMBUS METRO</td>
<td>COLUMBUS</td>
<td>COLUMBU</td>
<td>COLUMBU</td>
<td>UNITED</td>
</tr>
</tbody>
</table>
Measurements Collected: Trend Reports

Display and compare trends of multiple measurements.

Dev Area Visibility Codes N and P

Count


N = Blue, P = Red

(firstLOGIC)
DQ Project Risks

- Be psychologically prepared for them.

- Poorly defined business rules
  - Matching, column validations, definitions, etc.

- Poorly defined scope
  - Data sources, tables, columns get added at last minute

- Data not as specified
  - Part numbers that span two columns instead of one

- Additional cleansing tasks
  - You were so good at cleansing addresses, let’s apply GEO codes!
  - Good for the enterprise, but watch schedule impacts…
DQ Project Risks

• System resources over-taxed
  – Too slow to meet cleansing schedules

• Subject matter experts missing in action
  – Need a password entered? A software driver installed?

• System failures (acts of God)
  – Ever see a consultant trip over a power cord for a server

• Conflict between business and IT communities
  – Can paralyze a project if a decision is needed
In Closing, Changing Perceptions

- Report the progress. Communication is key.
- Reaffirm to the consumers…
  - Your commitment to their information
  - So the users will have confidence in the system and use it.
- Publicize the objectives of the project
  - What data was corrected, and future plans
  - Thus armed with information on what is validated and what is not, the users can make informed decisions as to how to use the data, and not just write it off.

A data quality project truly is the integration of people, process, and technology in pursuit of deriving value from your information assets.
Questions

Frank Dravis
Frankd@firstlogic.com
www.firstlogic.com, 001.608.782.5000