

## Experiences in Data Quality

### ABSTRACT

---

The MITRE Corporation is committed to helping the Federal government manage its data as an enterprise asset and make the best use of appropriate technologies and services in industry and academia. This presentation will provide an overview of the concepts, methodologies, and policies recommended for facilitating data sharing across and within Federal government agencies and why the quality of data and information is so important decision makers, practitioners and end- users. The presentation will discuss the continuous improvement of data quality by addressing the data quality concerns in legacy systems, system migrations, new systems, system upgrades and the adoption/utilization of SOA, MDM and Cloud Computing. Business implications of poor quality data, along with roles and responsibilities will be discussed as they relate to decision making and data stewardship. MITRE's experience with various agencies will be presented as example. Also discussed will be some MITRE evaluation results of data cleansing, profiling tools and appropriate environments for their application.

### BIOGRAPHY

---

#### **Annette Pence**

The MITRE Corporation

Annette Pence has more than 30 years of experience with the application of information technology – enabling solutions supporting aerospace, finance, classified command and control, and manufacturing. Her experience managing data spans the entire data lifecycle and includes system design, development, and integration. Ms. Pence is the head of the Information and Data Management Department at The MITRE Corporation's Center for Connected Government. As a Senior Principal, Information System Engineer, with the MITRE Corp., Ms. Pence is responsible for the strategic development and delivery of a targeted set of objectives for government customers, which includes the IRS, DHS, HHS, GPO, Department of Education, HHS and Centers for Medicare and Medicaid. Prior to MITRE, Ms. Pence served as Vice President, Data Management Services Division at Science Applications International Corporation (SAIC), Department Manager, Software Integration and Data Management, for the Reserve Component Automation Systems (RCAS), at Boeing Information Services, and Manager Database Administration for the National Aeronautics and Space Administration, Technical and Management Information Systems (TMIS). Her focus is raising the awareness and furthering the implementation of data management at the enterprise level for government agencies and enabling and facilitating the sharing of information across government agencies. A native of Washington, D.C., Ms. Pence earned a master's degree in Technology Management from George Mason University. During the Technology Management curriculum she was fortunate to study at the Kellogg College at Oxford University.

# Experiences in Data Quality

MIT IQIS 2010

Annette Pence  
July 14 - 16, 2010

Approved for Public Release: 10-0686

Distribution Unlimited

© 2010 The MITRE Corporation. All Rights Reserved.

**MITRE**

**As a public interest  
company, MITRE  
works in  
partnership with  
the government to  
address issues of  
critical national  
importance.**

**Apply Systems  
Thinking to Enable  
Government  
Effectiveness**



## MITRE is an Operator of FFRDCs

The MITRE Corporation operates four FFRDCs, each under the sponsorship of a government organization.

**Command, Control,  
Communications,  
and Intelligence (C3I)  
FFRDC**

*Sponsored by the  
Department of  
Defense  
(1958)*

**Center for  
Advanced  
Aviation System  
Development**

*Sponsored by the  
Federal Aviation  
Administration  
(1990)*

**Center for  
Enterprise  
Modernization**

*Co-Sponsored by  
the Internal Revenue  
Service and  
Department of  
Veterans Affairs  
(1998)*

**Homeland Security  
System Engineering  
and Development  
Institute (HS SEDI™)**

*Sponsored by the  
Department of  
Homeland Security  
(2009)*

The work performed by an FFRDC is defined by a **Sponsoring Agreement**, unique to each FFRDC, that describes the context in which work is to be performed.

MITRE

3

© 2010 The MITRE Corporation. All Rights Reserved.

## Center for Connected Government (CCG)

CCG houses the majority of MITRE's civilian agency work, and includes both SEDI and CEM FFRDCs.

- Center for Transforming Health (CTH)
- Center for Enterprise Modernization (CEM)
- Homeland Security Center (HS SEDI™)



MITRE

4

© 2010 The MITRE Corporation. All Rights Reserved.

## Currently Supported Federal Agencies

5

© 2010 The MITRE Corporation. All Rights Reserved.

## The Challenge

*Every government agency is data management challenged*

- Getting the right data to the right person
- Timely, dependable data access
- Massive information to store, manage, and access
- Stored within stove-pipe, application-centric data stores
- Data duplicated across multiple environments with unclear data definitions
- Data policies, responsibilities, and standards
- Characterized by questionable or uncertain data quality

↓  
 The Lack of Appropriate Data Quality Management

*Adversely impacts program and mission performance*

- Mission compromised by sub-optimal data management
- Increased cost
- Diminishes public confidence
- Minimizes information sharing
- Impacts service to citizens; creates inaccurate picture of agency performance

**MITRE has a responsibility to address these challenges**

6

© 2010 The MITRE Corporation. All Rights Reserved.

## What is Data Quality Management?

*“Planning, implementation and control activities that apply quality management techniques to measure, assess, improve and ensure the fitness of data for use.” [DMBOK]*

■ **Simply stated:**

- Accuracy – correctness of data values for their intended purpose
- Completeness – degree of inclusion of values present in a data set
- Consistency – conformance of data values to formats and constraints
- Timeliness – appropriateness of data use at a specific time
- Validity – extent that data values conform to specified acceptance criteria

---

**There are different levels of complexity and some components are more integral to some environments than others**

## Why, How, Who

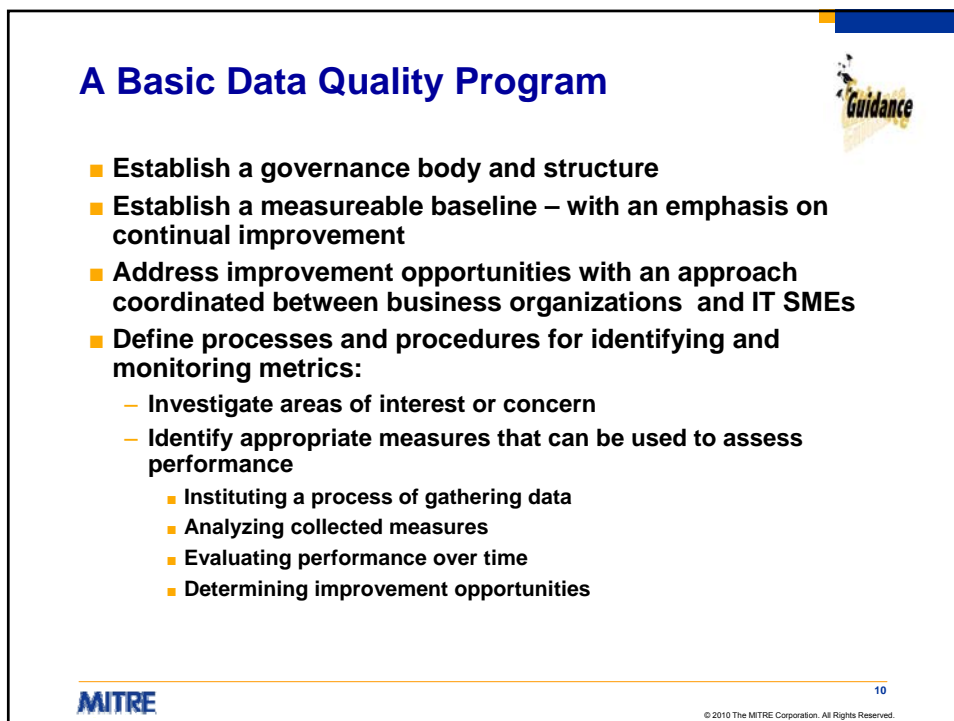
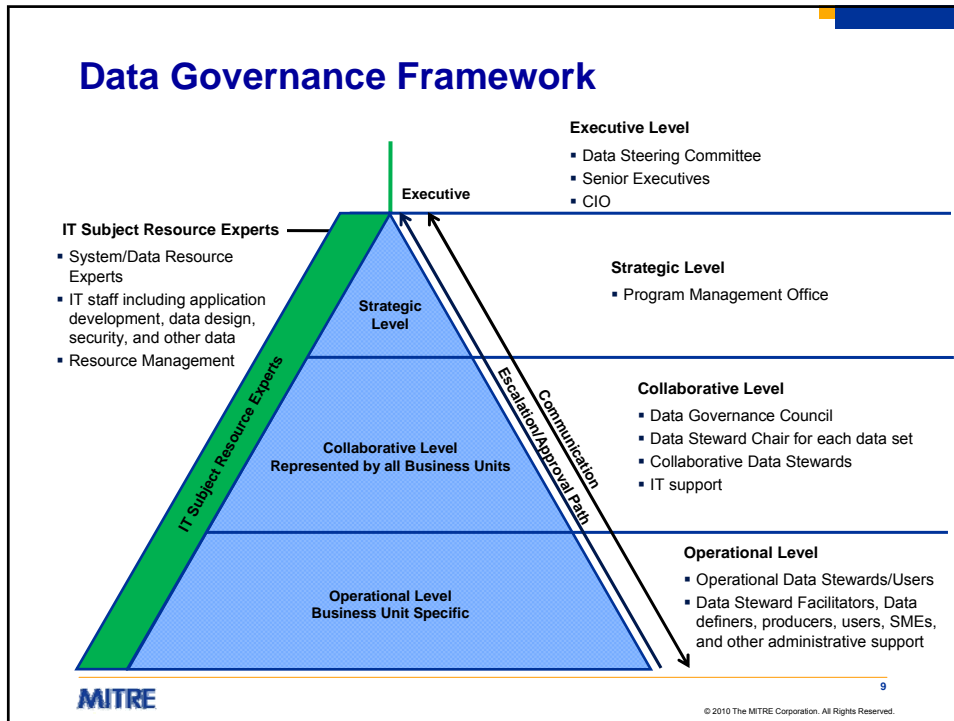
■ **Easy to explain why**

- Lower costs
- Better information for decision making
- Dependability for information sharing
- Easier to explain to technical and business owners than executives, but executives make funding decisions

■ **Harder to explain how**

- Standards
- Processes
- Tools

■ **Lots of politics regarding who does what, knows what, owns what, makes the decisions, is responsible ...**



# Back in Time



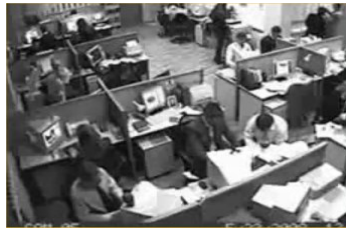
**Still relying on a few people who know the rules**



**Legacy system  
System migrations**



**Users that don't use system and keep their own files and records**



**KNOWLEDGE IS POWER**

**Legacy**

**Newbies**

**It'll get fixed in the database**

**MITRE**

13

© 2010 The MITRE Corporation. All Rights Reserved.

## There are Data Quality Problems Because...

- Data quality was relegated to cleansing and transformation
- Lots of undocumented processes
- No documented, understood, or acknowledged management processes
- Every man is an island
- Lack of implemented standards for data exchange
- Nobody had responsibility



## The Impact to the Agency's Business is...

- Problems paying bills
- Reports produced that are never used...wasted \$\$
- Multiple decisions made on different versions of the same information
- Inability to share information
- Limited trust

## Outcome

- Governance structure
  - Owner of quality
  - Approval authorization
  - Quality council
  - Data managers
  - Error reporting (specifically data) during system migrations and upgrades
- Some standards (mostly naming conventions)
- Some processes for determining consistency and validity
- Data definitions that could be shared
- Data element-to-business process mapping

*This was hard – they still have a really long way to go*

# We Just Need a Tool



I'm on top of  
data quality!

We've got the latest  
software!



We have the latest  
equipment!



We're state of the art  
here!





- Everybody knows there are rules, they're documented
- We make sure everything gets reviewed
- Different organizations have different procedures for data quality
- It's too far down in the weeds to look at each data element
- We inspect all data before it's used

***“WE JUST NEED  
A TOOL!”***



MITRE

19

© 2010 The MITRE Corporation. All Rights Reserved.

## So Tell Me:

- What do you want the tool to do?
- Who is going to own this tool?
- Who can use this tool?
- What about existing roles and procedures?



- Are there policies that drove the rules?
- Is there an overall plan for quality that defines completeness and accuracy in this environment?
- What do you do with all the review and inspection results?

MITRE

20

© 2010 The MITRE Corporation. All Rights Reserved.

## Outcome


- Established a quality council that included business owners, data stewards, and technicians
- Some education on best practices
- Guidelines for the organization
- Responsibility for suppliers of data
- Some perspective on the end users' view of the data
- Metrics and feedback to suppliers
- They did need a tool to automate manual labor intensive analysis processes
  - Profiling, defect detection, defect prevention
- Some team building – integrated teams

*A lot to work with on this one – just had to drag it out of them*


## IBM Information Server (IIS)

- Information Analyzer
  - Discovery
  - Profiling
  - Analysis
- QualityStage
  - Parsing
  - Standardization
  - Matching
- Integrated with IBM InfoSphere platform

## IBM InfoSphere Information Analyzer



**Subject Matter Experts**

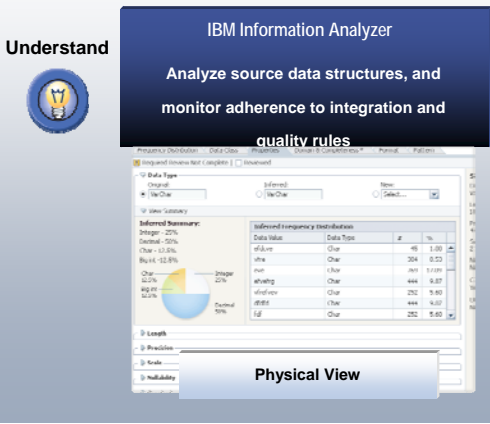


**Data Analysts**


**Understand**

**IBM Information Analyzer**

Analyze source data structures, and monitor adherence to integration and quality rules




- Column Analysis
- Completeness
- Consistency
- Pattern Consistency
- Frequency
- Primary Key
- Primary Key Analysis
- Single or Multicolumn
- Duplicate Analysis
- Foreign Key
- Foreign Key Analysis
- Duplicate Analysis
- Referential Integrity
- Cross Domain
- Redundancy Analysis
- Baseline




Page 23

© 2010 The MITRE Corporation. All Rights Reserved.

## IBM InfoSphere Quality Stage



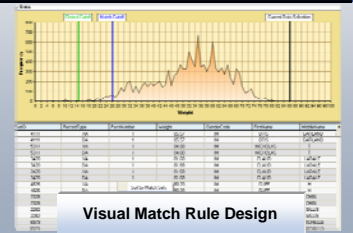
**Subject Matter Experts**



**Data Analysts**

**Cleanse**

**Standardize and correct source data fields, and match records together across sources to create a single view**



**Investigation**

Understand nature scope and detail of data quality challenges

**Standardization**

Ensure that data is formatted and conforms to organization wide standards

**Matching**

Identify duplicate records within and across data sets

**Survive**


Eliminate duplicate records and creating the best record view of the data

Investigate

Standardize

Match

Survive



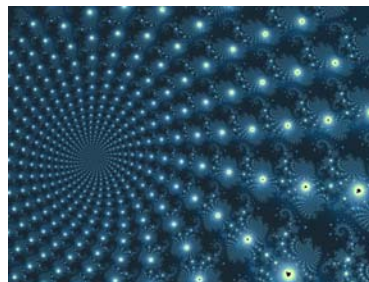
24

© 2010 The MITRE Corporation. All Rights Reserved.

## An Enterprise Approach

## A Very Mature Environment

- Lots of good processes and procedures
- Well developed data management organization
- Data quality management success in isolated areas
- Determined data quality as a critical need
- Willing to plan and implement holistic program



## How Does It All Come Together?

- **Data Quality Toolkit**
  - Profiling
  - Cleansing
  - Defect detection
  - Defect prevention
  - Transformation
  - Matching
- **Database Server**
  - Business rule implementation
  - Referential integrity constraints
  - Table level constraints
  - Column level constraints
- **Application Server**
  - Business rule implementation
  - Application logic
  - Column validation
  - Context validation
- **Application and Database IT support**
  - Business logic support
  - Software feature and capability implementation
  - DBMS feature and capability implementation
  - Toolkit utilization

## What About SOA, is That Going to Mess Us Up?

- We have lots of data and lots of integration
- We need to be able to communicate what good quality data is
- Sometimes we have to clean it up after it's used before it's used again



What's going to happen when we implement Master Data Management?



What should we do about transparency and having to share?

## Outcome

- Leverage the existing data management governance structure
- Develop a data quality plan and approach
- Manage your metadata!
- Develop and use dashboards
- Formally define, implement, and communicate roles for data stewards
- Get training:
  - Quality philosophies and practices
  - Data entry criteria and submission importance
  - Identifying requirements, data availability, and proper utilization of data
  - Implementation of business rules in database and application software configurations

*A data person's dream environment – able to do what we do best*

## Best Practice Recommendations

- One size does not fit all
- Determine your needs
- Get the appropriate tool for your needs
- Get executive buy-in
- Allocate enough time for training
- Adopt a framework for conflict resolution and decision making (quality council)
- Understand data stewardship
- Identify, enable, and use data stewards
- Don't try to implement a data quality program without understanding the data requirements
- Think in phases



**Thank You and Good Luck**

**Annette Pence**

**The MITRE Corporation**

**Senior Principal Information Systems Engineer**

**Information and Data Management Department Head**

**703-983-6098**

**[apence@mitre.org](mailto:apence@mitre.org)**