Experiences in Data Quality

MIT IQIS 2010

Annette Pence

July 14 - 16, 2010

Approved for Public Release: 10-0686

Distribution Unlimited

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

MITRE is an Operator of FFRDCs

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

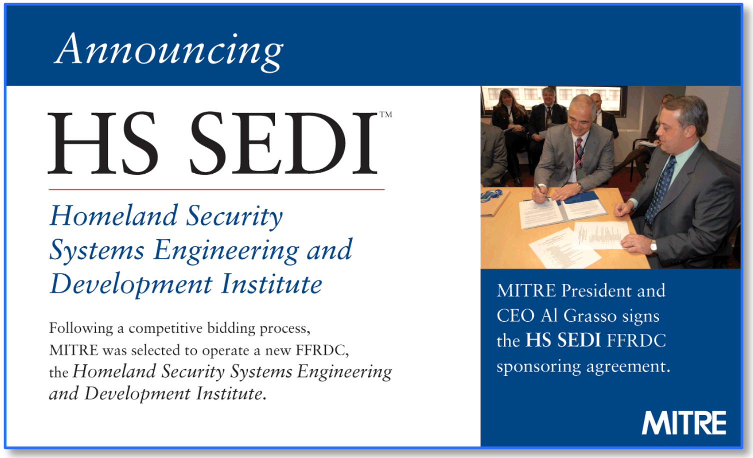
Command, Control, Communications, and Intelligence (C3I) FFRDC	**Center for Advanced Aviation System Development**	**Center for Enterprise Modernization**	**Homeland Security System Engineering and Development Institute (HS SEDI^{TM})**
Sponsored by the Department of Defense (1958)	***Sponsored by the Federal Aviation Administration (1990)***	***Co-Sponsored by the Internal Revenue Service and Department of Veterans Affairs (1998)***	***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.

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^{TM})



Announcing

HS SEDI™

Homeland Security Systems Engineering and Development Institute

Following a competitive bidding process, MITRE was selected to operate a new FFRDC, the *Homeland Security Systems Engineering and Development Institute.*

MITRE President and CEO Al Grasso signs the **HS SEDI** FFRDC sponsoring agreement.

MITRE

Currently Supported Federal Agencies

U.S. Intelligence Community

FDA

Food and Drug Administration

GPO

U.S. GOVERNMENT PRINTING OFFICE

KEEPING AMERICA INFORMED

DISA

DEFENSE INFORMATION SYSTEMS AGENCY

DEPARTMENT OF DEFENSE

US-VISIT

United States Visitor and Immigrant Status Indicator Technology

CMS

CENTERS for MEDICARE & MEDICAID SERVICES

NIST

National Institute of Standards and Technology

USDA

Department of Agriculture

U.S. Immigration and Customs Enforcement

Homeland Security

U.S. Customs and Border Protection

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

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

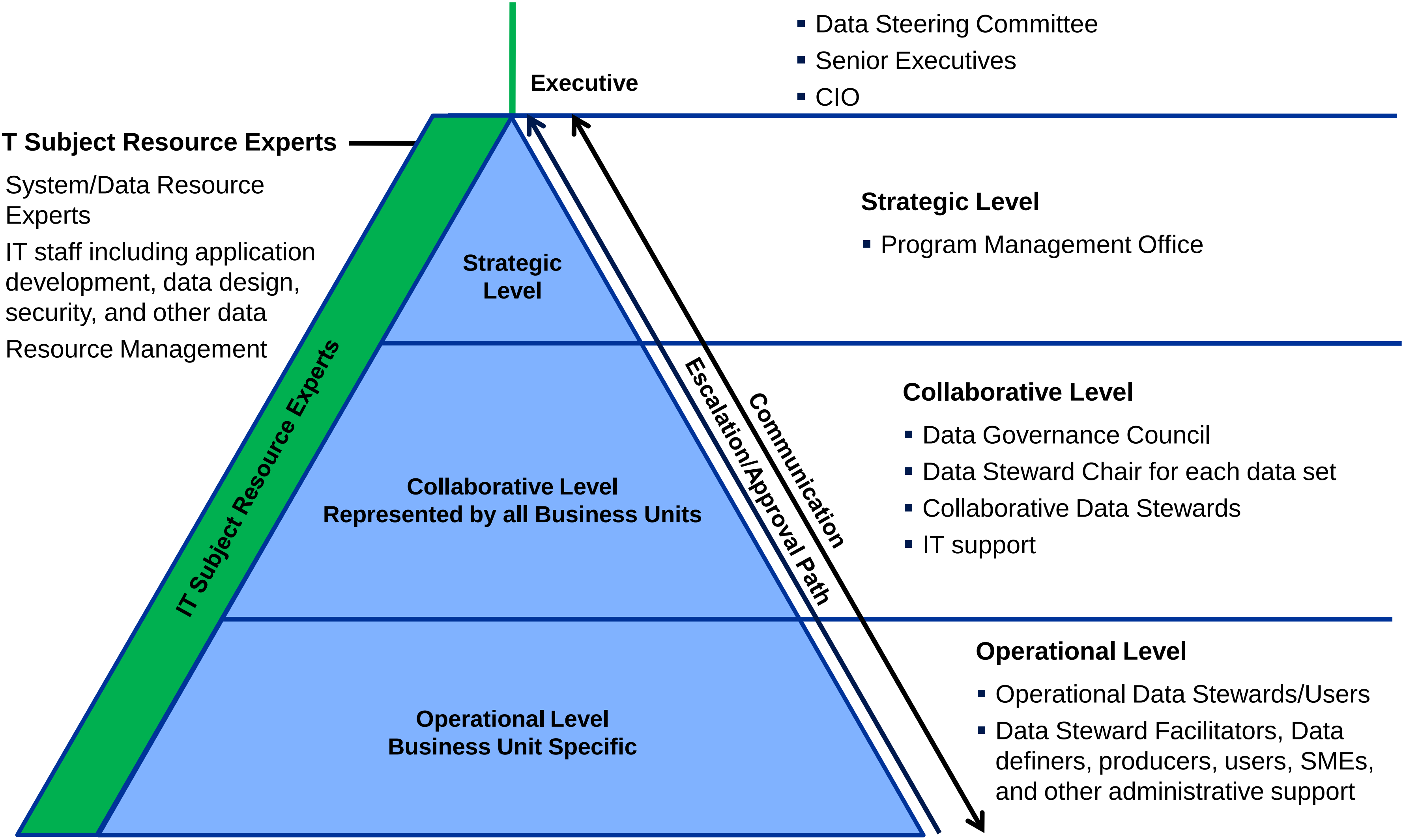
Why, How, Who

- 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 …

Data Governance Framework

IT Subject Resource Experts

- System/Data Resource Experts
- IT staff including application development, data design, security, and other data
- Resource Management



Executive

IT Subject Resource Experts

Strategic Level

Collaborative Level
Represented by all Business Units

Operational Level
Business Unit Specific

Escalation/Approval Path

Communication

Executive Level

- Data Steering Committee
- Senior Executives
- CIO

Strategic Level

- Program Management Office

Collaborative Level

- Data Governance Council
- Data Steward Chair for each data set
- Collaborative Data Stewards
- IT support

Operational Level

- Operational Data Stewards/Users
- Data Steward Facilitators, Data definers, producers, users, SMEs, and other administrative support

A Basic Data Quality Program

- Establish a governance body and structure
- Establish a measureable baseline – with an emphasis on continual improvement
- Address improvement opportunities with an approach coordinated between business organizations and IT SMEs
- Define processes and procedures for identifying and monitoring metrics:
  - Investigate areas of interest or concern
  - Identify appropriate measures that can be used to assess performance
    - Instituting a process of gathering data
    - Analyzing collected measures
    - Evaluating performance over time
    - Determining improvement opportunities

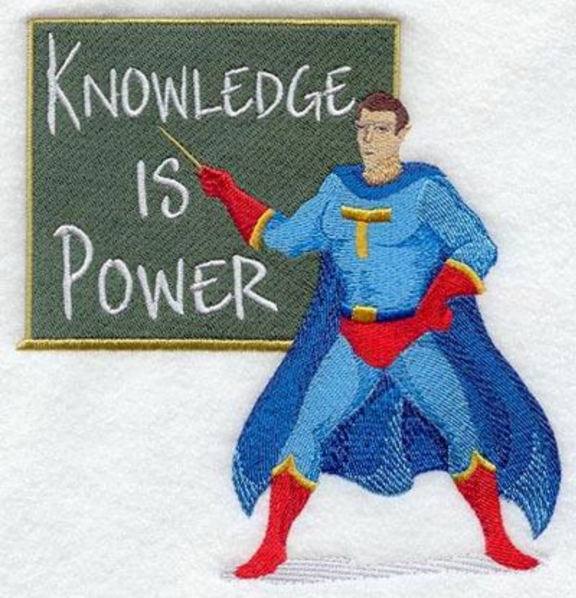
Back in Time

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



If they want to know, they have to ask me

Legacy

Newbies

It'll get fixed in the database

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!

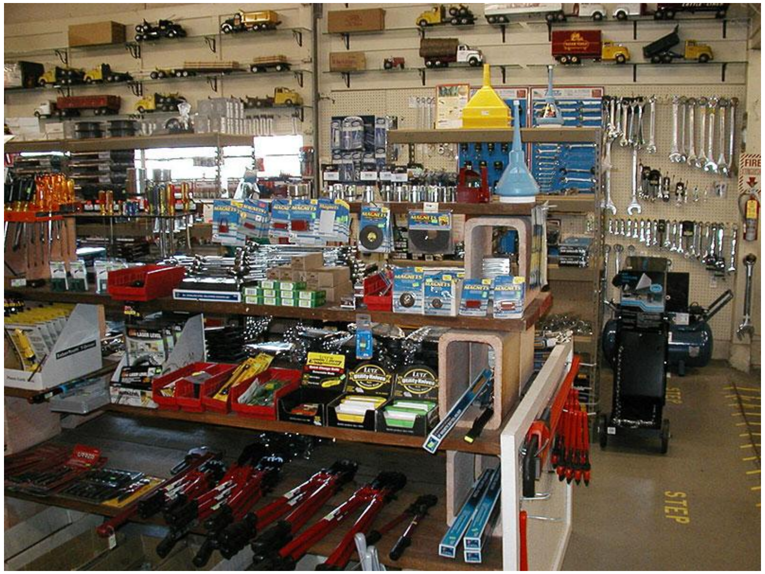
We're state of the art here!

- 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!"

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?

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
  - Profilng
  - Analysis
- QualityStage
  - Parsing
  - Standardization
  - Matching
- Integrated with IBM InfoSphere platform

IBM InfoSphere Information Analyzer

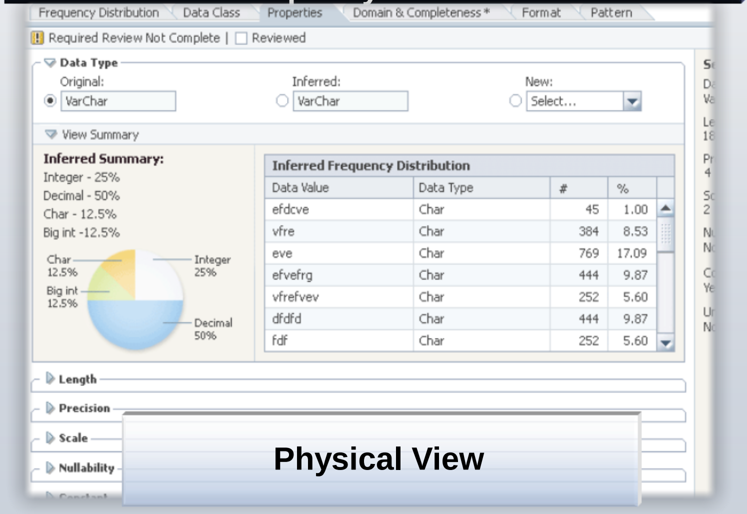
Subject Matter Experts

Data Analysts

Understand

IBM Information Analyzer

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



Required Review Not Complete | □ Reviewed

Data Type

Original: ● VarChar

Inferred: ○ VarChar

New: ○ Select...

Inferred Summary:
Integer - 25%
Decimal - 50%
Char - 12.5%
Big int -12.5%

Inferred Frequency Distribution

Data Value	Data Type	#	%
efdcve	Char	45	1.00
vfre	Char	384	8.53
eve	Char	769	17.09
efvefrg	Char	444	9.87
vfrefvev	Char	252	5.60
dfdfd	Char	444	9.87
fdf	Char	252	5.60

Physical View

- 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

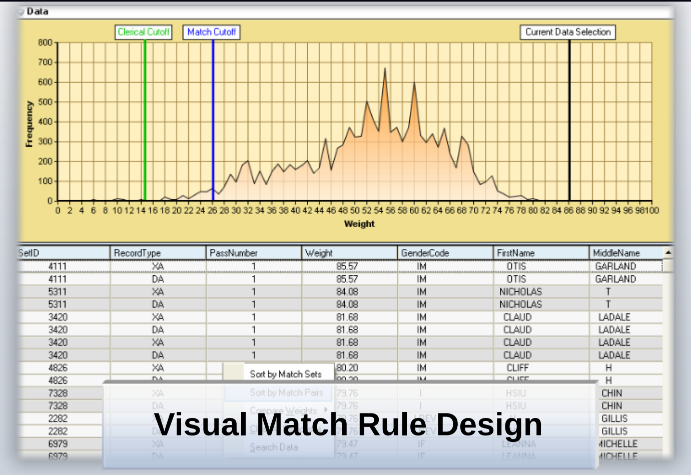
IBM InfoSphere Quality Stage

Subject Matter Experts

Data Analysts

Cleanse

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



SetID	RecordType	PassNumber	Weight	GenderCode	FirstName	MiddleName
4111	XA	1	85.57	IM	OTIS	GARLAND
4111	DA	1	85.57	IM	OTIS	GARLAND
5311	XA	1	84.08	IM	NICHOLAS	T
5311	DA	1	84.08	IM	NICHOLAS	T
3420	XA	1	81.68	IM	CLAUD	LADALE
3420	DA	1	81.68	IM	CLAUD	LADALE
3420	XA	1	81.68	IM	CLAUD	LADALE
3420	DA	1	81.68	IM	CLAUD	LADALE
4826	XA	[illegible]	80.20	IM	CLIFF	H
4826	DA	[illegible]	80.20	IM	CLIFF	H
7328	XA	[illegible]	79.76	I	HSIU	CHIN
7328	DA	[illegible]	79.76	I	HSIU	CHIN
2282	XA	[illegible]	79.76	[illegible]	[illegible]	GILLIS
2282	DA	[illegible]	79.76	[illegible]	[illegible]	GILLIS
6979	XA	[illegible]	79.47	IF	LEANNA	MICHELLE
6979	DA	[illegible]	79.47	IF	LEANNA	MICHELLE

Visual Match Rule Design

Investigation

Understand nature scope and detail of data quality challenges

Standardization

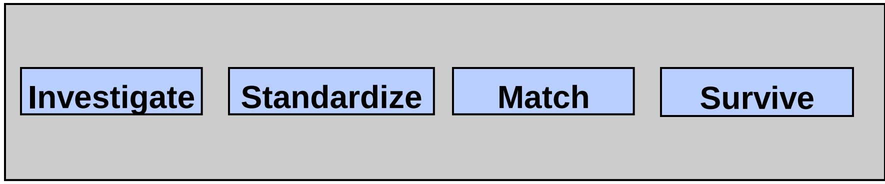
Ensure that data is formatted and conforms to organization wide standards

Matching

Identify duplicate records within and across data sets

Survive

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



Investigate

Standardize

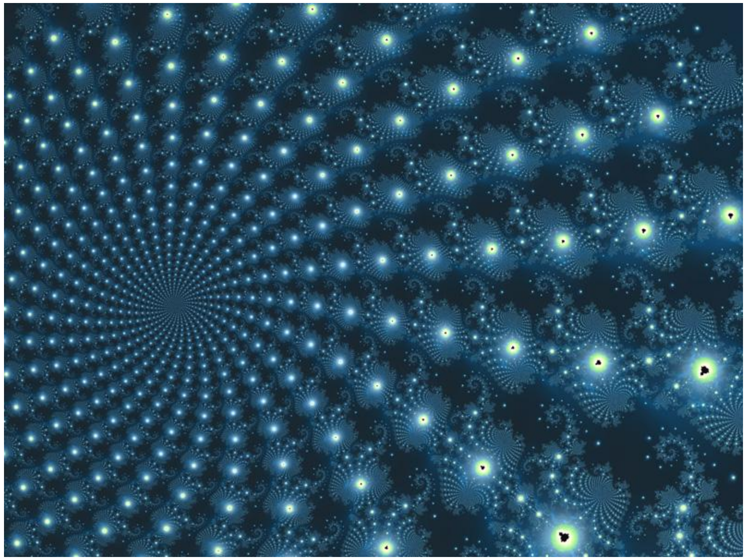
Match

Survive

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