Application of the Federal Data Quality Framework & the Federal Enterprise Architecture Reference Models

ABSTRACT

Problems with data quality in a complex networked world can result in tangible and intangible damage ranging from loss of information consumer confidence to loss of finances, property, or life. Federal agencies and Communities of Interest often have a number of data quality disciplines at their disposal, but rarely will they implement all disciplines at once because improving data quality is a process and not an event.

The Federal Data Quality Framework is designed to provide the guidance for consistent understanding and practices of data quality across government agencies and Communities of Interest by leveraging their existing Enterprise Architectures. A scenario-based activity or two will be provided to understand how to apply the 13 processes of the Federal Data Quality Framework with Enterprise Architecture. Points of emphasis will include:

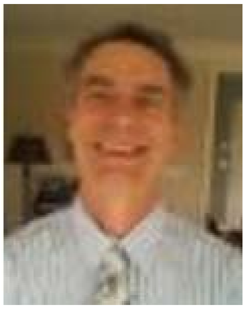
- Using Federal Enterprise Architecture Reference Models to systematically improve data quality
- 13 powerful processes to improve agency data and information
- Course exercises to demonstrate use

BIOGRAPHY

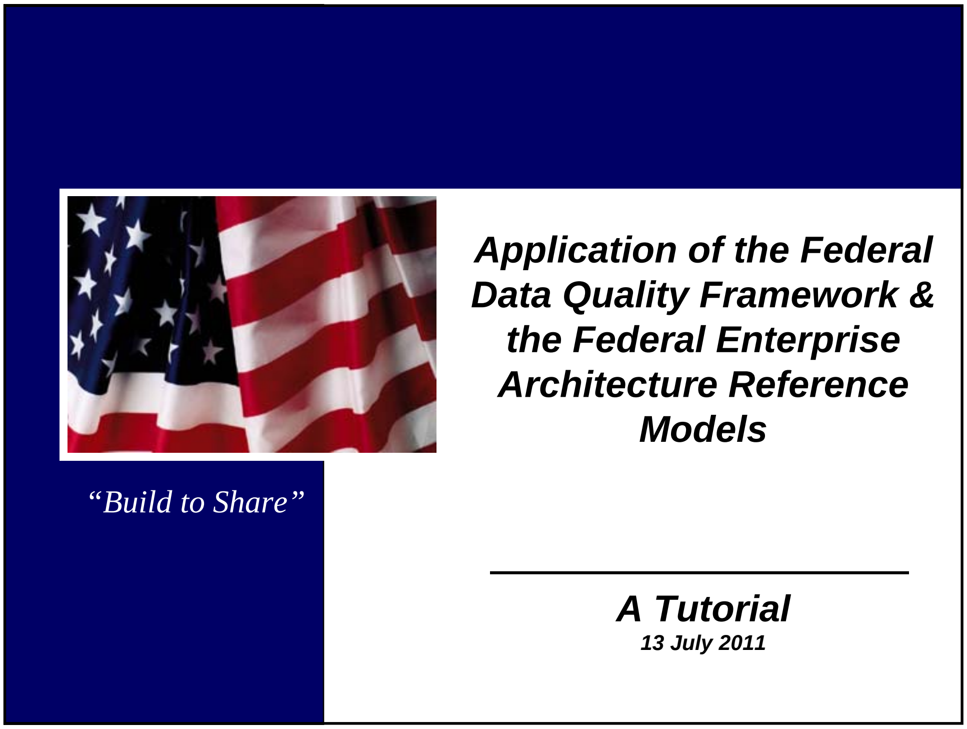
Suzanne Acar
Federal Data Architecture Subcommittee

Suzanne Acar possesses over 25 years of government experience in enterprise data management and data architectures. She currently serves as a leading advisor in enterprise data management at the Federal Bureau of Investigation (FBI) and co-chairs the inter-agency Federal Data Architecture Subcommittee (DAS) of the Federal CIO Council. Dr. Acar teaches courses on enterprise data architectures at two institutes using material that she developed. In the past, she led award winning enterprise data management programs that furthered the interoperability goals of the U.S. Department of the Interior (DOI) and the U.S. Army. She serves as the government advisor for the Data Management Association (DAMA) National Capitol Region Chapter as well as MIT's annual Information Quality Symposium. The DAMA International Government Award and the Federal 100 Award are among the honors and awards granted Dr. Acar.

Mark Amspoker
ATSC



Mark Amspoker, an ISO8000-110 certified master data quality manager at ATSC, has directed a number of data quality improvement practices within the federal government, including the Department of Housing and Urban Development, Department of Transportation, Department of the Interior, Small Business Administration, DOD's Defense Logistics Agency and the Federal Trade Commission. He is the principal author of the U.S. Data Architecture Subcommittee (DAS) "Federal Data Quality Framework" and presents annually at MIT's International Symposium on Information Quality in Boston, Massachusetts.



Application of the Federal Data Quality Framework & the Federal Enterprise Architecture Reference Models

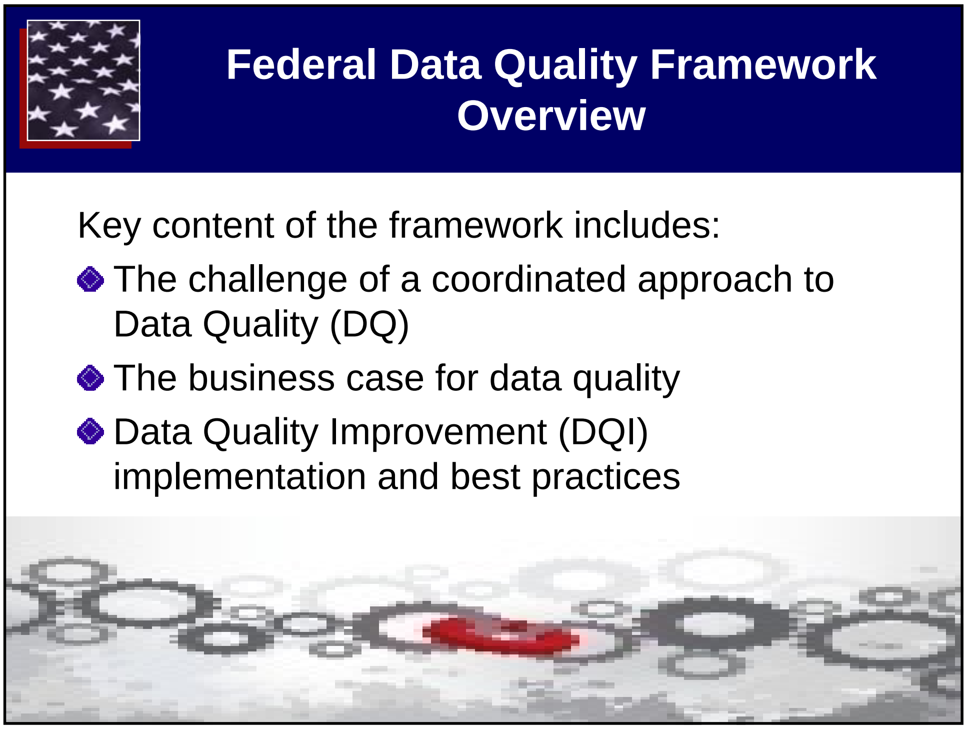
"Build to Share"

A Tutorial

13 July 2011

Agenda

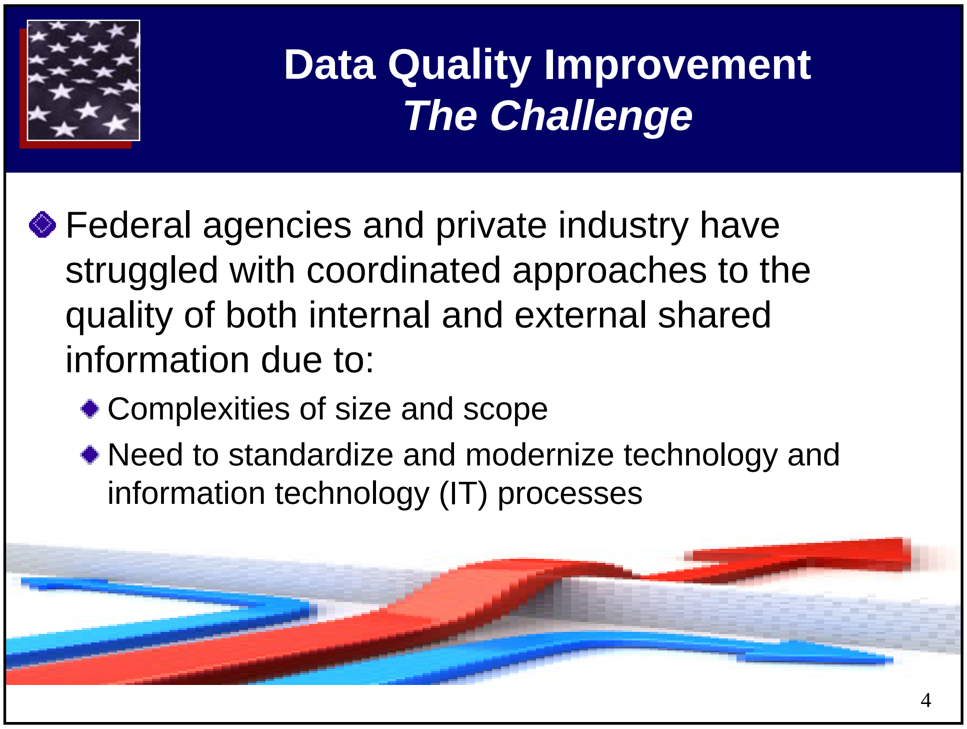
- Federal Data Quality Framework overview
- Tutorial Lessons: Exercises based on actual Data Quality projects & results
- How a Data Quality practice enables Federal Enterprise Architecture objectives
- Summary



Federal Data Quality Framework Overview

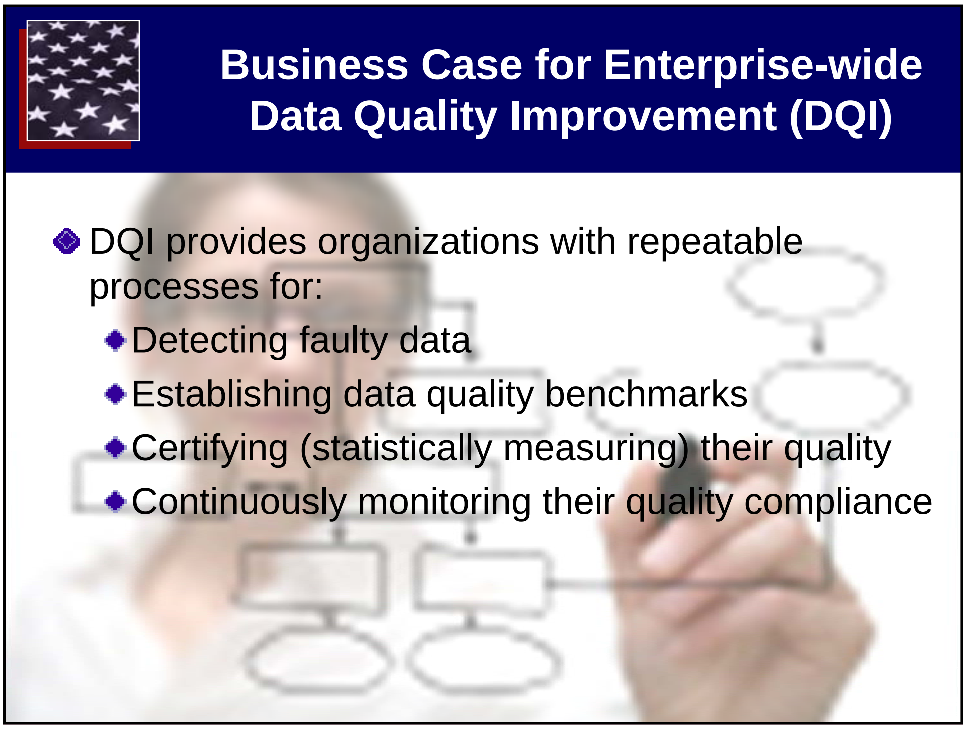
Key content of the framework includes:

- The challenge of a coordinated approach to Data Quality (DQ)
- The business case for data quality
- Data Quality Improvement (DQI) implementation and best practices



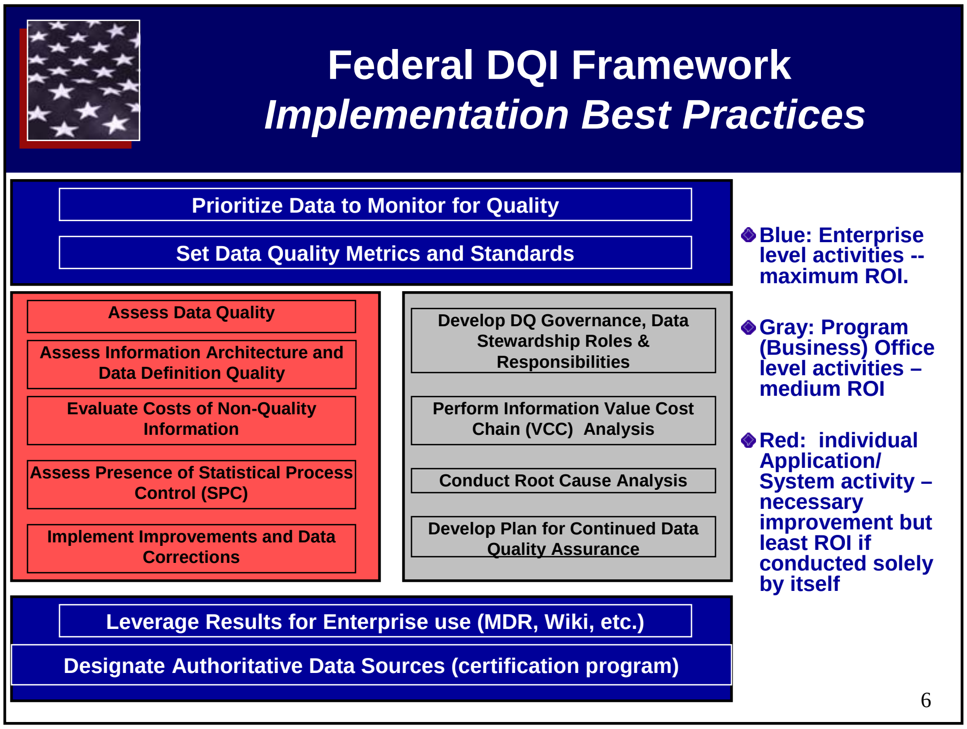
Data Quality Improvement *The Challenge*

- Federal agencies and private industry have struggled with coordinated approaches to the quality of both internal and external shared information due to:
  - Complexities of size and scope
  - Need to standardize and modernize technology and information technology (IT) processes



Business Case for Enterprise-wide Data Quality Improvement (DQI)

- DQI provides organizations with repeatable processes for:
  - Detecting faulty data
  - Establishing data quality benchmarks
  - Certifying (statistically measuring) their quality
  - Continuously monitoring their quality compliance



Federal DQI Framework *Implementation Best Practices*

Prioritize Data to Monitor for Quality

Set Data Quality Metrics and Standards

Assess Data Quality

Assess Information Architecture and Data Definition Quality

Evaluate Costs of Non-Quality Information

Assess Presence of Statistical Process Control (SPC)

Implement Improvements and Data Corrections

Develop DQ Governance, Data Stewardship Roles & Responsibilities

Perform Information Value Cost Chain (VCC) Analysis

Conduct Root Cause Analysis

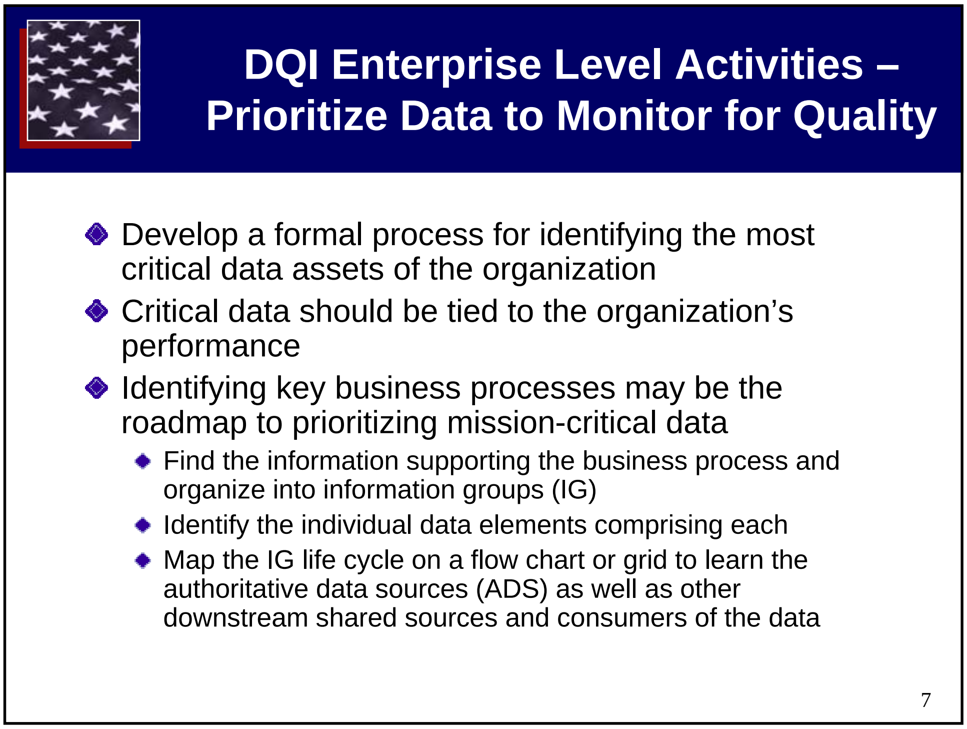
Develop Plan for Continued Data Quality Assurance

Leverage Results for Enterprise use (MDR, Wiki, etc.)

Designate Authoritative Data Sources (certification program)

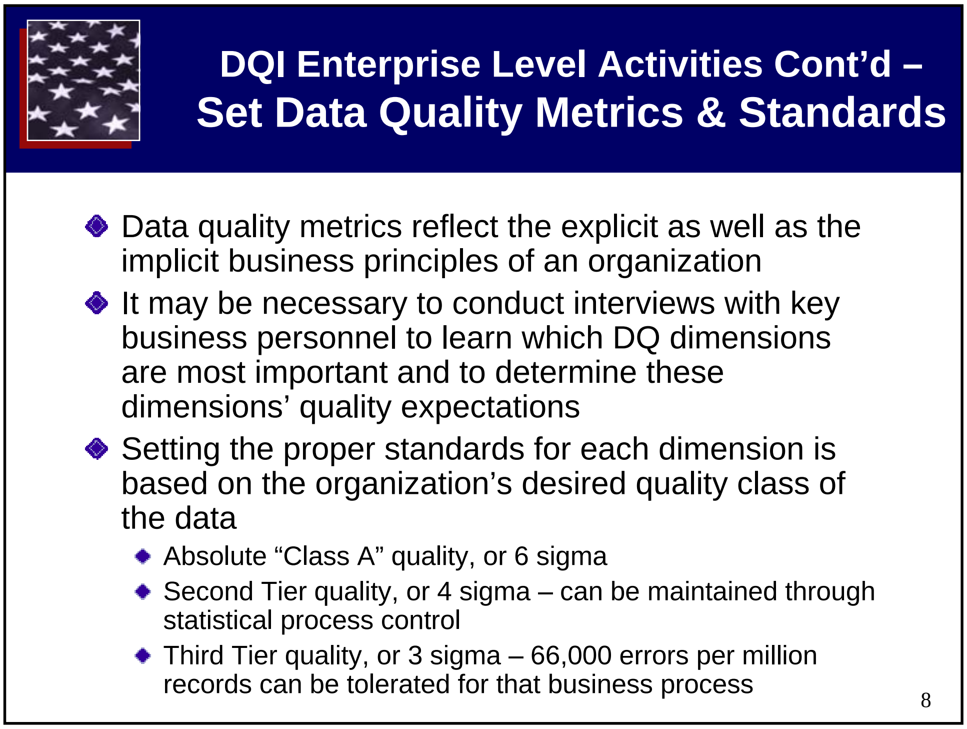
- Blue: Enterprise level activities -- maximum ROI.
- Gray: Program (Business) Office level activities – medium ROI
- Red: individual Application/ System activity – necessary improvement but least ROI if conducted solely by itself

6



DQI Enterprise Level Activities – Prioritize Data to Monitor for Quality

- Develop a formal process for identifying the most critical data assets of the organization
- Critical data should be tied to the organization’s performance
- Identifying key business processes may be the roadmap to prioritizing mission-critical data
  - Find the information supporting the business process and organize into information groups (IG)
  - Identify the individual data elements comprising each
  - Map the IG life cycle on a flow chart or grid to learn the authoritative data sources (ADS) as well as other downstream shared sources and consumers of the data

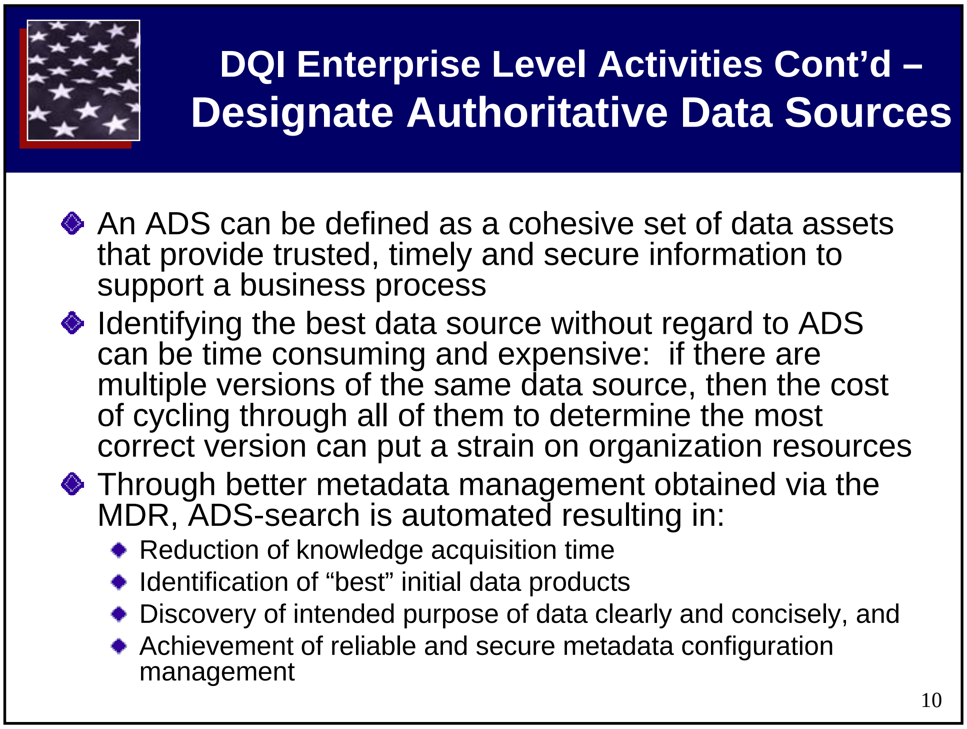


DQI Enterprise Level Activities Cont’d – Set Data Quality Metrics & Standards

- Data quality metrics reflect the explicit as well as the implicit business principles of an organization
- It may be necessary to conduct interviews with key business personnel to learn which DQ dimensions are most important and to determine these dimensions’ quality expectations
- Setting the proper standards for each dimension is based on the organization’s desired quality class of the data
  - Absolute “Class A” quality, or 6 sigma
  - Second Tier quality, or 4 sigma – can be maintained through statistical process control
  - Third Tier quality, or 3 sigma – 66,000 errors per million records can be tolerated for that business process

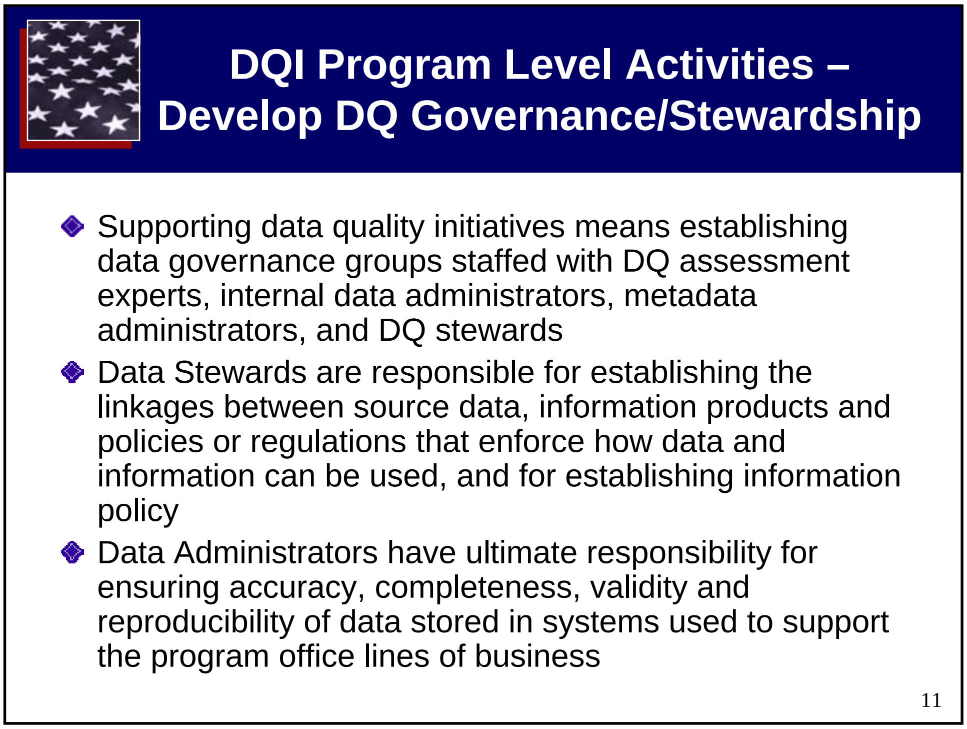
DQI Enterprise Level Activities Cont'd – Leverage Results for Enterprise Use

- Data quality assessment information – best practices, procedures, training materials, standards, DQ artifacts – should be made available to future data quality projects, so that the information and experience from earlier efforts can be leveraged to yield greater success for subsequent efforts
- A Metadata Repository (MDR) holds DQ findings and other metadata, has the ability to be cross-referenced and has some mechanism of version control
- Educating the organization about data quality successes can be accomplished through classroom training, computer-based training, an announcement on the agency's Intranet, an internal newsletter, or simple e-mail notification



DQI Enterprise Level Activities Cont'd – Designate Authoritative Data Sources

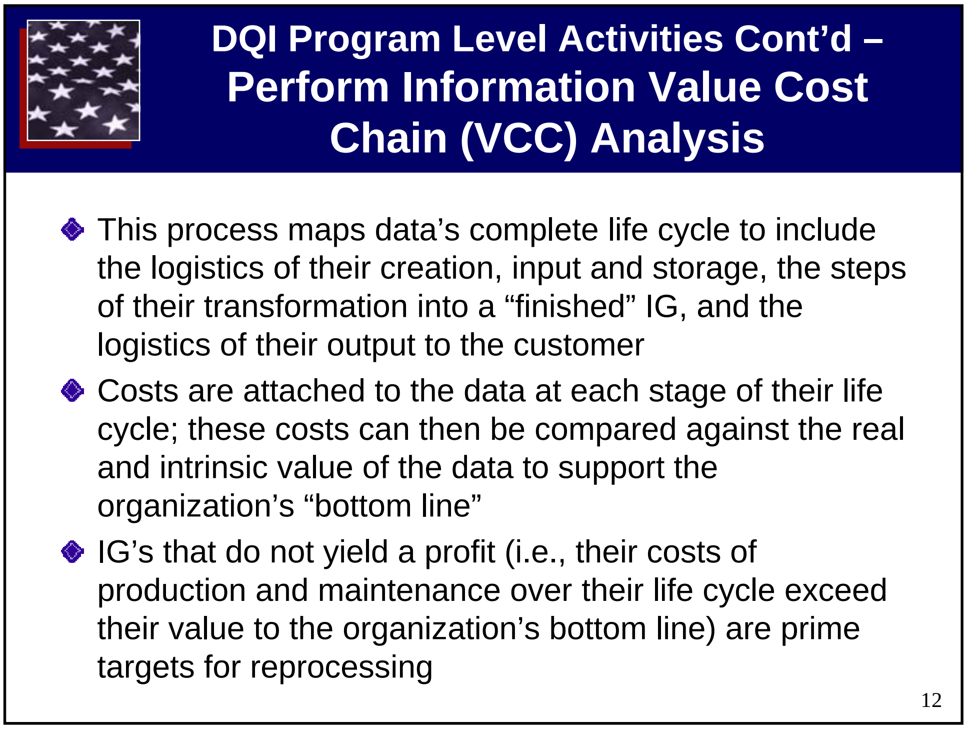
- An ADS can be defined as a cohesive set of data assets that provide trusted, timely and secure information to support a business process
- Identifying the best data source without regard to ADS can be time consuming and expensive: if there are multiple versions of the same data source, then the cost of cycling through all of them to determine the most correct version can put a strain on organization resources
- Through better metadata management obtained via the MDR, ADS-search is automated resulting in:
  - Reduction of knowledge acquisition time
  - Identification of "best" initial data products
  - Discovery of intended purpose of data clearly and concisely, and
  - Achievement of reliable and secure metadata configuration management



DQI Program Level Activities – Develop DQ Governance/Stewardship

- Supporting data quality initiatives means establishing data governance groups staffed with DQ assessment experts, internal data administrators, metadata administrators, and DQ stewards
- Data Stewards are responsible for establishing the linkages between source data, information products and policies or regulations that enforce how data and information can be used, and for establishing information policy
- Data Administrators have ultimate responsibility for ensuring accuracy, completeness, validity and reproducibility of data stored in systems used to support the program office lines of business

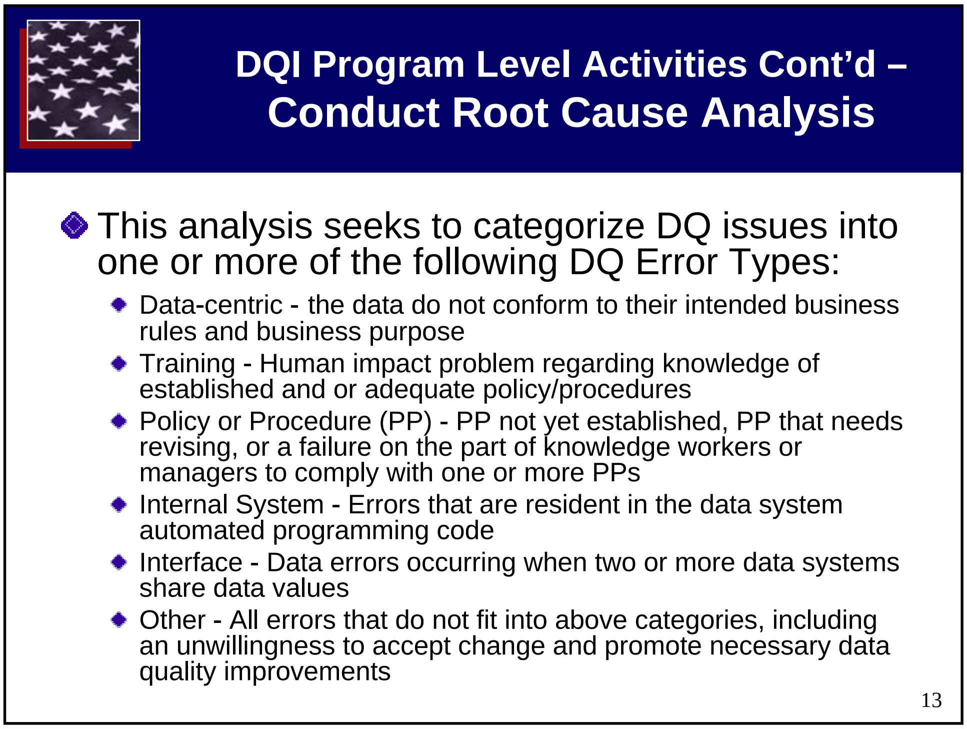
11



DQI Program Level Activities Cont'd – Perform Information Value Cost Chain (VCC) Analysis

- This process maps data's complete life cycle to include the logistics of their creation, input and storage, the steps of their transformation into a "finished" IG, and the logistics of their output to the customer
- Costs are attached to the data at each stage of their life cycle; these costs can then be compared against the real and intrinsic value of the data to support the organization's "bottom line"
- IG's that do not yield a profit (i.e., their costs of production and maintenance over their life cycle exceed their value to the organization's bottom line) are prime targets for reprocessing

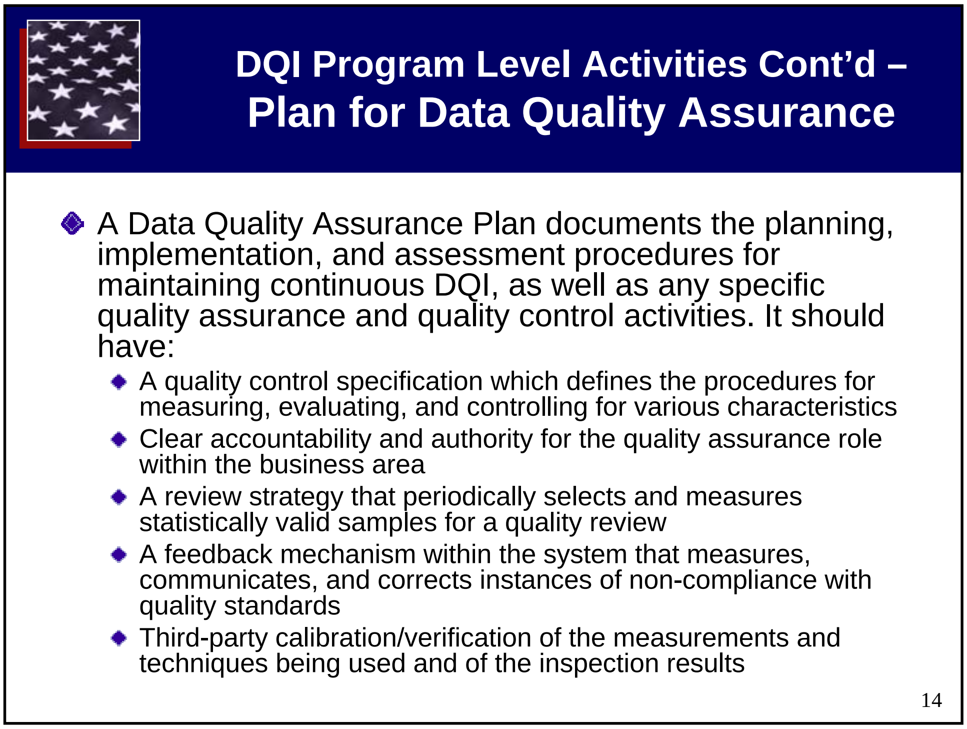
12



DQI Program Level Activities Cont'd – Conduct Root Cause Analysis

- This analysis seeks to categorize DQ issues into one or more of the following DQ Error Types:
  - Data-centric - the data do not conform to their intended business rules and business purpose
  - Training - Human impact problem regarding knowledge of established and or adequate policy/procedures
  - Policy or Procedure (PP) - PP not yet established, PP that needs revising, or a failure on the part of knowledge workers or managers to comply with one or more PPs
  - Internal System - Errors that are resident in the data system automated programming code
  - Interface - Data errors occurring when two or more data systems share data values
  - Other - All errors that do not fit into above categories, including an unwillingness to accept change and promote necessary data quality improvements

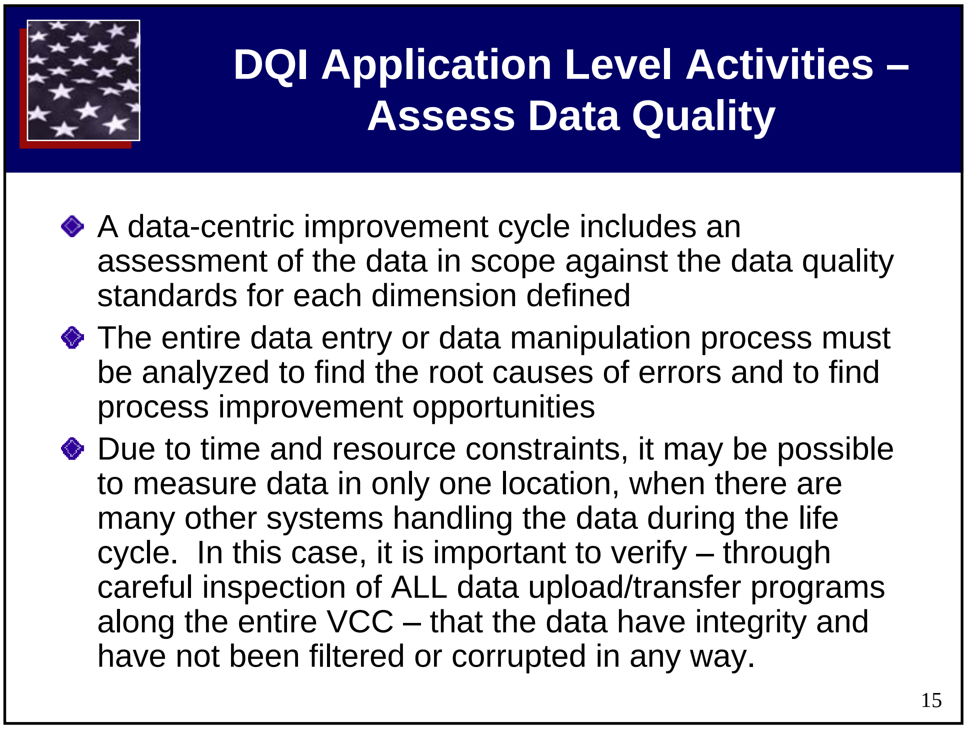
13



DQI Program Level Activities Cont'd – Plan for Data Quality Assurance

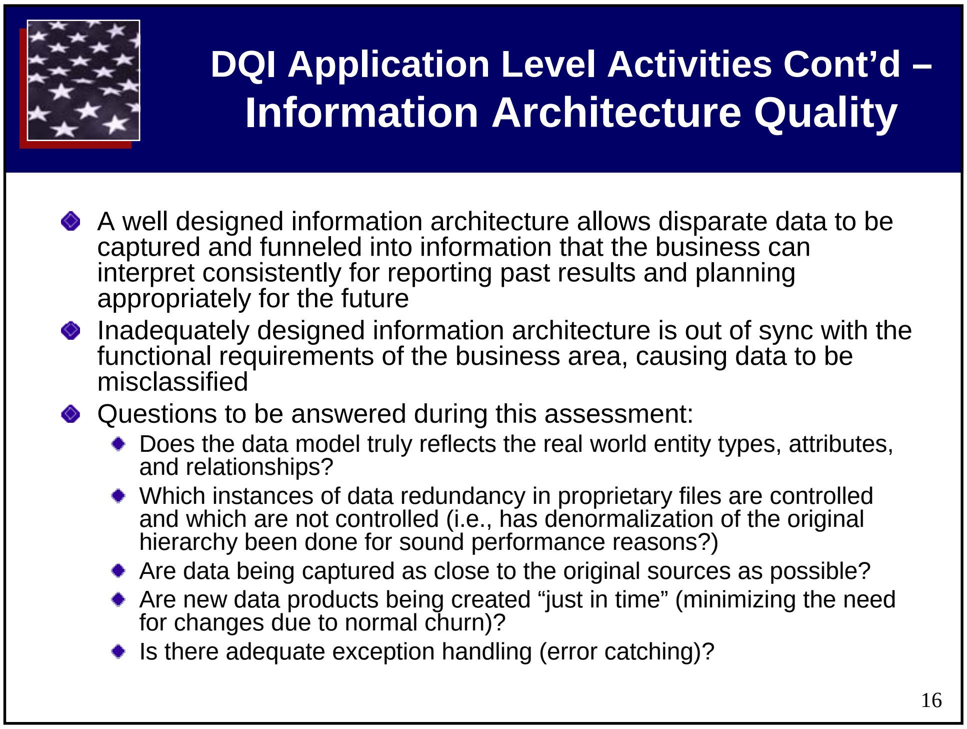
- A Data Quality Assurance Plan documents the planning, implementation, and assessment procedures for maintaining continuous DQI, as well as any specific quality assurance and quality control activities. It should have:
  - A quality control specification which defines the procedures for measuring, evaluating, and controlling for various characteristics
  - Clear accountability and authority for the quality assurance role within the business area
  - A review strategy that periodically selects and measures statistically valid samples for a quality review
  - A feedback mechanism within the system that measures, communicates, and corrects instances of non-compliance with quality standards
  - Third-party calibration/verification of the measurements and techniques being used and of the inspection results

14



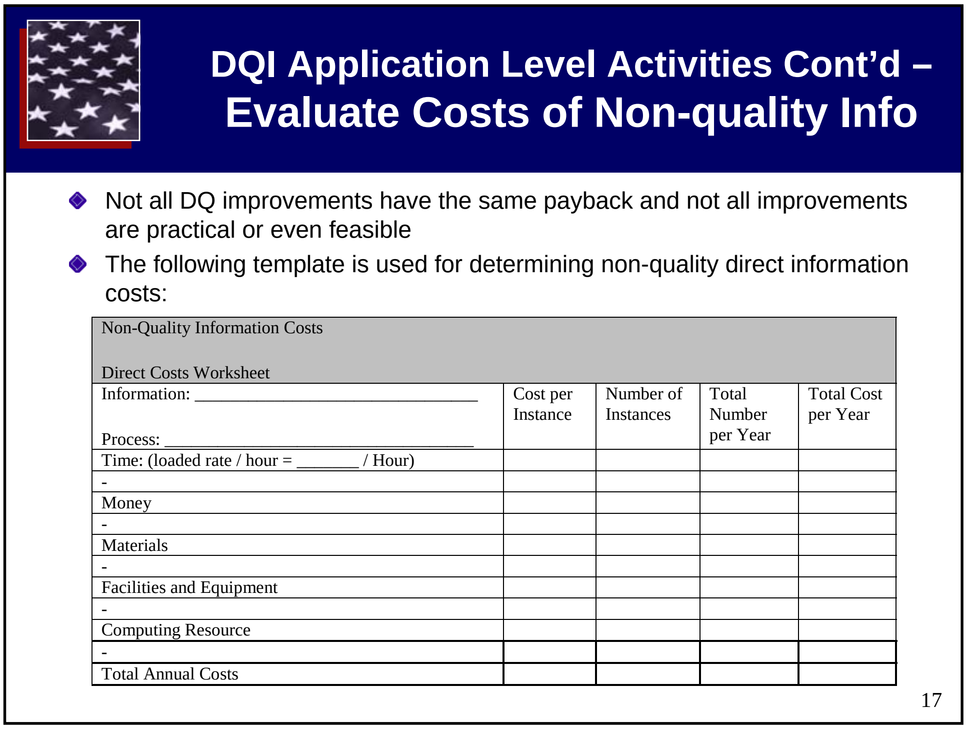
DQI Application Level Activities – Assess Data Quality

- A data-centric improvement cycle includes an assessment of the data in scope against the data quality standards for each dimension defined
- The entire data entry or data manipulation process must be analyzed to find the root causes of errors and to find process improvement opportunities
- Due to time and resource constraints, it may be possible to measure data in only one location, when there are many other systems handling the data during the life cycle. In this case, it is important to verify – through careful inspection of ALL data upload/transfer programs along the entire VCC – that the data have integrity and have not been filtered or corrupted in any way.



DQI Application Level Activities Cont'd – Information Architecture Quality

- A well designed information architecture allows disparate data to be captured and funneled into information that the business can interpret consistently for reporting past results and planning appropriately for the future
- Inadequately designed information architecture is out of sync with the functional requirements of the business area, causing data to be misclassified
- Questions to be answered during this assessment:
  - Does the data model truly reflects the real world entity types, attributes, and relationships?
  - Which instances of data redundancy in proprietary files are controlled and which are not controlled (i.e., has denormalization of the original hierarchy been done for sound performance reasons?)
  - Are data being captured as close to the original sources as possible?
  - Are new data products being created “just in time” (minimizing the need for changes due to normal churn)?
  - Is there adequate exception handling (error catching)?

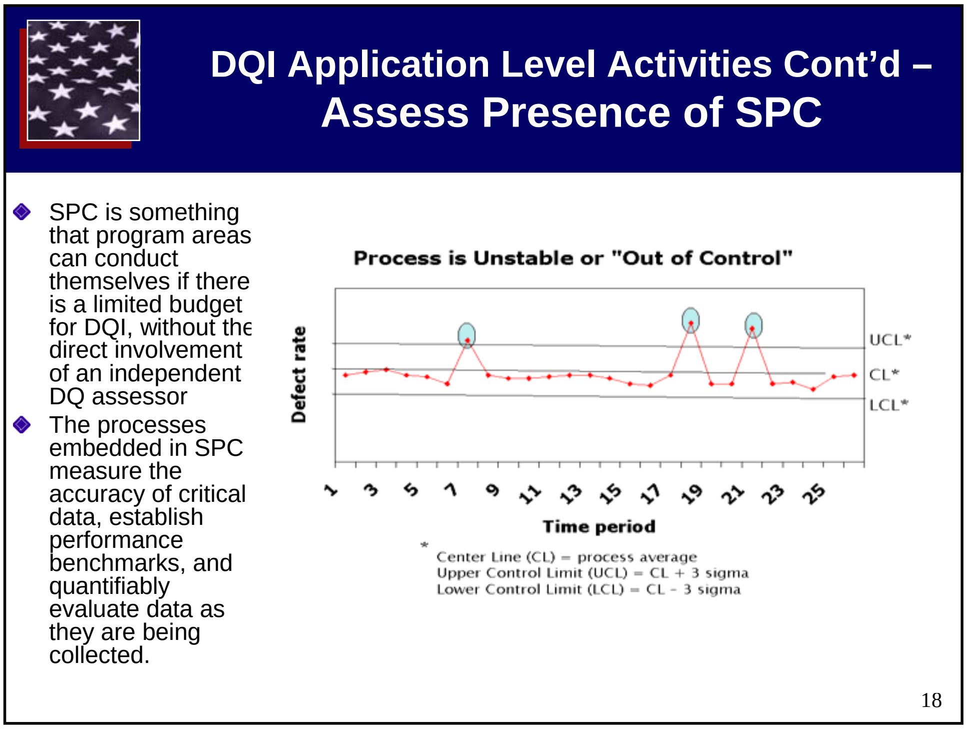


DQI Application Level Activities Cont'd – Evaluate Costs of Non-quality Info

- Not all DQ improvements have the same payback and not all improvements are practical or even feasible
- The following template is used for determining non-quality direct information costs:

Non-Quality Information Costs Direct Costs Worksheet				
Information: _______________________________ Process: ___________________________________	Cost per Instance	Number of Instances	Total Number per Year	Total Cost per Year
Time: (loaded rate / hour = _______ / Hour)				
-				
Money				
-				
Materials				
-				
Facilities and Equipment				
-				
Computing Resource				
-				
Total Annual Costs				

17



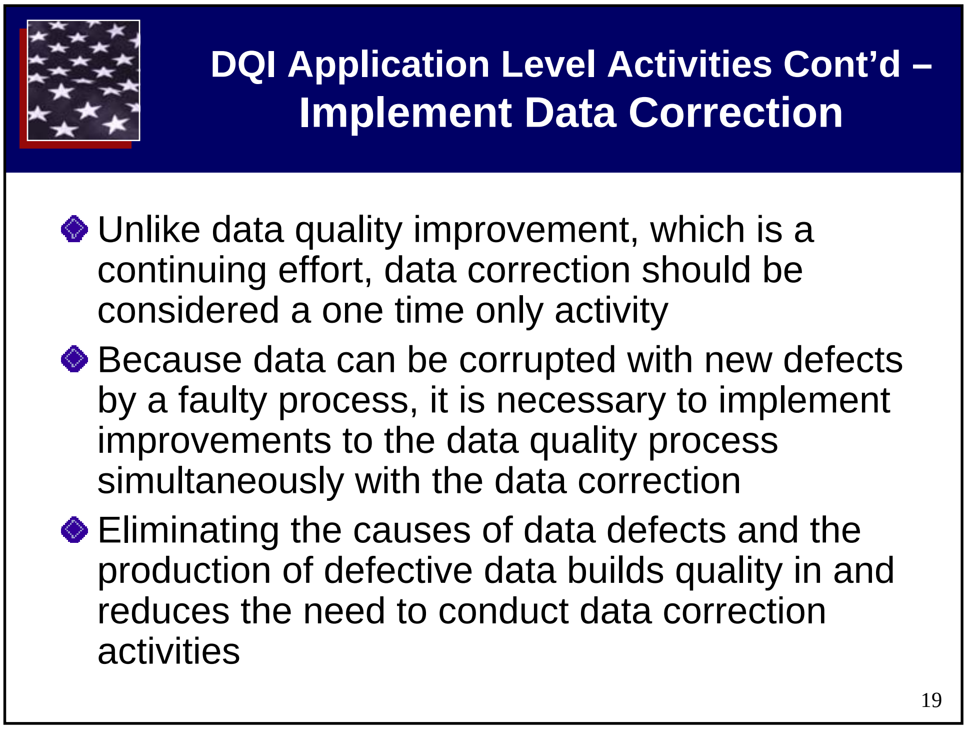
DQI Application Level Activities Cont'd – Assess Presence of SPC

- SPC is something that program areas can conduct themselves if there is a limited budget for DQI, without the direct involvement of an independent DQ assessor
- The processes embedded in SPC measure the accuracy of critical data, establish performance benchmarks, and quantifiably evaluate data as they are being collected.

Process is Unstable or "Out of Control"

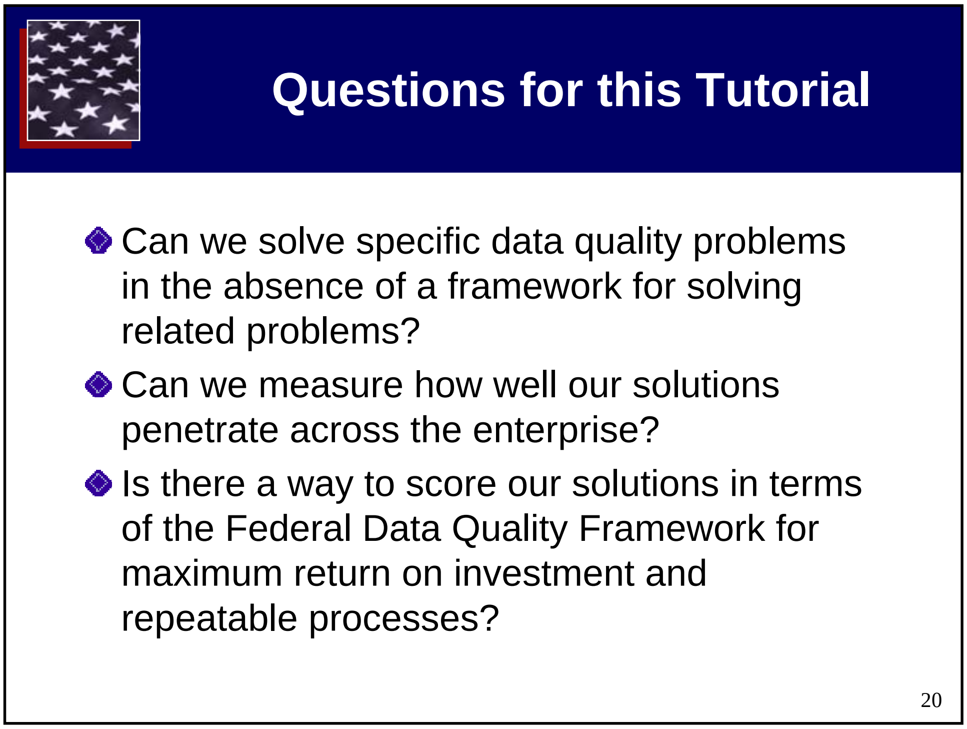
* Center Line (CL) = process average
Upper Control Limit (UCL) = CL + 3 sigma
Lower Control Limit (LCL) = CL - 3 sigma

18



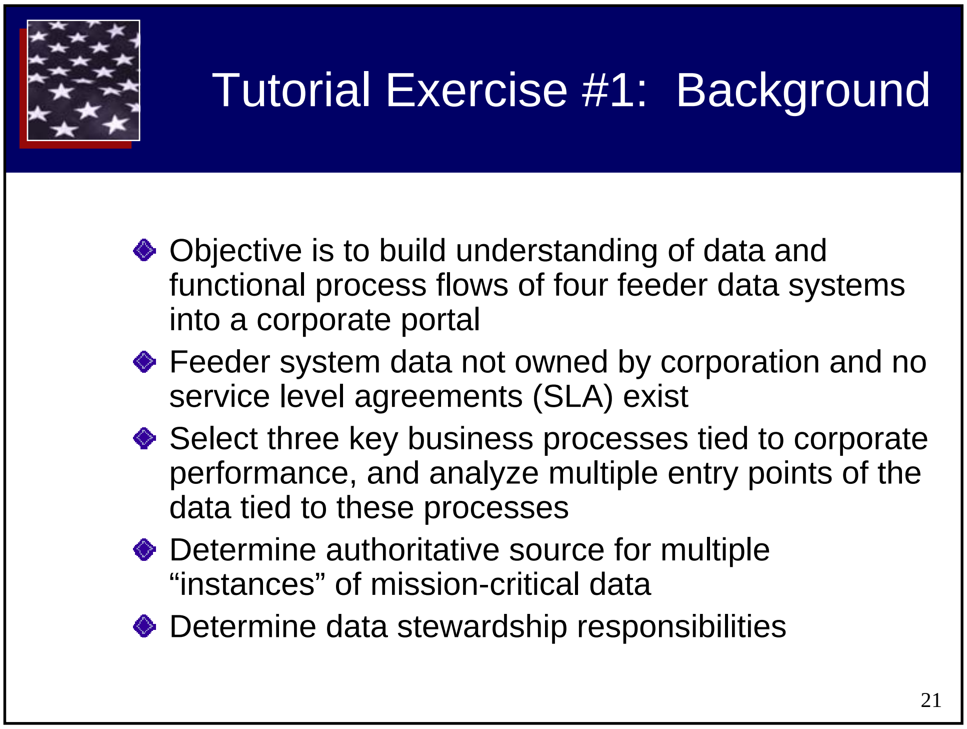
DQI Application Level Activities Cont'd – Implement Data Correction

- Unlike data quality improvement, which is a continuing effort, data correction should be considered a one time only activity
- Because data can be corrupted with new defects by a faulty process, it is necessary to implement improvements to the data quality process simultaneously with the data correction
- Eliminating the causes of data defects and the production of defective data builds quality in and reduces the need to conduct data correction activities



Questions for this Tutorial

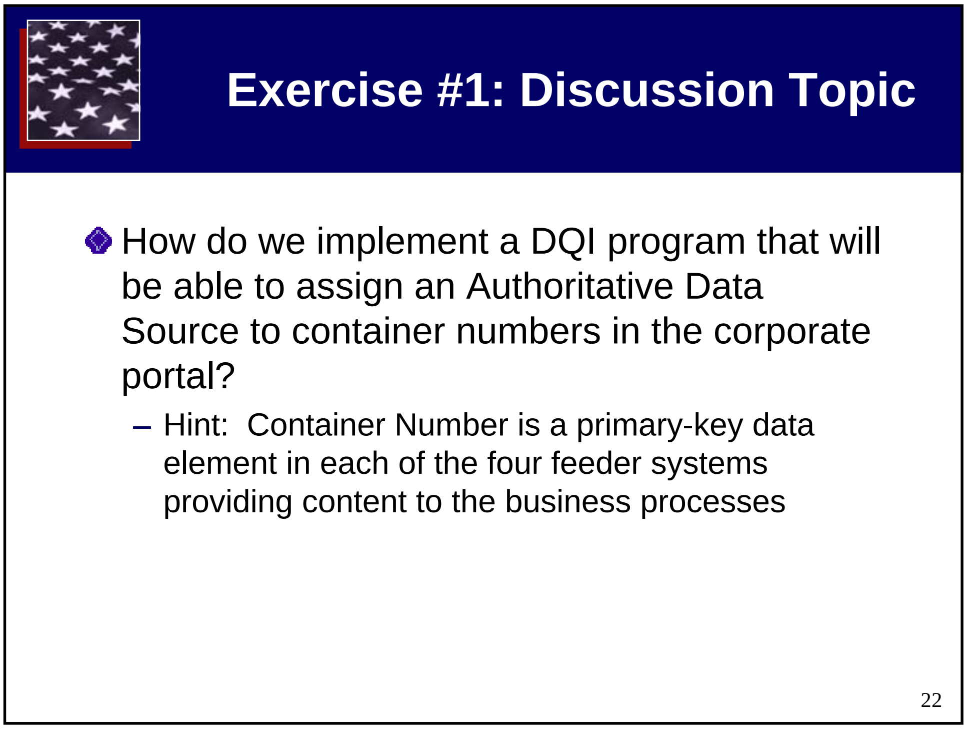
- Can we solve specific data quality problems in the absence of a framework for solving related problems?
- Can we measure how well our solutions penetrate across the enterprise?
- Is there a way to score our solutions in terms of the Federal Data Quality Framework for maximum return on investment and repeatable processes?



Tutorial Exercise #1: Background

- Objective is to build understanding of data and functional process flows of four feeder data systems into a corporate portal
- Feeder system data not owned by corporation and no service level agreements (SLA) exist
- Select three key business processes tied to corporate performance, and analyze multiple entry points of the data tied to these processes
- Determine authoritative source for multiple “instances” of mission-critical data
- Determine data stewardship responsibilities

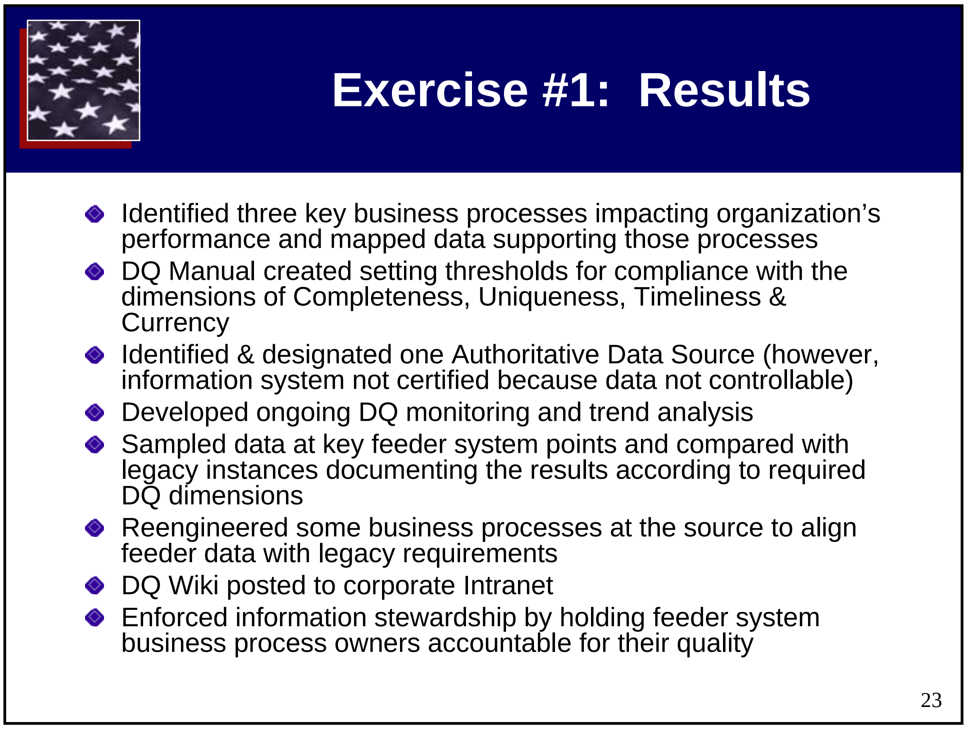
21



Exercise #1: Discussion Topic

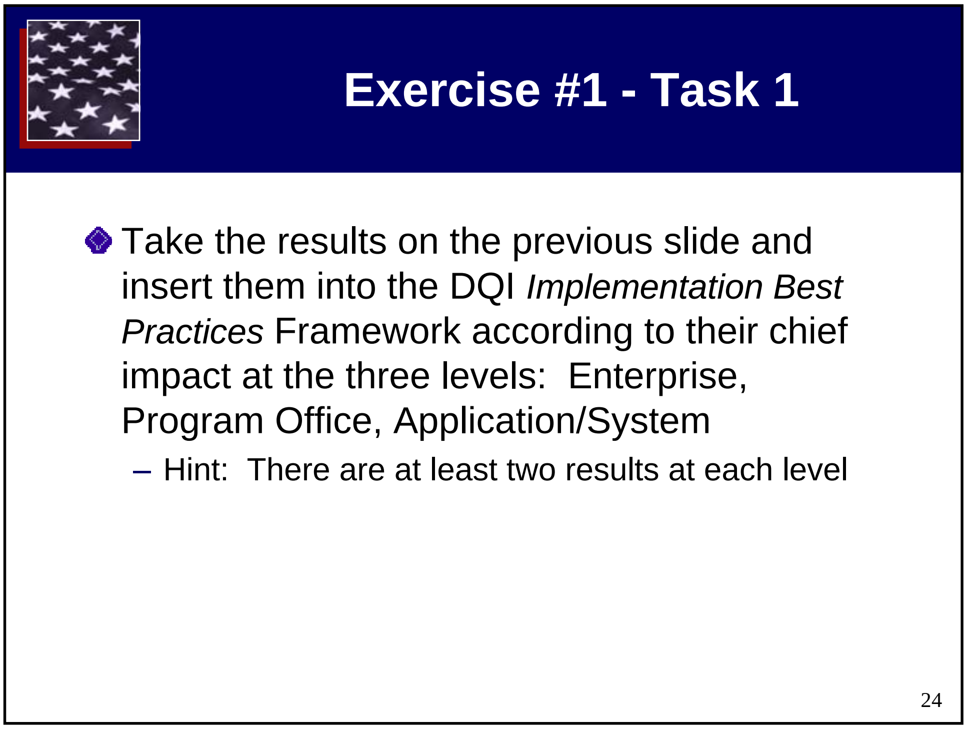
- How do we implement a DQI program that will be able to assign an Authoritative Data Source to container numbers in the corporate portal?
  - Hint: Container Number is a primary-key data element in each of the four feeder systems providing content to the business processes

22



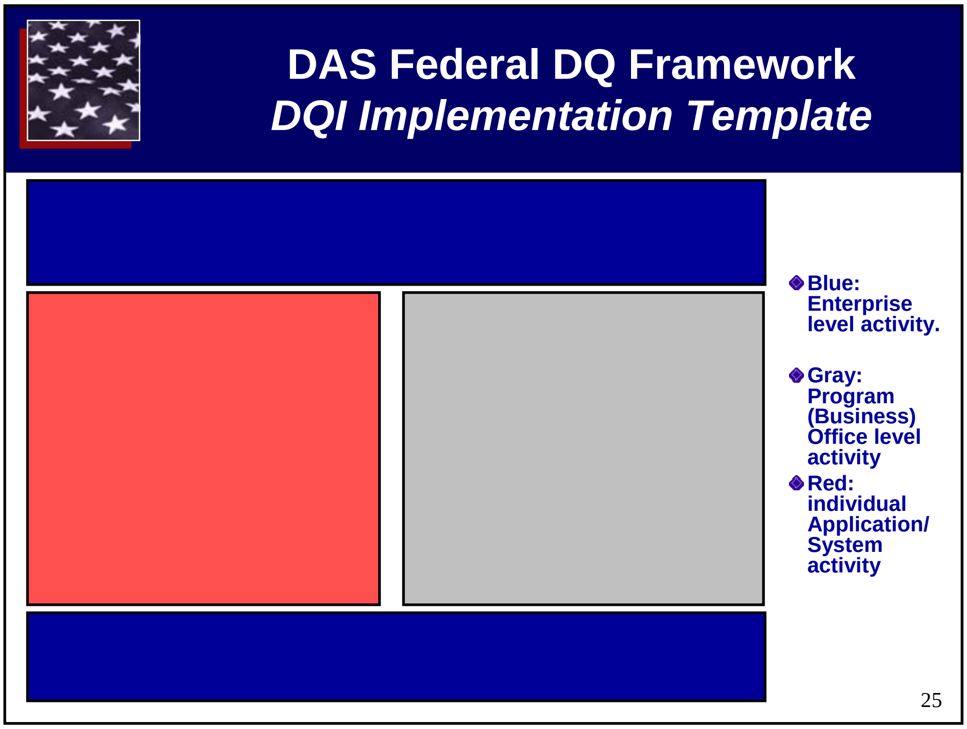
Exercise #1: Results

- Identified three key business processes impacting organization's performance and mapped data supporting those processes
- DQ Manual created setting thresholds for compliance with the dimensions of Completeness, Uniqueness, Timeliness & Currency
- Identified & designated one Authoritative Data Source (however, information system not certified because data not controllable)
- Developed ongoing DQ monitoring and trend analysis
- Sampled data at key feeder system points and compared with legacy instances documenting the results according to required DQ dimensions
- Reengineered some business processes at the source to align feeder data with legacy requirements
- DQ Wiki posted to corporate Intranet
- Enforced information stewardship by holding feeder system business process owners accountable for their quality



Exercise #1 - Task 1

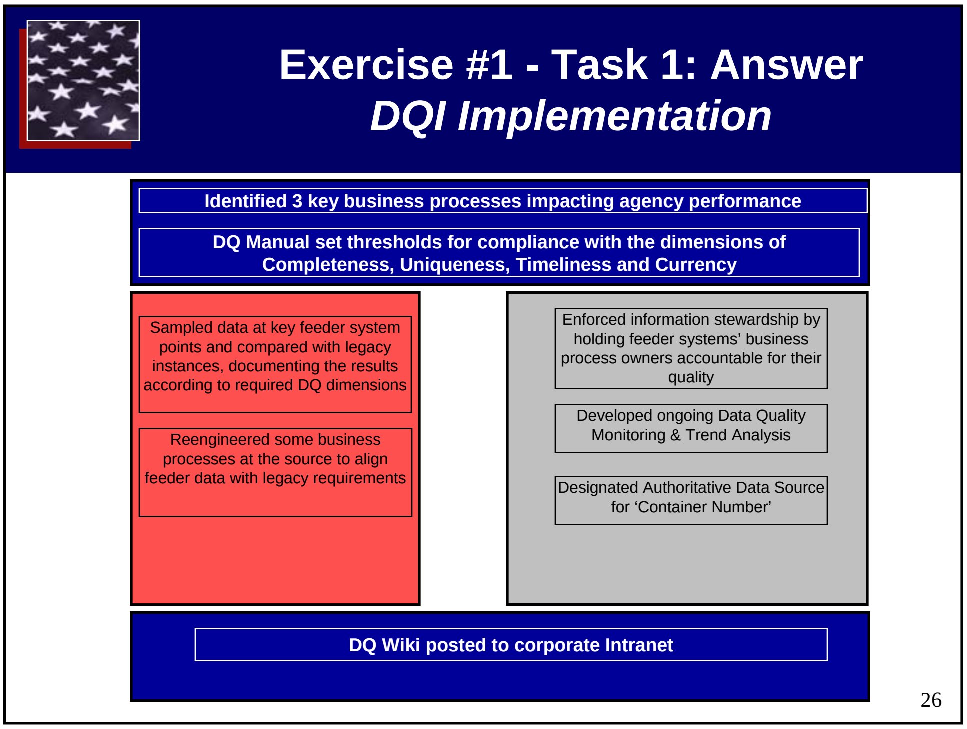
- Take the results on the previous slide and insert them into the DQI *Implementation Best Practices* Framework according to their chief impact at the three levels: Enterprise, Program Office, Application/System
  - Hint: There are at least two results at each level



DAS Federal DQ Framework
DQI Implementation Template

- **Blue: Enterprise level activity.**
- **Gray: Program (Business) Office level activity**
- **Red: individual Application/ System activity**

25



Exercise #1 - Task 1: Answer
DQI Implementation

Identified 3 key business processes impacting agency performance

DQ Manual set thresholds for compliance with the dimensions of Completeness, Uniqueness, Timeliness and Currency

Sampled data at key feeder system points and compared with legacy instances, documenting the results according to required DQ dimensions

Reengineered some business processes at the source to align feeder data with legacy requirements

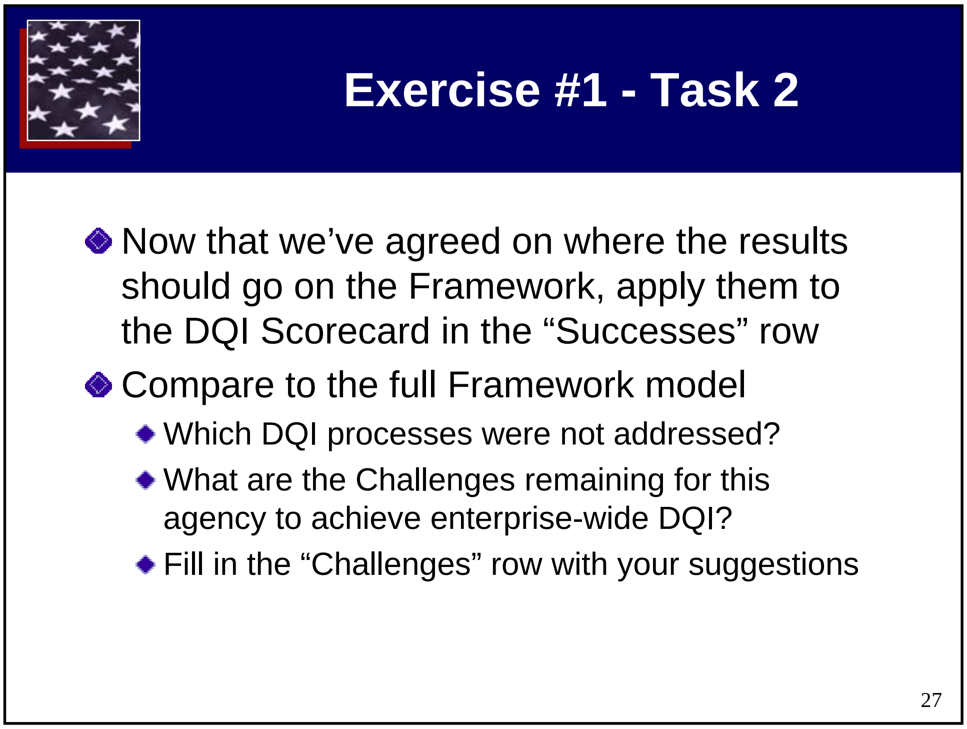
Enforced information stewardship by holding feeder systems' business process owners accountable for their quality

Developed ongoing Data Quality Monitoring & Trend Analysis

Designated Authoritative Data Source for 'Container Number'

DQ Wiki posted to corporate Intranet

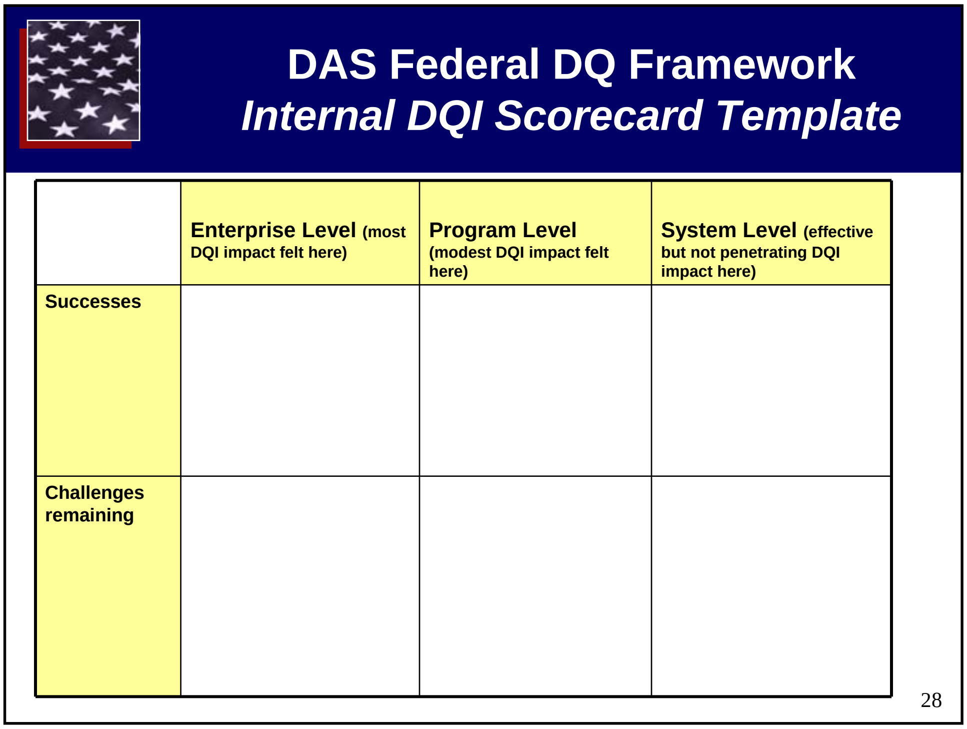
26



Exercise #1 - Task 2

- Now that we've agreed on where the results should go on the Framework, apply them to the DQI Scorecard in the "Successes" row
- Compare to the full Framework model
  - Which DQI processes were not addressed?
  - What are the Challenges remaining for this agency to achieve enterprise-wide DQI?
  - Fill in the "Challenges" row with your suggestions

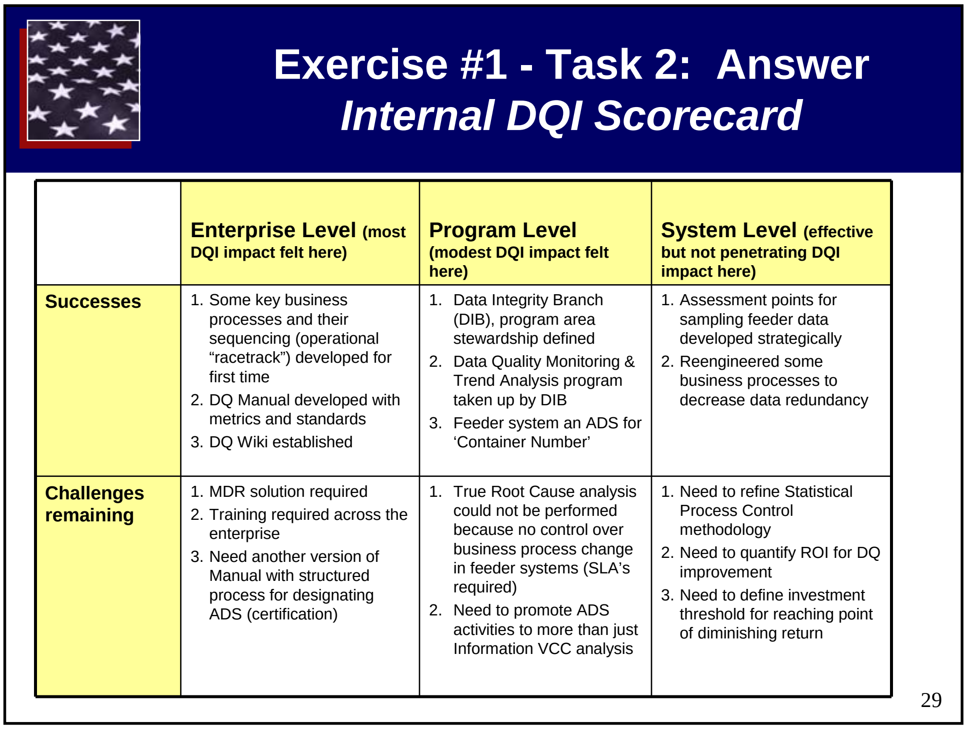
27



DAS Federal DQ Framework
Internal DQI Scorecard Template

	Enterprise Level **(most DQI impact felt here)**	**Program Level** **(modest DQI impact felt here)**	**System Level** **(effective but not penetrating DQI impact here)**
Successes			
Challenges remaining			

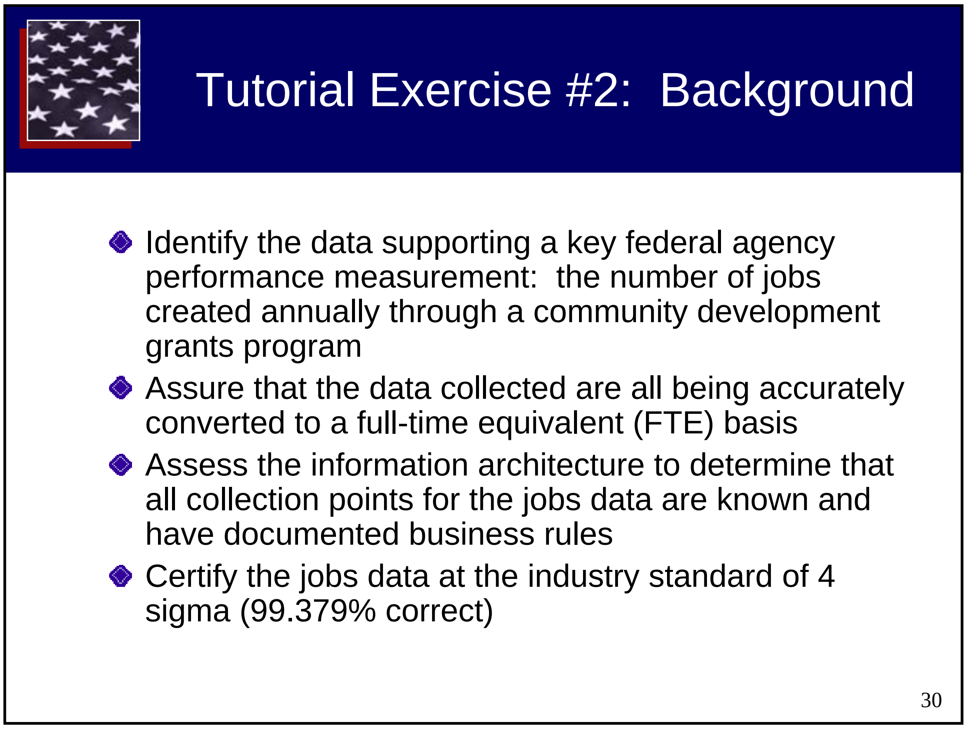
28



Exercise #1 - Task 2: Answer *Internal DQI Scorecard*

	Enterprise Level (most DQI impact felt here)	**Program Level (modest DQI impact felt here)**	**System Level (effective but not penetrating DQI impact here)**
Successes	1. Some key business processes and their sequencing (operational "racetrack") developed for first time 2. DQ Manual developed with metrics and standards 3. DQ Wiki established	1. Data Integrity Branch (DIB), program area stewardship defined 2. Data Quality Monitoring & Trend Analysis program taken up by DIB 3. Feeder system an ADS for 'Container Number'	1. Assessment points for sampling feeder data developed strategically 2. Reengineered some business processes to decrease data redundancy
Challenges remaining	1. MDR solution required 2. Training required across the enterprise 3. Need another version of Manual with structured process for designating ADS (certification)	1. True Root Cause analysis could not be performed because no control over business process change in feeder systems (SLA's required) 2. Need to promote ADS activities to more than just Information VCC analysis	1. Need to refine Statistical Process Control methodology 2. Need to quantify ROI for DQ improvement 3. Need to define investment threshold for reaching point of diminishing return

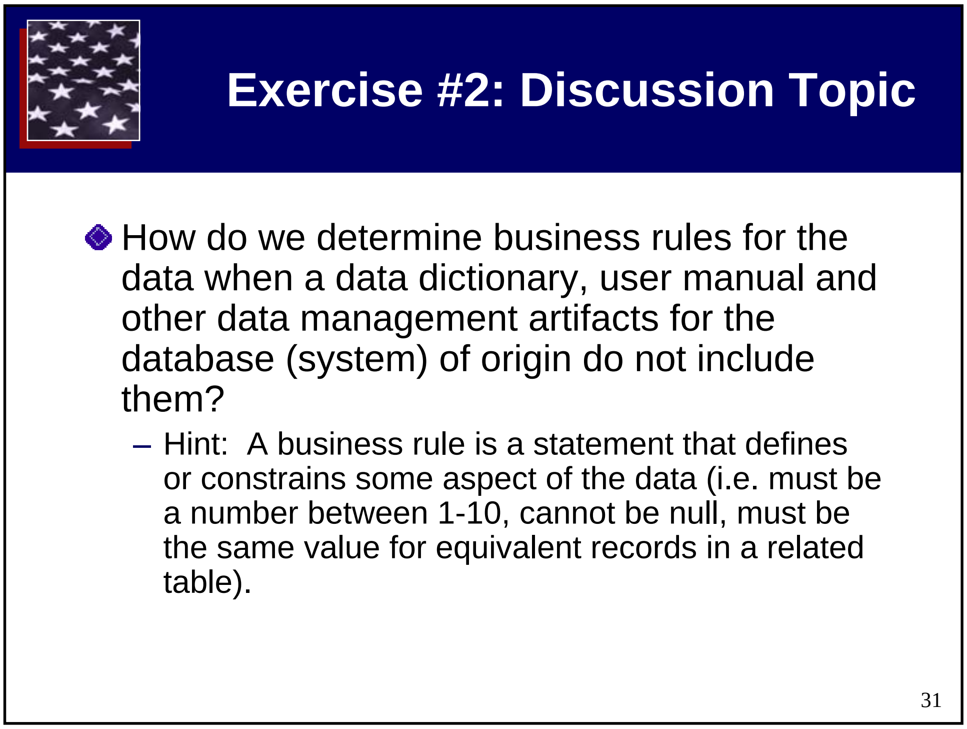
29



Tutorial Exercise #2: Background

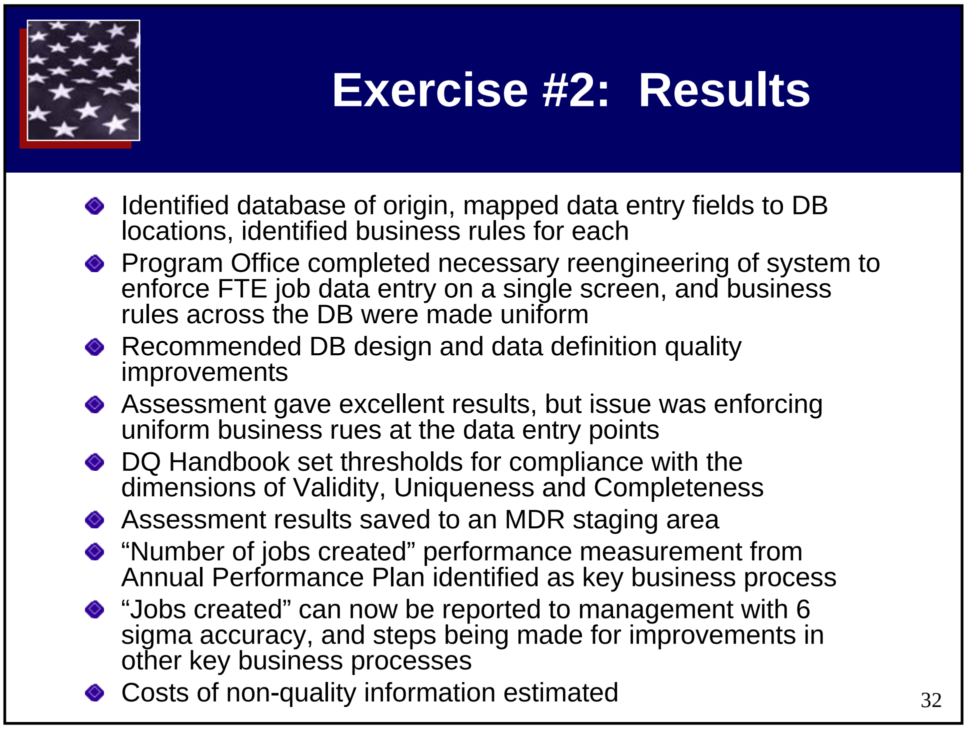
- Identify the data supporting a key federal agency performance measurement: the number of jobs created annually through a community development grants program
- Assure that the data collected are all being accurately converted to a full-time equivalent (FTE) basis
- Assess the information architecture to determine that all collection points for the jobs data are known and have documented business rules
- Certify the jobs data at the industry standard of 4 sigma (99.379% correct)

30



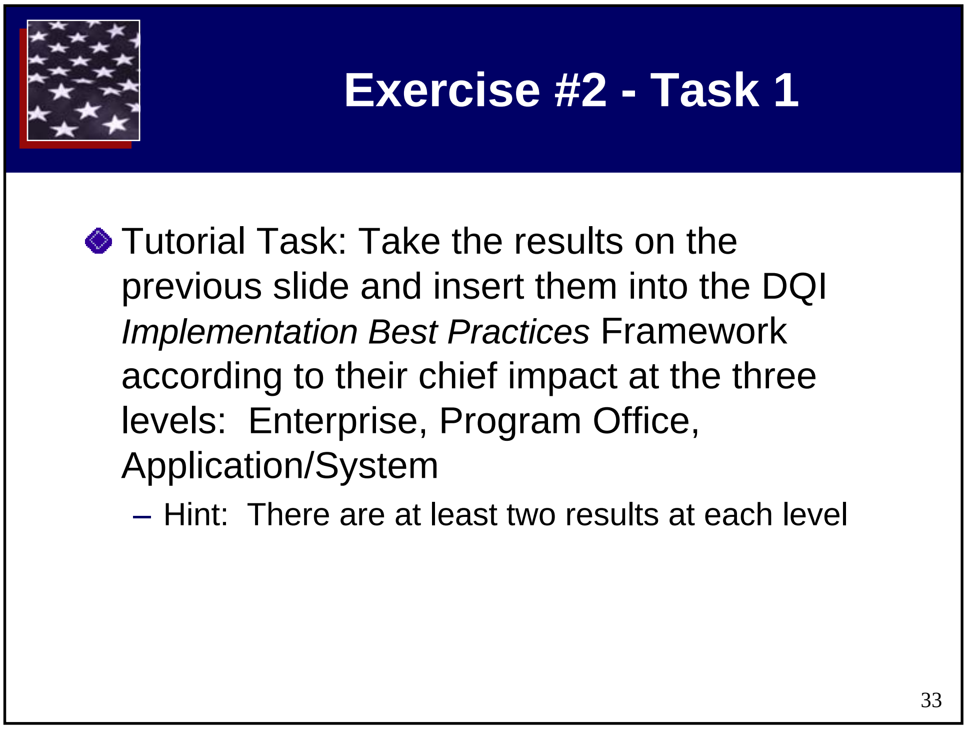
Exercise #2: Discussion Topic

- How do we determine business rules for the data when a data dictionary, user manual and other data management artifacts for the database (system) of origin do not include them?
  - Hint: A business rule is a statement that defines or constrains some aspect of the data (i.e. must be a number between 1-10, cannot be null, must be the same value for equivalent records in a related table).



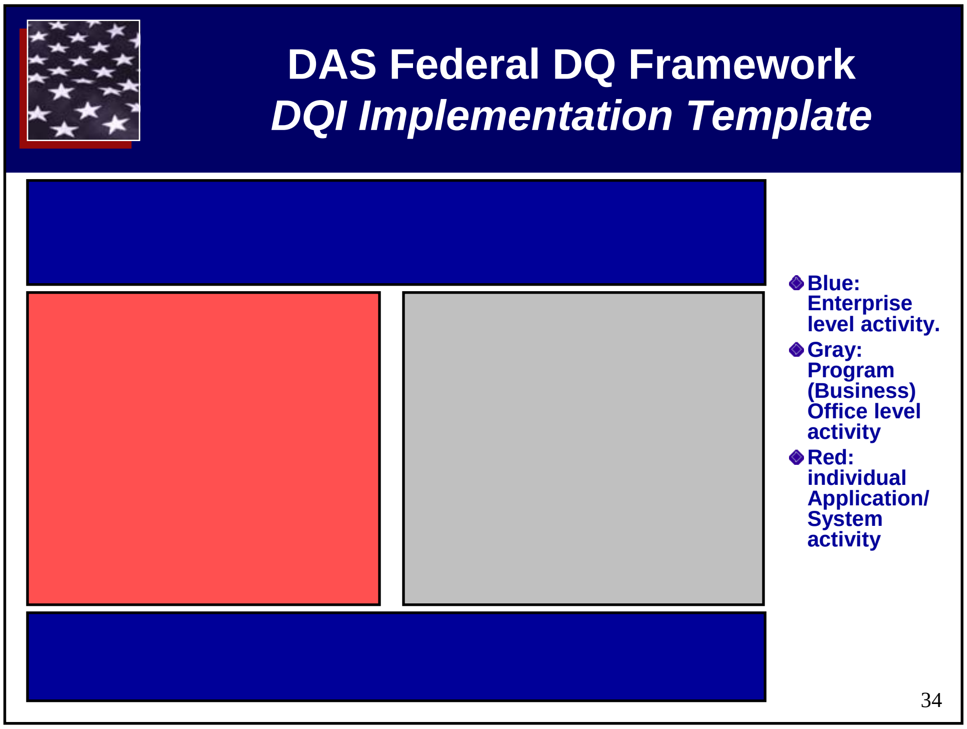
Exercise #2: Results

- Identified database of origin, mapped data entry fields to DB locations, identified business rules for each
- Program Office completed necessary reengineering of system to enforce FTE job data entry on a single screen, and business rules across the DB were made uniform
- Recommended DB design and data definition quality improvements
- Assessment gave excellent results, but issue was enforcing uniform business rues at the data entry points
- DQ Handbook set thresholds for compliance with the dimensions of Validity, Uniqueness and Completeness
- Assessment results saved to an MDR staging area
- “Number of jobs created” performance measurement from Annual Performance Plan identified as key business process
- “Jobs created” can now be reported to management with 6 sigma accuracy, and steps being made for improvements in other key business processes
- Costs of non-quality information estimated



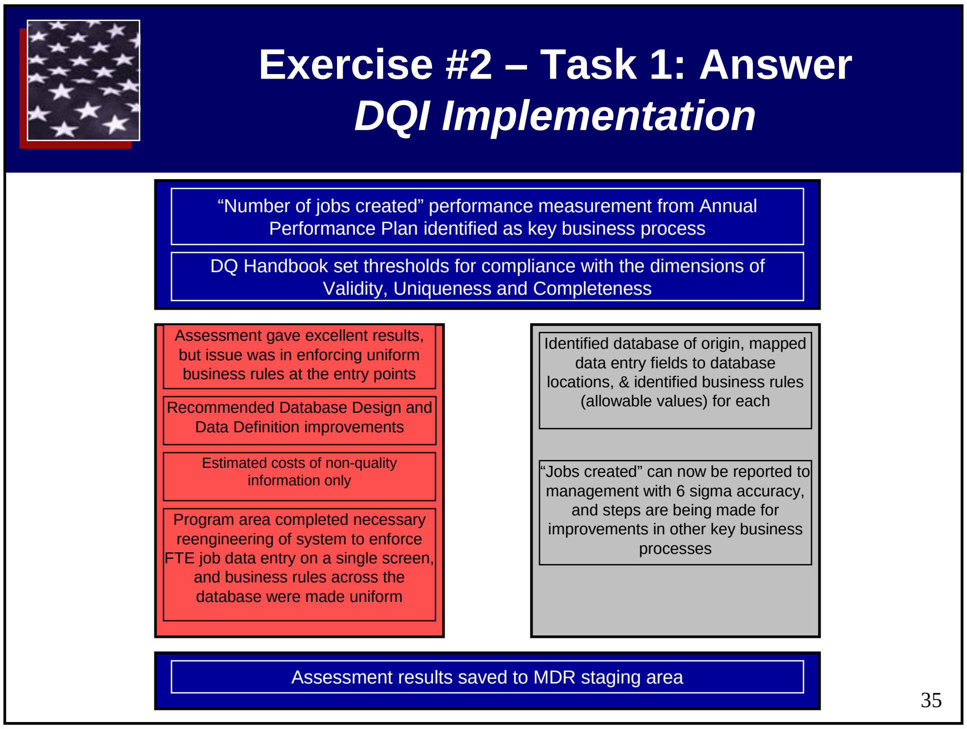
Exercise #2 - Task 1

- Tutorial Task: Take the results on the previous slide and insert them into the DQI *Implementation Best Practices* Framework according to their chief impact at the three levels: Enterprise, Program Office, Application/System
  - Hint: There are at least two results at each level



DAS Federal DQ Framework
DQI Implementation Template

- **Blue: Enterprise level activity.**
- **Gray: Program (Business) Office level activity**
- **Red: individual Application/ System activity**



Exercise #2 – Task 1: Answer *DQI Implementation*

"Number of jobs created" performance measurement from Annual Performance Plan identified as key business process

DQ Handbook set thresholds for compliance with the dimensions of Validity, Uniqueness and Completeness

Assessment gave excellent results, but issue was in enforcing uniform business rules at the entry points

Recommended Database Design and Data Definition improvements

Estimated costs of non-quality information only

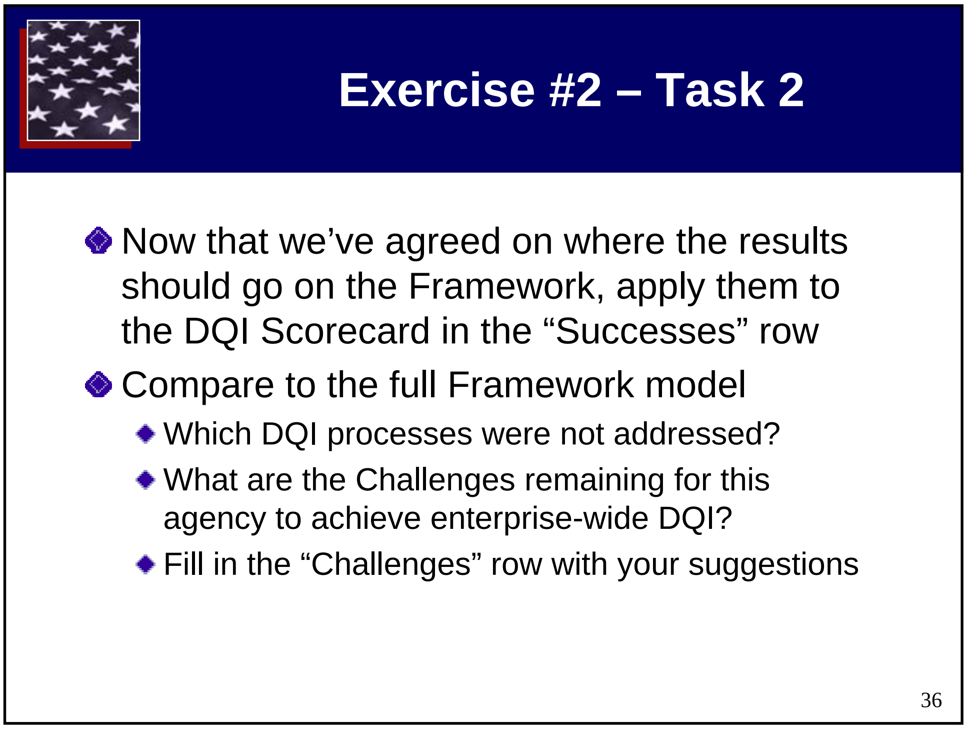
Program area completed necessary reengineering of system to enforce FTE job data entry on a single screen, and business rules across the database were made uniform

Identified database of origin, mapped data entry fields to database locations, & identified business rules (allowable values) for each

"Jobs created" can now be reported to management with 6 sigma accuracy, and steps are being made for improvements in other key business processes

Assessment results saved to MDR staging area

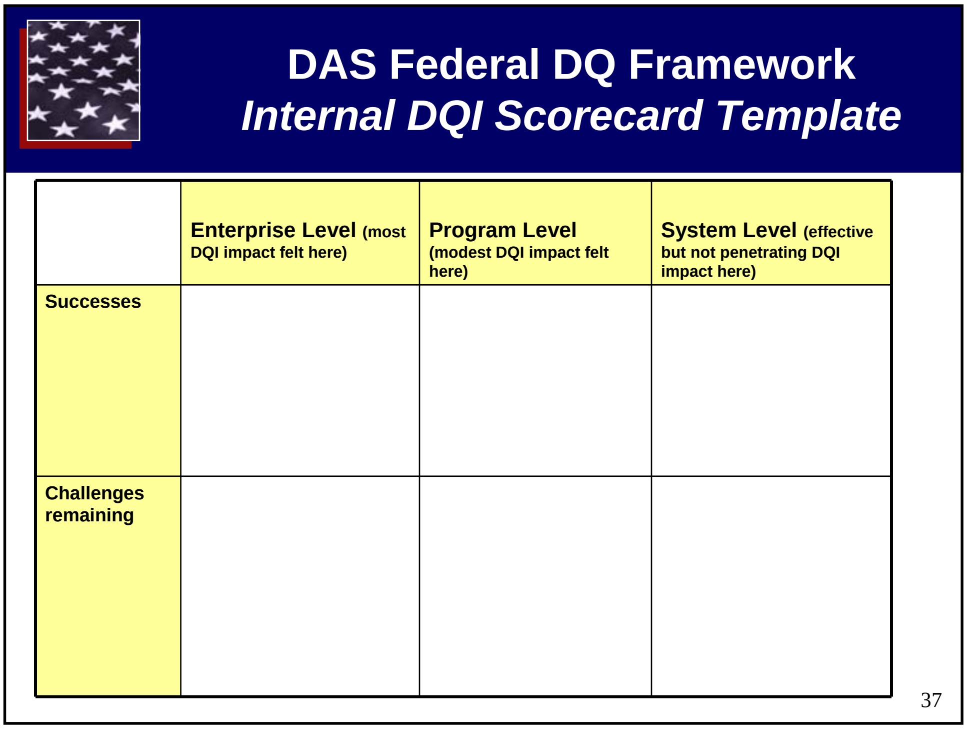
35



Exercise #2 – Task 2

- Now that we've agreed on where the results should go on the Framework, apply them to the DQI Scorecard in the "Successes" row
- Compare to the full Framework model
  - Which DQI processes were not addressed?
  - What are the Challenges remaining for this agency to achieve enterprise-wide DQI?
  - Fill in the "Challenges" row with your suggestions

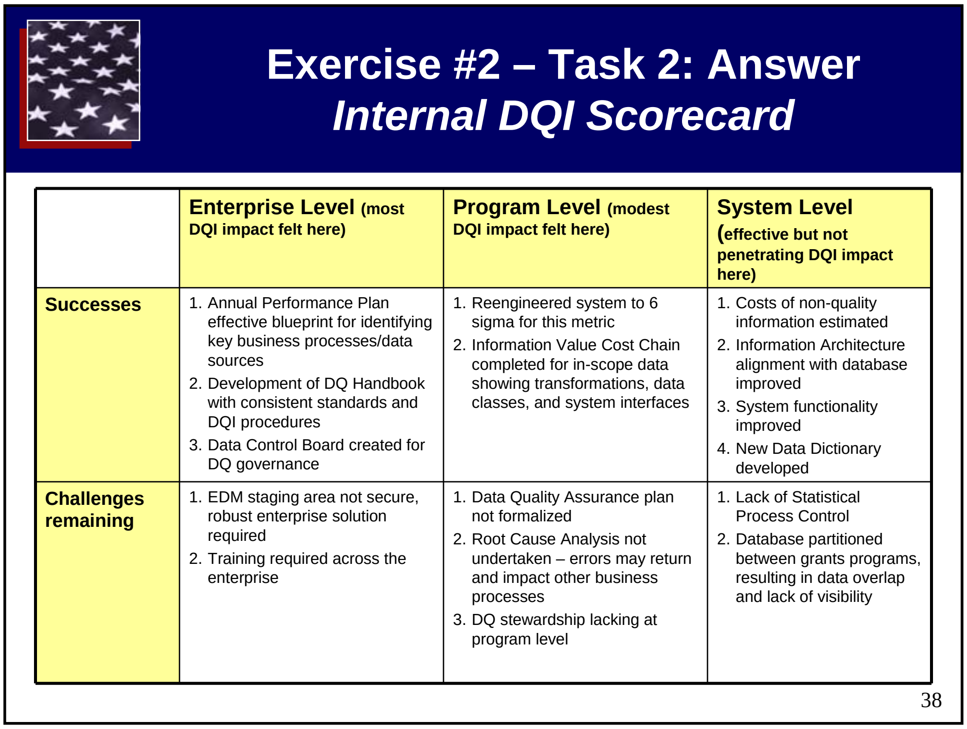
36



DAS Federal DQ Framework
Internal DQI Scorecard Template

	Enterprise Level (most DQI impact felt here)	**Program Level** (modest DQI impact felt here)	**System Level** (effective but not penetrating DQI impact here)
Successes			
Challenges remaining			

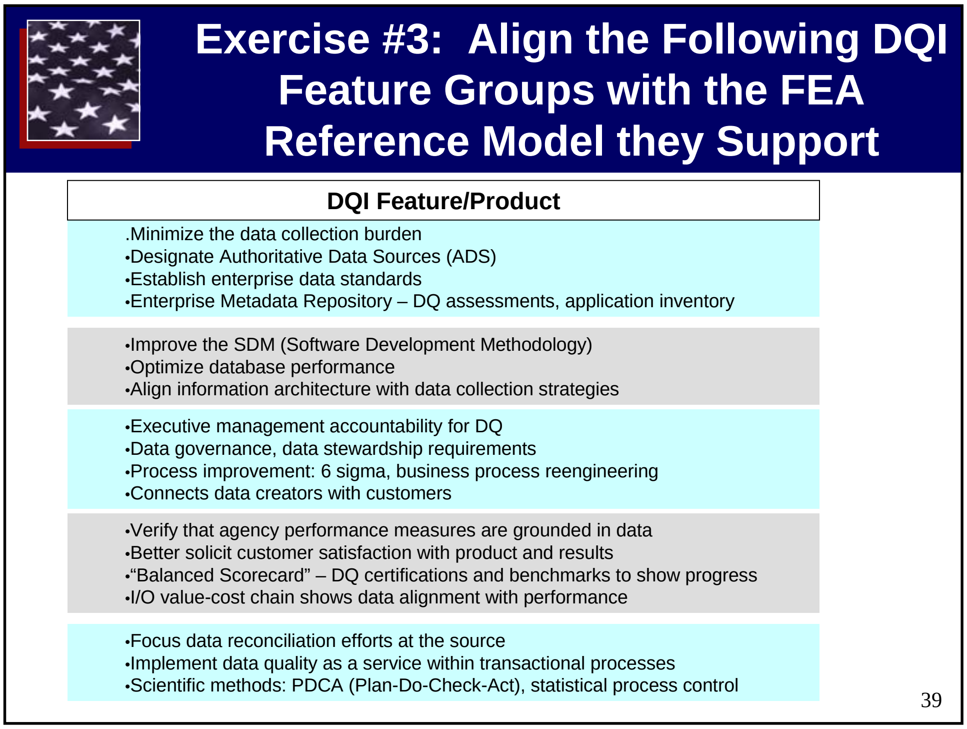
37



Exercise #2 – Task 2: Answer
Internal DQI Scorecard

	Enterprise Level (most DQI impact felt here)	**Program Level** (modest DQI impact felt here)	**System Level** (effective but not penetrating DQI impact here)
Successes	1. Annual Performance Plan effective blueprint for identifying key business processes/data sources 2. Development of DQ Handbook with consistent standards and DQI procedures 3. Data Control Board created for DQ governance	1. Reengineered system to 6 sigma for this metric 2. Information Value Cost Chain completed for in-scope data showing transformations, data classes, and system interfaces	1. Costs of non-quality information estimated 2. Information Architecture alignment with database improved 3. System functionality improved 4. New Data Dictionary developed
Challenges remaining	1. EDM staging area not secure, robust enterprise solution required 2. Training required across the enterprise	1. Data Quality Assurance plan not formalized 2. Root Cause Analysis not undertaken – errors may return and impact other business processes 3. DQ stewardship lacking at program level	1. Lack of Statistical Process Control 2. Database partitioned between grants programs, resulting in data overlap and lack of visibility

38



Exercise #3: Align the Following DQI Feature Groups with the FEA Reference Model they Support

DQI Feature/Product

.Minimize the data collection burden
•Designate Authoritative Data Sources (ADS)
•Establish enterprise data standards
•Enterprise Metadata Repository – DQ assessments, application inventory

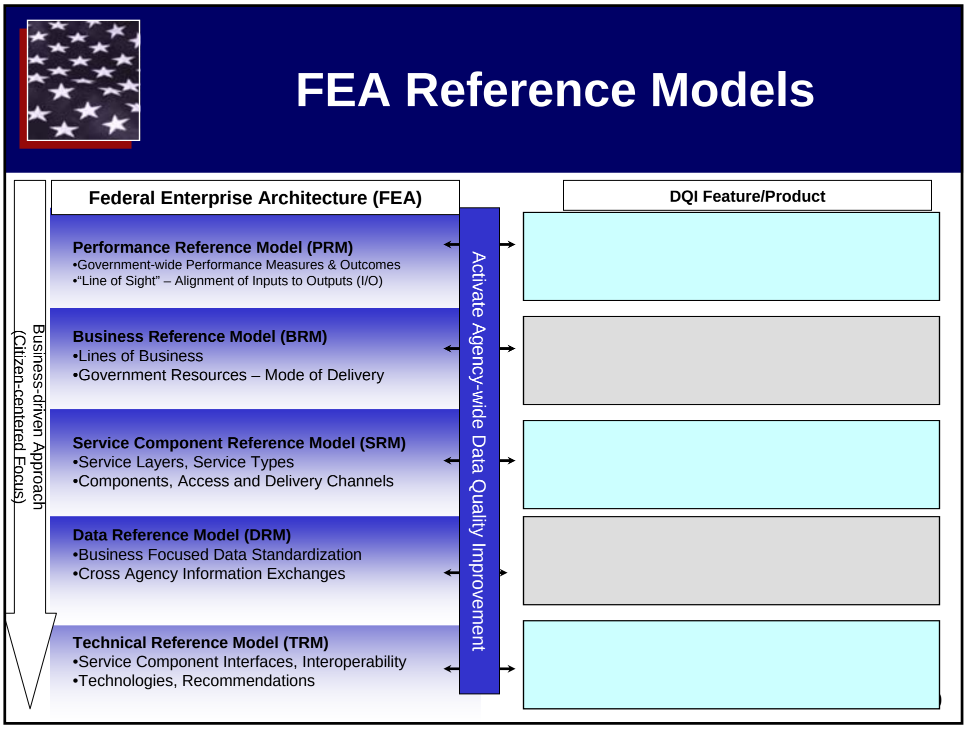
•Improve the SDM (Software Development Methodology)
•Optimize database performance
•Align information architecture with data collection strategies

•Executive management accountability for DQ
•Data governance, data stewardship requirements
•Process improvement: 6 sigma, business process reengineering
•Connects data creators with customers

•Verify that agency performance measures are grounded in data
•Better solicit customer satisfaction with product and results
•"Balanced Scorecard" – DQ certifications and benchmarks to show progress
•I/O value-cost chain shows data alignment with performance

•Focus data reconciliation efforts at the source
•Implement data quality as a service within transactional processes
•Scientific methods: PDCA (Plan-Do-Check-Act), statistical process control

39



FEA Reference Models

Federal Enterprise Architecture (FEA)

Performance Reference Model (PRM)
•Government-wide Performance Measures & Outcomes
•"Line of Sight" – Alignment of Inputs to Outputs (I/O)

Business Reference Model (BRM)
•Lines of Business
•Government Resources – Mode of Delivery

Service Component Reference Model (SRM)
•Service Layers, Service Types
•Components, Access and Delivery Channels

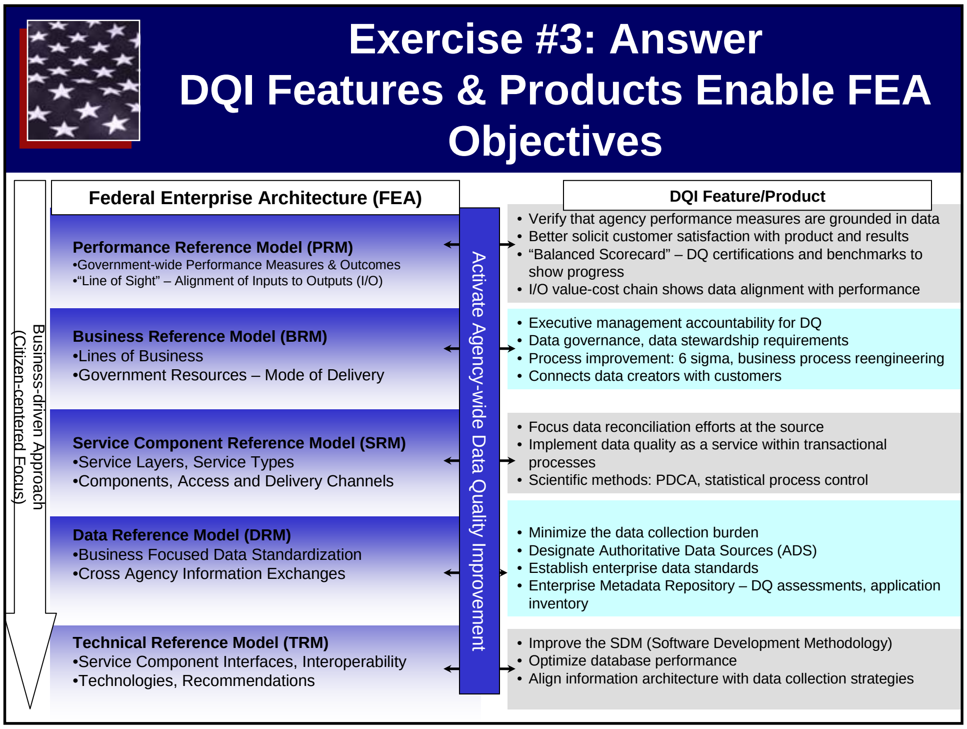
Data Reference Model (DRM)
•Business Focused Data Standardization
•Cross Agency Information Exchanges

Technical Reference Model (TRM)
•Service Component Interfaces, Interoperability
•Technologies, Recommendations

Business-driven Approach
(Citizen-centered Focus)

Activate Agency-wide Data Quality Improvement

DQI Feature/Product

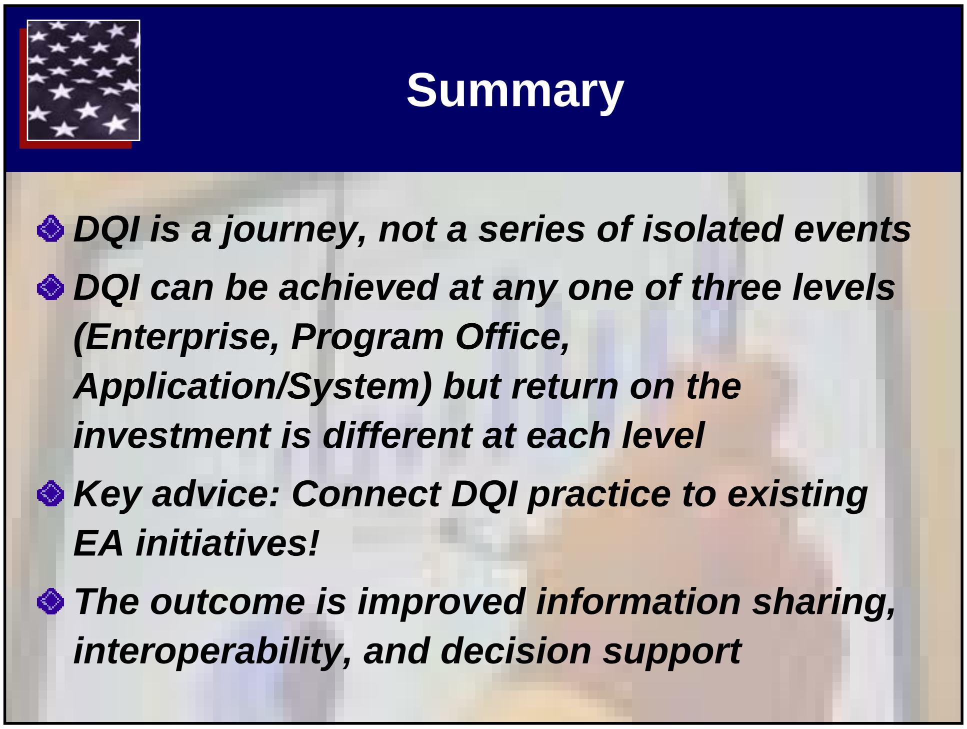


Exercise #3: Answer
DQI Features & Products Enable FEA Objectives

Business-driven Approach (Citizen-centered Focus)

Activate Agency-wide Data Quality Improvement

Federal Enterprise Architecture (FEA)	DQI Feature/Product
Performance Reference Model (PRM) •Government-wide Performance Measures & Outcomes •"Line of Sight" – Alignment of Inputs to Outputs (I/O)	• Verify that agency performance measures are grounded in data • Better solicit customer satisfaction with product and results • "Balanced Scorecard" – DQ certifications and benchmarks to show progress • I/O value-cost chain shows data alignment with performance
Business Reference Model (BRM) •Lines of Business •Government Resources – Mode of Delivery	• Executive management accountability for DQ • Data governance, data stewardship requirements • Process improvement: 6 sigma, business process reengineering • Connects data creators with customers
Service Component Reference Model (SRM) •Service Layers, Service Types •Components, Access and Delivery Channels	• Focus data reconciliation efforts at the source • Implement data quality as a service within transactional processes • Scientific methods: PDCA, statistical process control
Data Reference Model (DRM) •Business Focused Data Standardization •Cross Agency Information Exchanges	• Minimize the data collection burden • Designate Authoritative Data Sources (ADS) • Establish enterprise data standards • Enterprise Metadata Repository – DQ assessments, application inventory
Technical Reference Model (TRM) •Service Component Interfaces, Interoperability •Technologies, Recommendations	• Improve the SDM (Software Development Methodology) • Optimize database performance • Align information architecture with data collection strategies



Summary

- ***DQI is a journey, not a series of isolated events***
- ***DQI can be achieved at any one of three levels (Enterprise, Program Office, Application/System) but return on the investment is different at each level***
- ***Key advice: Connect DQI practice to existing EA initiatives!***
- ***The outcome is improved information sharing, interoperability, and decision support***

Questions

Contact info:

Mark Amspoker
Sr. Data Analyst, ATS Corporation
E-mail: mamspoker@atsc.com

Suzanne Acar
FBI Senior Advisor, and
Co-chair, Federal Data Architecture Subcommittee, and
Chair, Data.gov Agency POC Working Group
E-mail: suzanne.acar@ic.fbi.gov

43