



# **Proceedings of the Second MIT Information Quality Industry Symposium**

Massachusetts Institute of Technology  
Cambridge, MA, USA

July 16-17, 2008

Edited by

**WooYoung Chung**

St. John Fisher College, USA

**Suzanne Acar**

U.S. Department of the Interior, USA

**Linda Kresl**

Mentor Graphics, USA

Copyright © 2008 MIT Information Quality Program, Massachusetts Institute of Technology

All presentations included in these proceedings are published with explicit permission from their author or copyrights holder. These presentations are protected by the international and U.S. copyright law. Any copying, distribution, exhibition in public or private meetings, and/or pursuing any derivative work, by any third party, in part or as a whole, is strictly prohibited without explicit written permission of the original copyrights holder. Any third party who wishes to quote or refer to the information presented in these proceedings must comply with the fair use provisions of the U.S. copyright law 17 U.S.C. §107 and must properly cite the author and these proceedings as its source.

## FOREWORD

WELCOME to the 2<sup>nd</sup> *MIT Industry Symposium on Information Quality (MIT2008IQIS)*. *Information Quality* (IQ) is an acknowledged prerequisite to organizational success. However, organizations also face the challenge of balancing quality requirements including an increasing volume of information and the demand for faster delivery, while contending with constantly changing business environments. The *Symposium* is designed to help bridge these issues through discussions among practitioners, vendors, and academicians. In addition to presentations and workshops, the Symposium includes vendor presentations, product announcements, and consultancy methods that complement the annual *International Conference on Information Quality (ICIQ)*.

Acknowledgments are extended to all Symposium participants. We thank you for your contribution in establishing IQ as a multi-disciplinary field and industry. Members of the Symposium organizing committee worked hard; their contributions made our work on the program a pleasure. We also wish to thank Lockheed Martin Corporation, UTi Worldwide, and Acxiom Corporation for their continuing sponsorship of the MIT IQ Consortium.

Thanks are also due to, among others, DAMA, EWSolutions and DM Forum for their support in promoting the Symposium.

At the MIT 2007 Information Quality Industry Symposium, Dr. Bruce Davidson (Cedars-Sinai Health System) and CDR Stanley Dobbs (U.S. Navy) were recognized for their outstanding contributions.

We would like to express our gratitude to Fori Wang, Li-Hsin Chang, and other MIT staff for their assistance in producing the Symposium proceedings, keeping the website up and managing the various aspects of the Symposium operation. The final Symposium program and other information about the Symposium are available at <http://mitiq.mit.edu/IQIS>.

## Symposium Organizing Committee

<u>Program Co-Chairs</u>	
• Suzanne Acar	• Department of the Interior
• Linda Kresl	• Mentor Graphics
• Marcus Gebauer	• WestLb BankGermany
<u>Publicity Chairs</u>	
• Jeff Fried	• FAST
• Peter Aiken	• VCU/Data Blueprint
• Michael Mielke	• De Bahn & DGIQGermany
<u>Logistics &amp; Facilities Chair</u>	
• Jarl S. Magnusson	• DNV
<u>Vendor Exhibit Chair</u>	
• Lisa Dodson	• SAS/DataFlux
• Harald Smith	• IBM
<u>Placement Chair</u>	
• Alba Alemán	• Citizant Inc.
<u>Proceedings Chair</u>	
• Bill Burke	• Wilmington Trust Company
<u>General Chairs</u>	
• Anne Marie Smith	• EWSolutions/Northcentral Univ.
• WooYoung Chung	• St. John Fisher College
• Rich Wang	• MIT Information Quality Program
<u>Advisor Committee</u>	
• Peter Aiken	• VCU/Data Blueprint
• Bruce Davidson	• Cedars-Sinai Health System
• Yang Lee	• Northeastern University
• Stuart Madnick	• MIT Sloan School of Management



TABLE OF CONTENTS		
<b>WEDNESDAY, July 16, 2008</b>		
<b>8:30AM – 12PM</b>	<b>Complimentary Pre-Symposium Invited Tutorial Track 1</b>  <b>Information Quality: What's Enterprise Architecture Got To Do with It? .....</b> Kathie Sowell, Custom Enterprise Solutions, LLC  <b>Information Quality for Business Intelligence.....</b> Earl Hadden, Intelligence Commerce Network	<b>E51-345</b>  <b>Pg. 1</b>  <b>Pg. 11</b>
<b>8:30AM - 12PM</b>	<b>Complimentary Pre-Symposium Invited Tutorial Track 2</b>  <b>Maturing Data Quality to Information Quality to Business Quality .....</b> Larry English, Information Impact International Inc.  <b>Data Driven: Profiting from Your Most Important Business Asset .....</b> Thomas Redman, Navesink Consulting Group	<b>E51-372</b>  <b>Pg. 23</b>  <b>Pg. 41</b>
<b>8:30AM- 12 PM</b>	<b>Complimentary Pre-Symposium Invited Tutorial Track 3</b>  <b>The NATO Codification System as the Foundation of the ECCMA Open Technical Dictionary.....</b> Steven Arnett, U.S. National Codification Bureau  <b>Meeting the Requirements of ISO 8000-110:2008 Master Data Quality .....</b> Peter Benson, Electronic Commerce Code Management Association	<b>E51-376</b>  <b>Pg. 74</b>  <b>Pg. 115</b>
<b>8:30AM</b>	<b>Pre-Symposium UHC Workshop, by invitation only</b>	
<b>12 - 12:45PM</b>	<b>First-of-its-kind Master and Ph.D. IQ degree programs &amp; certificate courses at UALR</b> John Talburt, University of Arkansas at Little Rock Elizabeth Pierce, University of Arkansas at Little Rock	
<b>1 - 2PM</b>	<b>Chairs' Welcome and Opening Remarks</b> Anne Marie Smith, WooYoung Chung, Richard Wang Conference Co-Chairs Suzanne Acar, Linda Kresl, Michael GeBauer Program Co-Chairs  <b>Keynote Speech</b> <b>Arkansas Leading the Information Quality Economy and Education</b> Michael Beebe, The Honorable Governor of Arkansas	<b>E51-345</b>
<b>2 - 2:30PM</b>	<b>COFFEE BREAK</b>	

2:30 - 4PM	<b>Session 1A HEALTHCARE IQ</b> Moderator: Bruce Davison, Cedars-Sinai Health System  <b>Data Quality in Healthcare Comparative Databases .....</b> Pg. 127 Steve Meurer, University HealthSystem Consortium <b>Galaxy Data Quality Program .....</b> Pg. 144 Laura Sebastian-Coleman, Ingenix/United Health Analytics <b>Can Accountants Help Reduce Medication Errors? .....</b> Pg. 151 Scott Boss, Bentley College Janis Gogan, Bentley College James Hunton, Bentley College	E51-345
2:30 - 4PM	<b>Session 1B EA, IQ and IPMAP</b> Moderator: Edwin Nassiff, Lockheed Martin Corporation  <b>Using Data Quality Methods at the Federal Railroad Administration: Improving the Archiving and Retrieval of Safty Information.....</b> Pg. 161 Scott Bernard, US Department of Transportation – FRA Mark Trimble, US Department of Transportation – FRA <b>Embedding Information Quality in the Lockheed Martin Enterprise Architecture Framework (LEAF): An IPMAP Approach .....</b> Pg. 171 Edwin Nassiff, Lockheed Martin Corporation Paul Pierson, Lockheed Martin Corporation John Slone, Lockheed Martin Corporation	E51-372
2:30 - 4PM	<b>Session 1C VENDOR IQ SOLUTIONS</b> Moderator: Lisa Dodson, SAS/DataFlux  <b>Patterns in Data Quality .....</b> Pg. 179 Michael Overturf, Pitney Bowes Group 1 Software Navin Sharma, Pitney Bowes Group 1 Software <b>Rapid Corporate Growth and Information .....</b> Pg. 186 Steve Sarsfield, Harte-Hanks Trillium Software <b>What's in a Name?: Cutural Name Recognition and Data Quality.....</b> Pg. 199 Mala Narasimharajan, IBM Software Group	E51-376
4:30 – 6PM	<b>Session 2A INVITED FEDERAL IQ PRACTICE</b> Moderator: Suzanne Acar, U.S. Department of the Interior  <b>Meaningful Engagement for Information Quality .....</b> Pg. 207 Glenn Norton, U.S. Department of Homeland Security	E51-345
4:30 – 6PM	<b>Session 2B IQ CONSULTING METHODS</b> Moderator: Jeff Fried, FAST  <b>Ten Steps to Data Quality and Trusted Information .....</b> Pg. 217 Danette McGilvray, Granite Falls Consulting Inc. <b>Using Conceptual Data Modeling to Ensure High Data and Information Quality .....</b> Pg. 236 Pete Stiglich, EWSolutions <b>Improving Your Data Warehouse's IQ .....</b> Pg. 278 Derek Strauss, Gavroshe USA, Inc	E51-372

6:15 – 7:45PM	<b>DAMA International / ICCP Beta Data &amp; Information Quality Exam</b>  <b>NOTE: MIT IQ Program provides the venue as a benefit to all parties. However, this should NOT be perceived or promoted as an MIT endorsement.</b>	E51-345
---------------	---	---------

<b>THURSDAY, July 17, 2008</b>		
<b>8:30AM</b>	<b>CONTINENTAL BREAKFAST</b>	
<b>9:30 – 11AM</b>	<b>Session 3A INVITED INDUSTRY IQ PRACTICE</b> Moderator: Anne Marie Smith, MIT2008IQIS Co-Chair  <b>Implementing a Successful Enterprise Data Quality Initiative .....</b> Pg. 292 David Marco, EWSolutions	E51-345
<b>9:30 – 11AM</b>	<b>Session 3B INVITED INDUSTRY IQ PRACTICE</b> Moderator: Linda Kresl, MIT2008IQIS Co-Chair  <b>Successfully Applying Data Quality in Data Governance .....</b> Pg. 321 Ivan Chong, Informatica	E51-372
<b>9:30 – 11AM</b>	<b>Session 3C INVITED INDUSTRY IQ PRACTICE</b> Moderator: Paul Prabhaker, Northern Illinois University  <b>Geography Reference Data Services.....</b> Justin Magruder, Freddie Mac Diane Schmidt, Freddie Mac Xinhua Chris Deng, Freddie Mac <b>Federal Data Quality Guide .....</b> Mark Amspoker, HUD & Citizant, Inc.	E51-376       Pg. 336       Pg. 346
<b>11AM – 12:50PM</b>	<b>VENDOR EXHIBIT &amp; PRESENTATION LUNCH BREAK</b>	E51-345, 361, 273, 276
<b>1 – 2:30PM</b>	<b>Session 4A CIO IQ PANEL</b> Moderator: WooYoung Chung, St. John Fisher College  <b>Data Governance for Improved Information Quality.....</b> Pg. 358 Joseph Bugajski, Burton Group Robert Grossman, Open Data Group <b>Raising the Bar: DQ/IQ to "Enterprise IQ" .....</b> Pg. 378 Garry Darrer, Getinge USA, Inc.	E51-345
<b>1 – 2:30PM</b>	<b>Session 4B EMERGING IQ ISSUES</b> Moderator: Peter Aiken, Data Blueprint & VCU  <b>"Fit for Use" to a Fault.....</b> Pg. 396 Deborah Henderson, CapGemini; DAMA Foundation Tamdum Lett, Sullivan & Cromwell, LLP Anne Marie Smith, EWSolutions Cora Zeeman, University of Toronto	E51-372

	<b>Quality Information as a Service for SOA .....</b> <b>Pg. 410</b> Linda Kresl, Mentor Graphics <b>Unified Architecture for Integrating Intelligence Data .....</b> <b>Pg. 425</b> Suzanne Yoakum-Stover, Potomac Institute for Policy Studies Tatiana Malyuta, New York City College of Technology	
<b>1 – 2:30PM</b>	<b>Session 4C INDUSTRY IQ PRACTICE</b> Moderator: Elizabeth Pierce, University of Arkansas at Little Rock  <b>Application of Practical Nominalism to Data Management .....</b> <b>Pg. 436</b> Fulton Wilcox, Colts Neck Solutions LLC <b>Data Governance with a Focus on Information Quality .....</b> <b>Pg. 450</b> Gwen Thomas, The Data Governance Institute	<b>E51-376</b>
<b>2:30 – 3PM</b>	<b>COFFEE BREAK</b>	
<b>3 – 4:30PM</b>	<b>Session 5A IQ STANDARDS</b> Moderator: Willa Pickering, Lockheed Martin Corporation  <b>ISO 8000 the International Standard for Data Quality .....</b> <b>Pg. 459</b> Peter Benson, Electronic Commerce Code Management Association (ECCMA) <b>Data Integration and Data Quality: Pharmaceutical Industry Case .....</b> <b>Pg. 465</b> Sergiy Sirichenko, Regeneron Pharmaceuticals, Inc Vadim Tantsyura, Regeneron Pharmaceuticals, Inc Olive Yuan, Regeneron Pharmaceuticals, Inc	<b>E51-345</b>
<b>3 – 4:30PM</b>	<b>Session 5B INTERNATIONAL IQ PANEL</b> Moderator: Michael Mielke, Die Bahn & DGIQ, Germany  <b>MDM Enterprise Analyzer: a framework to support centralized and local master data quality analysis.....</b> <b>Pg. 478</b> Kai-Uwe Baryga, SYDECON	<b>E51-372</b>
	<b>END OF MIT 2008 IQ INDUSTRY SYMPOSIUM</b>	
<b>4:30 – 6PM</b>	<b>DAMA International / ICCP Beta Data &amp; Information Quality Exam</b>  NOTE: MIT IQ Program provides the venue as a benefit to all parties. However, this should NOT be perceived or promoted as an MIT endorsement.	<b>E51-345</b>
<b>8:30 – 1PM</b>	<b>POST-SYMPOSIUM OMG WORKSHOP ON DATA QUALITY STANDARDS</b> Peter Aiken Harsh Sharma	



The MIT 2008 Information Quality Industry Symposium



## Information Quality: What's Enterprise Architecture Got to Do with It?

P. Kathie Sowell  
Custom Enterprise Solutions, LLC



The MIT 2008 Information Quality Industry Symposium



### Objectives of this presentation

- Demonstrate the compatibility of the data quality discipline with Scenario-Based Enterprise Architecture

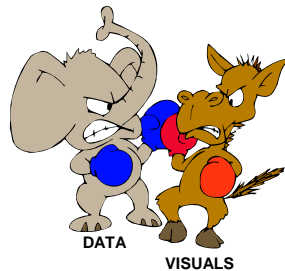


The MIT 2008 Information Quality Industry Symposium



## What's the Problem?

- “It’s the data, stupid” vs. “Let the pictures tell the story”



3



The MIT 2008 Information Quality Industry Symposium



## What does “It’s the data, stupid” really mean?

- The (only) important aspect of an enterprise architecture is the underlying data.
- It doesn’t matter how you express this data to humans.
- It doesn’t matter *if* you express this data to humans.
- It is only important that the data conform to the data standards you have set up
  - format standards for storing in a database
  - quality standards for usability

4



The MIT 2008 Information Quality Industry Symposium



## What does “Let the pictures tell the story” really mean?

- The underlying data has to be “good,” but
- Humans think and understand quickly and well via visuals (pictures).
- Humans are the ones analyzing enterprise architecture data.
- Humans are the ones making decisions based on analysis of the enterprise architecture data.
- Most of these decisionmakers are not computer scientists.

5

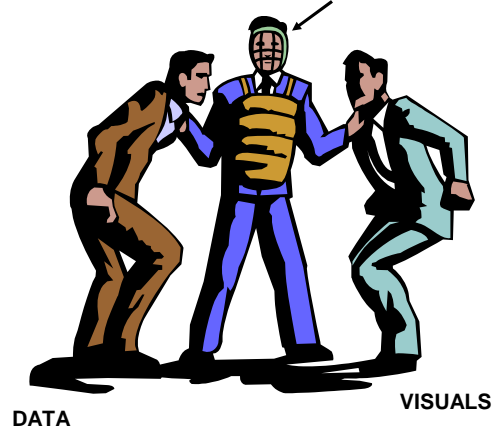


The MIT 2008 Information Quality Industry Symposium



But wait, they are both right.

**SCENARIO-BASED ENTERPRISE ARCHITECTURE:  
VISUAL, DATA-FOCUSED, TELLS A STORY**



6



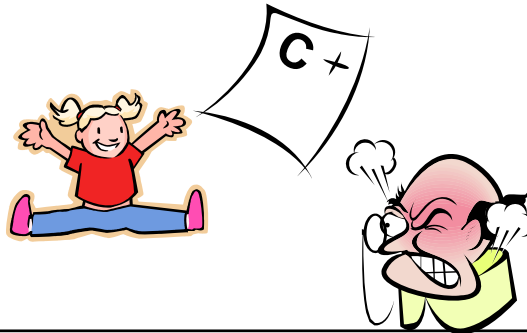
The MIT 2008 Information Quality Industry Symposium



## How can they both be right?

### The definition of data “quality” is circumstantial

- Data quality depends on where, when, why, how, and by whom the data needs to be used.
- One person’s “good enough” is another person’s disaster.



7



The MIT 2008 Information Quality Industry Symposium



### Enterprise architecture can help us account for this circumstantial definition of data quality

- Enterprise architecture “products” or “artifacts” are the visual renderings of selected data about your enterprise.
- Visual artifacts allow human stakeholders and decisionmakers to quickly grasp the logic of your message and analyze its validity and repercussions.
- The different circumstances under which data is to be used can be expressed as different story lines.
- To tell these different stories, we need a sequential, visual representation of our underlying enterprise architecture data.

Combine the discipline of data quality with Scenario-Based Enterprise Architecture

8





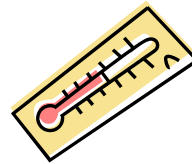
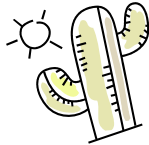
The MIT 2008 Information Quality Industry Symposium



## What is Scenario-Based Architecture?



- A representation of the various ways a given enterprise operates under different sets of conditions (circumstances)



- Examining a range of scenarios can help you determine if your enterprise (and its data) is robust enough to operate under the likely circumstances.

9



The MIT 2008 Information Quality Industry Symposium



## What are the basic components of a Scenario-Based Architecture?

- **Purpose Statement:** Tells what you intend to analyze via the architecture
- **Activity Model:** Shows the essential activities that occur, under any and all circumstances (i.e., irrespective of specific circumstances)
- **Node Connection Model:** Shows which business performers exchange information, irrespective of specific circumstances
- **Information Exchange Matrix:** Shows the detailed characteristics of the information exchanged
- **Scenario Sequence Models:** Illustrate multiple storylines showing the different ways the enterprise operates under specific conditions
- **Capability Progression Model:** Defines what it means to achieve certain levels of capability
- And, if you need details about technology used: \*
  - **Systems Connection Model**
  - **Systems Data Exchange Matrix**

\* For illustration, we will not consider technology factors here

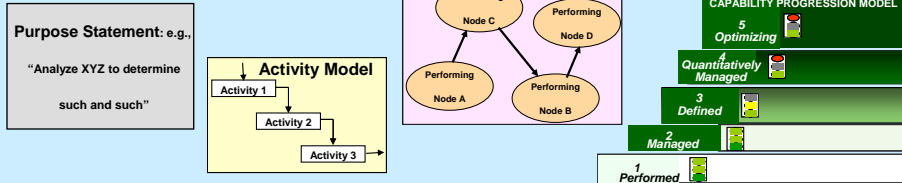
10



The MIT 2008 Information Quality Industry Symposium



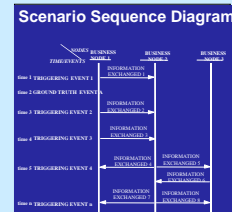
What does a Scenario-Based Architecture with these components look like?

**One each per architecture**

These represent the whole of the enterprise under consideration.

**One each per scenario**

Information Exchange Matrix					
Data Item	Sender	Receiver	Timeliness Reqmt.	Precision Reqmt.	Other Reqmt.



These tap into the whole of the enterprise information to select threads that illustrate specific story lines.

11



The MIT 2008 Information Quality Industry Symposium



Purpose Statement sets the stage for your enterprise analysis

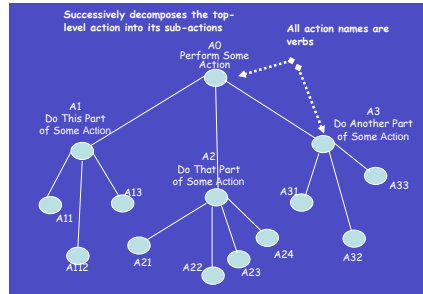
- Why you are developing the architecture
- What issues you will examine, what questions you hope to answer
- Who are your stakeholders, decisionmakers
- What artifacts (models) you will construct
- How you will approach and tailor the models
- How you will know when you are finished
- How you will know if you have succeeded



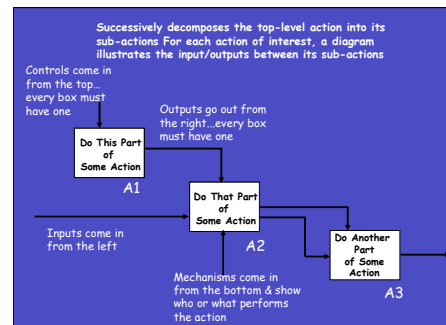
The MIT 2008 Information Quality Industry Symposium



The Activity Model shows the relevant\* actions that take place in your enterprise (irrespective of scenario)



Activity Hierarchy Tree



Activity Flow Model

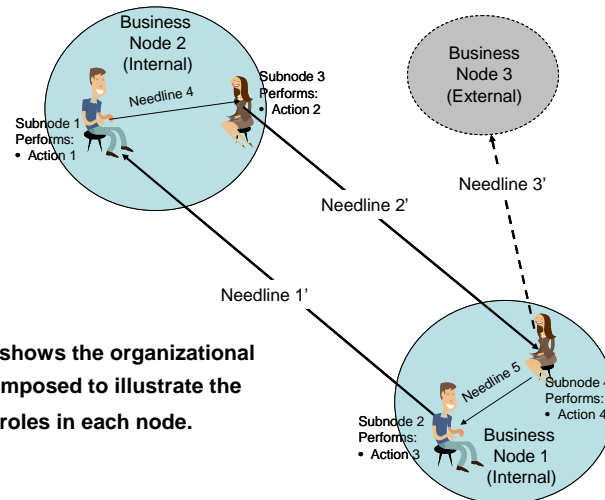
\* Relevant to the purpose & scope of the architecture



The MIT 2008 Information Quality Industry Symposium



The Node Connection Model shows which enterprise participants need to interact with each other (irrespective of scenario)



This example shows the organizational nodes decomposed to illustrate the human roles in each node.



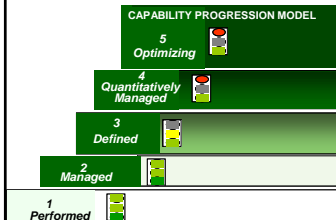
The MIT 2008 Information Quality Industry Symposium



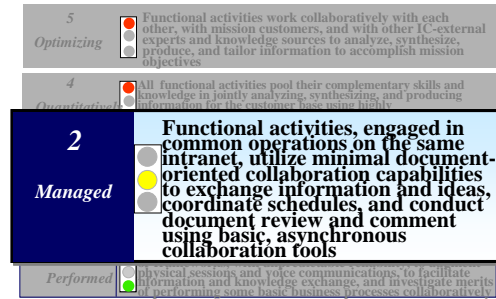
The Capability Progression Model (CPM) defines levels of ability in selected capability areas (irrespective of scenario)

#### Example: CPM of Collaboration Capability

##### Generic Format



*Level 2 of the Collaboration Capability is Determined to be the Target Capability -- and the First Segment of the Capability Profile*

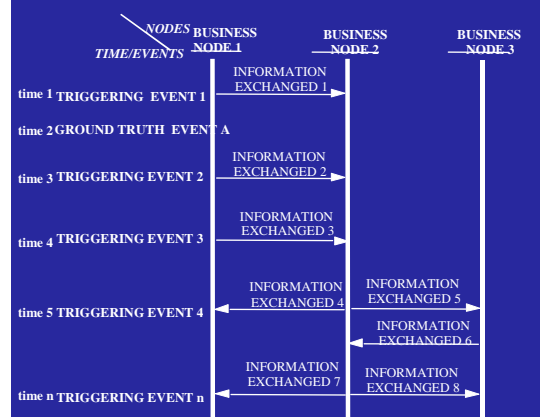


The MIT 2008 Information Quality Industry Symposium



A Scenario Sequence Diagram shows a series of events, and the information exchanges that occur in response to the events of a given scenario

### Scenario Sequence Diagram



12



The MIT 2008 Information Quality Industry Symposium



The Information Exchange Matrix captures the relevant quality (and other) characteristics of information *as it is used in a given scenario*

Identifier/ Name of Needline Supported	Identifier/ Name of Information Exchange	Nature of Transaction						Purpose/ Triggering Event	Information Source			Information Destination		
		Mission/ Scenario	Language (For Multi- National Operations)	Content	Size/ Units	Media (Voice, Text, Data, Imagery, Physical)	Collabo- rative or One- Way?	Interoper- ability Level Required	ID of Producing Node	Owner/ Organization of Node	Name of Producing Action	ID of Receiving Node	Owner/ Organization of Node	Name of Receiving Action
1	e.g., 1-a													
2	e.g., 2-b													
...	...													
n	e.g., n-c													

C O N T I N U E D	Identifier/ Name of Needline Supported	Identifier/ Name of Information Exchange	Performance Requirements			Information Assurance Attributes					Threats		
			Frequency (# per Unit Of Time)	Timeliness	Other	Classification/ Declassification Restrictions	Criticality/ Priority	Integrity Checks Required	Assured Authorization to Send/ Receive	Other	Physical	Electronic (jamming, hackers, etc.)	Political/ Economic
	1	e.g., 1-a		Time- liness	Data Qual- ity							Adversarial	Environmental
	2	e.g., 2-b											
	...	...											
	n	e.g., n-c											

Who needs what information or goods may differ by scenario.  
Required characteristics of that information or data may differ by scenario.

14



The MIT 2008 Information Quality Industry Symposium



### But where does data quality fit in?

- Data quality depends on where, when, why, how, and by whom the data needs to be used.
- The various scenarios illustrate where, when, why, how, and by whom the data needs to be used, one storyline at a time.
- The Capability Progression Model provides a scale for defining capabilities related to data quality (and other factors).
- The Information Exchange Matrix details the characteristics, including quality attributes, of information as it is used in these various circumstances
- Examination of the Information Exchange Matrix in context with the Capability Progression Scale allows the architect to define "success" for each scenario.

13

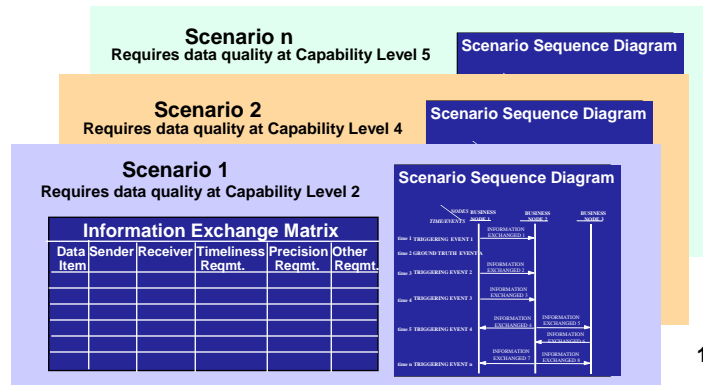


The MIT 2008 Information Quality Industry Symposium



By examining a representative range of these scenarios and their data quality requirements, the architect can measure the range of quality requirements for given information items. **For example:**

- “Depending on circumstances, data item X needs to be..”
  - from one minute to one hour old
  - validated by a level one manager to a level three manager
  - precise to a level of one decimal place to three decimal places



The MIT 2008 Information Quality Industry Symposium



## Summary

- Yes, it is the data that is important.
- Yes, it is the visual representation of that data that is important.
- The quality of the data depends on the circumstances.
- Visual, Scenario-Based Enterprise Architecture helps you explain the circumstances and the resulting data quality assessment to human decisionmakers.





The MIT Information Quality Industry Symposium, 2008



## **Information Quality for Business Intelligence**

### **Projects**

**Earl Hadden  
Intelligent Commerce Network LLC**



The MIT Information Quality Industry Symposium, 2008



### **Objectives of this presentation**

- Understand Information Quality Problems on BI/DW Projects
- Define Strategic and Tactical Approaches to addressing Information Quality Problems
- Demonstrate how TDQM methods can augment BI/DW methodologies



The MIT Information Quality Industry Symposium, 2008



## What's the Problem

- “Data quality is the most significant problem in our efforts to integrate information.” Al Albhorn, consultant to the Chief Architect, Department of Homeland Security
- The cost of non-quality is 5% of US GDP
- In service companies, information non-quality costs can cost up to 20% of gross revenue



The MIT Information Quality Industry Symposium, 2008



## Defining Information Quality

DQ Category	DQ Dimensions
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Access, Security
Contextual DQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of data
Representational DQ	Interpretability, Ease of understanding, Concise representation, Consistent representation





The MIT Information Quality Industry Symposium, 2008



### **Top 3 BI/DW Information Quality Problems**

1. Believability – international steel manufacturer with multiple production schedules
2. Completeness – health insurance provider with over 50% of claims records incomplete
3. Timeliness – multinational bank spent US\$15 million on DW, warehouse is available on the 15<sup>th</sup> day after the close of the month, business information required by the 5<sup>th</sup>



The MIT Information Quality Industry Symposium, 2008



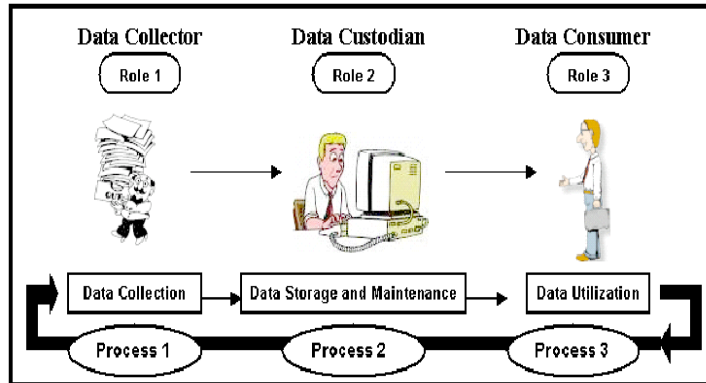
### **Data Quality Problems vs. Information Quality Problems**



The MIT Information Quality Industry Symposium, 2008



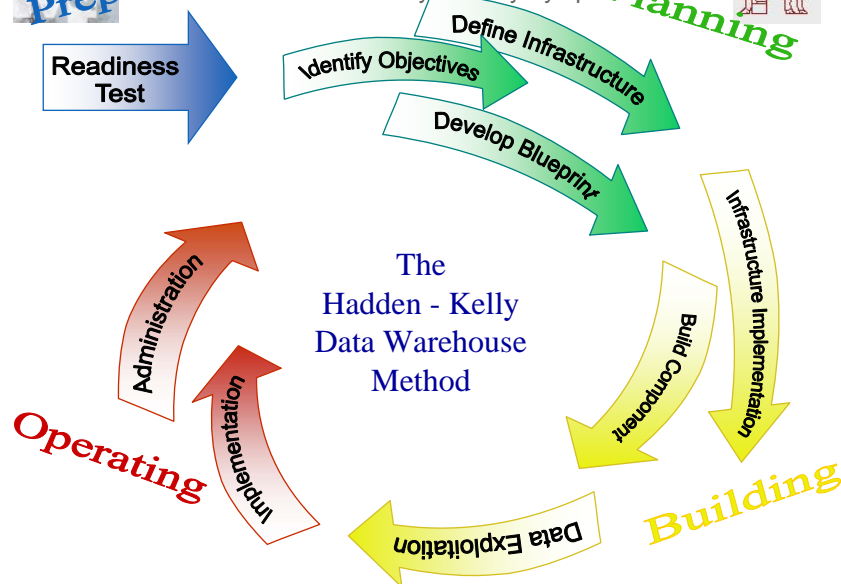
## Where's the Problem?



Source: Prof. Richard Wang



The MIT Information Quality Industry Symposium, 2008





The MIT Information Quality Industry Symposium, 2008



## Choosing your IQ Path

- Enterprise-wide, executive sponsorship

Advantages

- Broad sponsorship
- Organizations tend to “stay the course”

Disadvantages

- Hard sell
- Expensive
- Takes a long time to get measurable results



The MIT Information Quality Industry Symposium, 2008



## Choosing your IQ Path

- Subject based – “middle out”

Advantages

- Can be tied to a specific project with business goals, benefits
- Eliminates a lot of time wasted on data of lesser importance

Disadvantages

- Hard to get business units not directly receiving value to participate (therefore limits value)
- Adds time (and costs) to integration projects
- Takes a long time to get measurable results



The MIT Information Quality Industry Symposium, 2008



## Choosing your IQ Path

- “Bottom up” – data cleansing
  - Advantages
    - Limited to organization units directly involved in the project
    - Can be done in “stealth” mode
  - Disadvantages
    - May create as many problems as it solves – multiple versions of the truth, conflicting rules...
    - Tends to get lost when the deadline approaches



The MIT Information Quality Industry Symposium, 2008



## Choosing your IQ Path

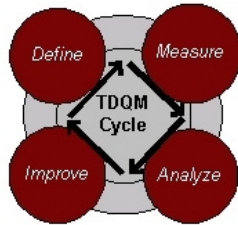
- Do nothing
  - Advantages
    - BI results match production system reports/queries
  - Disadvantages
    - Lack of “believability” compromises use of the BI solution



The MIT Information Quality Industry Symposium, 2008



## Enterprise-wide – executive sponsor



- **Define** and establish data quality to be
  - A multi-dimensional concept beyond accuracy
  - Both objective and subjective
- **Measure** DQ with software tools such as
  - *Integrity Analyzer* and *Information Quality Assessment*
- **Analyze** DQ with models, methods & principles
  - Modeling Information Manufacturing Systems to deliver high-quality information products
- **Improve**



The MIT Information Quality Industry Symposium, 2008



## Subject based – “middle out”

- Identify business objectives for the BI/DW project
- Identify organization units involved
- Identify other stakeholders interested in the outcome
- Identify information needed by the organization units and stakeholders to ensure the objectives are met



The MIT Information Quality Industry Symposium, 2008



## Subject based – “middle out”

- Establish IQ Environment (Policies, Roles & Responsibilities, etc.)
- Conduct preliminary information quality assessment
- Determine where the information is needed
- Identify technology that will be used to deliver the information
- Develop the project plan for the BI/DW implementation project



The MIT Information Quality Industry Symposium, 2008



## Subject based – “middle out”

“The only way to achieve integration is to work from a common data model.”

-- John Zachman



The MIT Information Quality Industry Symposium, 2008



## Subject based – “middle out”

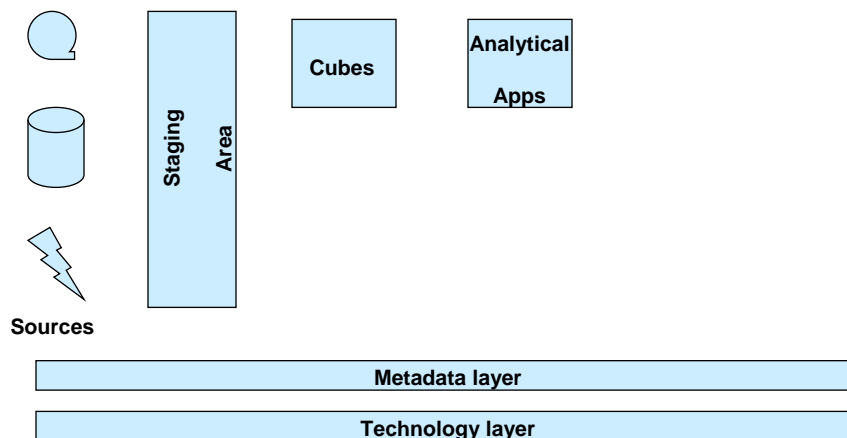
- Identify attributes required to provide desired information
- Define IQ standards for each attribute
- Perform source analysis for each attribute
- Establish sourcing logic (if there are multiple candidate sources)
- Define extract and transform specifications



The MIT Information Quality Industry Symposium, 2008



## Typical BI Architectural Model

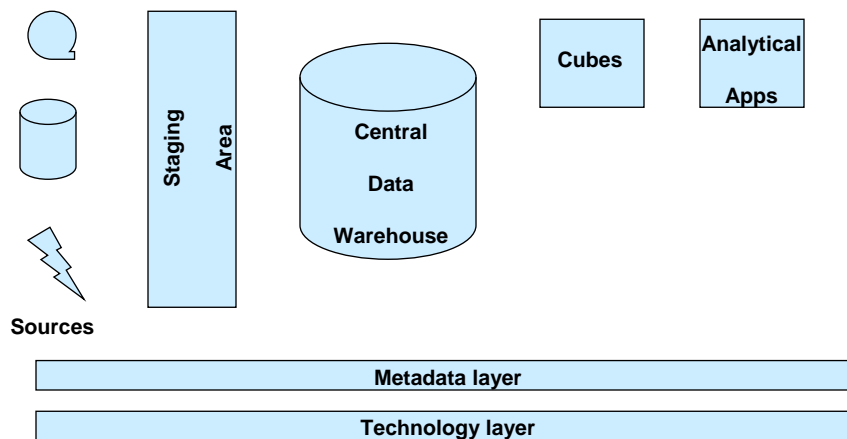




The MIT Information Quality Industry Symposium, 2008



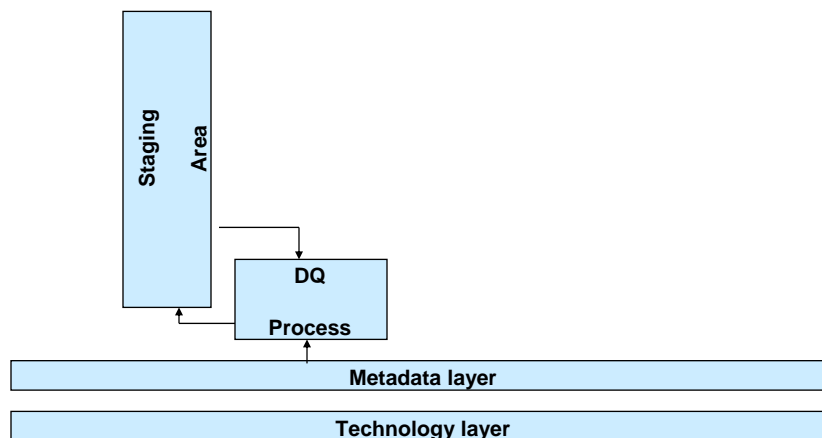
## Typical DW Architectural Model



The MIT Information Quality Industry Symposium, 2008



## DQ Process







The MIT Information Quality Industry Symposium, 2008



## Subject based – “middle out”

- Note quality failure, notify data steward, load non-quality data, indicate non-quality attributes to consumers
- Automated information quality correction, load corrected data
- Level 1 manual intervention – hold data until corrected, then load corrected data
- Level 2 manual intervention -- hold data until corrected, then load corrected data
- For each of these activities, notify Information Product Manager



The MIT Information Quality Industry Symposium, 2008



## Subject based – “middle out”

- Provide business metadata to consumers
- Provide training in IQ to consumers
- Audit compliance with IQ standards with information generated



The MIT Information Quality Industry Symposium, 2008



## Questions?

Earl Hadden

[Earl@Hadden-Kelly.com](mailto:Earl@Hadden-Kelly.com)

(919) 593 1804

# **MATURING FROM DATA QUALITY TO INFORMATION QUALITY TO BUSINESS QUALITY:**

## ***Keys to Business Performance Excellence***

**MIT 2008 IQ Industry Symposium**

**Cambridge, MA**

**July 16-17, 2008**

by:

**Larry P. English**



**INFORMATION IMPACT  
International, Inc.**

871 Nialla Lane, Suite 100  
Brentwood, TN USA 37027  
Tel: +1 (615) 837-1211

Fax: +1 (615) 837-8804  
E-mail: [Larry.English@infoimpact.com](mailto:Larry.English@infoimpact.com)  
Web: <http://www.infoimpact.com>

0668 [0667P0, 0673DW, 0693P1]

IQ 1

© INFORMATION IMPACT Confidential & Proprietary

## **Larry P. English President and Principal**



Mr. English is an internationally recognized speaker, educator, author and consultant in information and knowledge management and information quality improvement. He also provides consulting and education in information stewardship, strategic information visioning, information technology evaluation, information resource management and data administration, data modeling and facilitation, and value-centric application development methods. Mr. English has developed the Total Quality data Management (TIQM®) methodology applying Kaizen® quality principles to information quality management. He chairs Information Quality Conferences around the world and he is a co-founder of the International Association of Information and Data Quality (IAIDQ).

Prior to founding INFORMATION IMPACT International, Inc. ([www.infoimpact.com](http://www.infoimpact.com)), Brentwood, TN, over nineteen years ago, Mr. English was Vice President of an international IRM consulting firm. Before that, he was manager of systems development and then for information management with a large publishing firm. Before positions as Senior Instructor for a computer manufacturer and Information Systems Training Coordinator for a major insurance firm, Mr. English began his career with Sears, Roebuck, and Co., as a programmer and systems analyst.

He was featured as one of the "21 Voices for the 21st Century" in the January, 2000 issue of *Quality Progress*. DAMA awarded him the 1998 "Individual Achievement Award" for his contributions to the field of information resource management. Mr. English has served as an Adjunct Associate Professor in computer science. He is a member of the American Society for Quality and is a former advisor for DAMA. He has also been an active member of various ANSI (American National Standards Institute) standards committees, and he is an editorial advisor for *DM Review*.

A magna cum laude graduate of Hardin-Simmons University, Mr. English holds a Masters Degree from the Southern Baptist Theological Seminary where he was a Luther Rice Scholar and a Garrett Fellow. He is listed in Outstanding Young Men in America and Who's Who Worldwide. He has provided consulting and educational services in more than 30 countries on five continents to such organizations as Aera Energy, Air Canada, American Express, Belgacom, Boeing, British Telecom, Coca-Cola Foods, Dow Chemical, Eastman Kodak, Eli Lilly, the FDIC, Hewlett-Packard, The Hartford, IBM, L. L. Bean, NTT DATA, Optical Fibres, Sprint, Telenor, Toyota Motor Sales, UNUM Life Insurance Co., the U.S. Navy, Western Health Alliance and Weyerhaeuser.

A frequent keynote speaker, Mr. English writes the monthly "Plain English about Information Quality" column for *DM Review*, and is the author of the highly acclaimed *Improving Data Warehouse and Business Information Quality*, also available in Japanese, and numerous articles for publications in the US and Europe.

0604

IQ 2

© INFORMATION IMPACT Confidential & Proprietary

COPYRIGHT © 1987- 2008

INFORMATION IMPACT International, Inc., Confidential & Proprietary  
871 Nialta Lane, Ste 100  
Brentwood, TN 37027  
Tel: +1-615-837-1211  
Fax: +1-615-837-8804  
Email: [info@infoimpact.com](mailto:info@infoimpact.com)  
Web Site: <http://www.infoimpact.com>

This material is the sole property of INFORMATION IMPACT International, Inc., World rights reserved. This document is based on trade secrets or copyrighted material owned by Information Impact International, Inc. No part of this document may be stored in a retrieval system, transmitted or reproduced in any way, including but not limited to photocopy, photograph, magnetic or other record, without the prior agreement and written permission of INFORMATION IMPACT International, Inc.



TIQM® and TQdM® are registered trademarks of INFORMATION IMPACT International, Inc.

Is a registered trademark of INFORMATION IMPACT International, Inc.

IQMM™ is a trademark of INFORMATION IMPACT International, Inc.

RADD™ is a trademark of INFORMATION IMPACT International, Inc.

0603 [0602]

IQ 3

© INFORMATION IMPACT Confidential & Proprietary

## MATURING FROM DATA QUALITY TO INFORMATION QUALITY TO BUSINESS QUALITY

- ❑ The Stages of IQ Management Maturity
- ❑ Taking Inventory: Where are You?
- ❑ Establishing a Vision: “Begin with the End in Mind”
- ❑ Planning your Next Steps: “Put First Things First”
- ❑ Controlling Processes to “Hold the Gain”
- ❑ Moving to Certainty: Measuring the Value Delivered

05264

IQ 4

© INFORMATION IMPACT Confidential & Proprietary

**COMMON MISCONCEPTIONS**

1. IQ is “data cleansing”
2. IQ is data assessment
3. IQ is “fitness of purpose”
4. Quality is *best-of-breed* or *zero defects*
5. IQ *problems* are created by the information producers
6. IQ improvement is what the Information Quality Team does
7. IQ problems can be edited out
8. TQM or TIQM® is a program / project
9. IQ is quality of data in databases
10. IQ is too expensive

IQ = Information Quality  
TQM = Total Quality Mgt  
TIQM® = Total Information Quality Mgt

0802 [5027, 5408] IQ 5 © INFORMATION IMPACT Confidential & Proprietary

**TOTAL INFORMATION QUALITY MANAGEMENT**

- ❑ Information Quality is **NOT\*** about what is in databases  
(\*well, it is, but that is not all)
- ❑ Information Quality (IQ) is **ABOUT business, service and manufacturing performance excellence by improving processes to** increase information quality

Information Quality addresses:

- Quality of information *definition, models, DB designs*
- Quality of information *content*
- Quality of information *presentation*
- Quality of *business communication*

➡ **Total** Information Quality Management results in:

- Increased *Customer* satisfaction
- Increased *Employee* satisfaction and *productivity*
- *Decreased* costs and *increased* profits / surplus

5027 [0802, 5408] IQ 6 © INFORMATION IMPACT Confidential & Proprietary

## THE DISCIPLINE OF INFORMATION QUALITY MANAGEMENT

The application of *proven Quality Management principles, processes and practices* to information as a *product* of the enterprise processes (business, manufacturing & service) to meet or exceed information consumers' expectations

*Larry P. English*

5233Q [5240MQ, 5236M, 5237KM]

IQ 7

© INFORMATION IMPACT Confidential & Proprietary



## INFORMATION QUALITY

“Consistently  
meeting<sup>\*</sup>

*all* knowledge workers' and end-customers'  
expectations”

through information and information services so:

- *Knowledge workers* accomplish enterprise objectives
- *Customers* are successful

Larry P. English, TIQM®

### ↖ *Components* of Information Quality:

- Information Product Specifications and Information Architecture (Definition & Rules)
- Data Content
- Information Presentation

<sup>\*</sup>World-class organizations do not stop here—  
they strive to “delight” their customers

3851 [3891Gov]

IQ 8

© INFORMATION IMPACT Confidential & Proprietary

**P2.2** → **THE FUNDAMENTAL QUALITY PRINCIPLES**

- **Customer Focus**
  - Market focus
  - Customer satisfaction
  - Supplier / Customer Partnership
- **Process Improvement** to reduce waste
  - Process definition
  - Product specification (customer-focused)
  - Team work
  - Continuous Process Improvement (CPI)
  - Business Process Re-engineering (BPR)
- **Proven, scientific Methods**
  - Statistical quality control
  - PDSA or PDCA (Shewhart cycle)
- **Management Accountability**

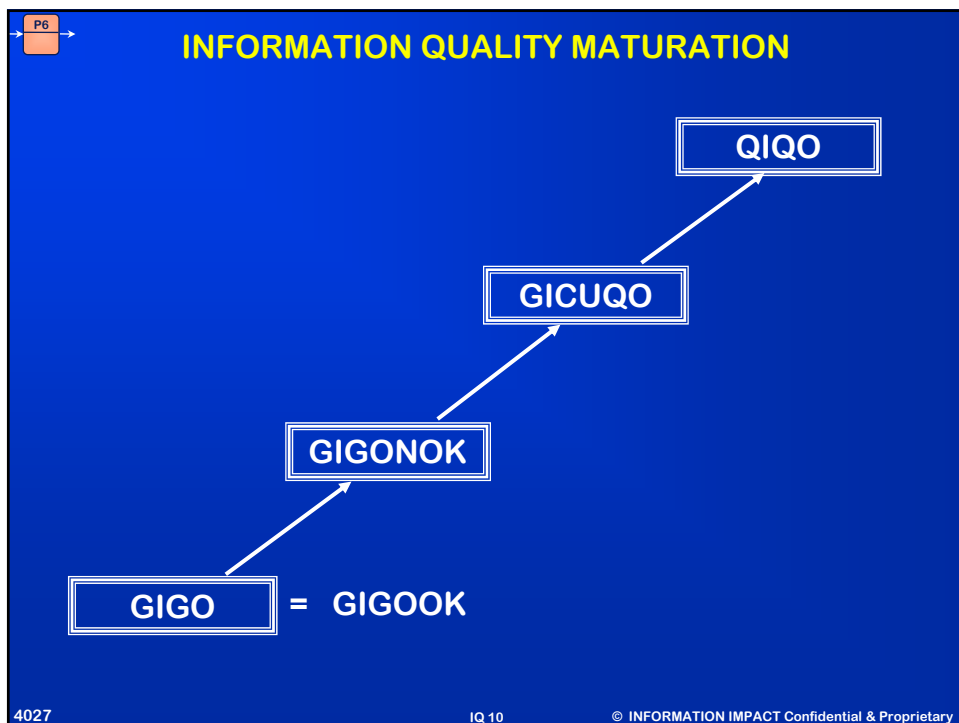


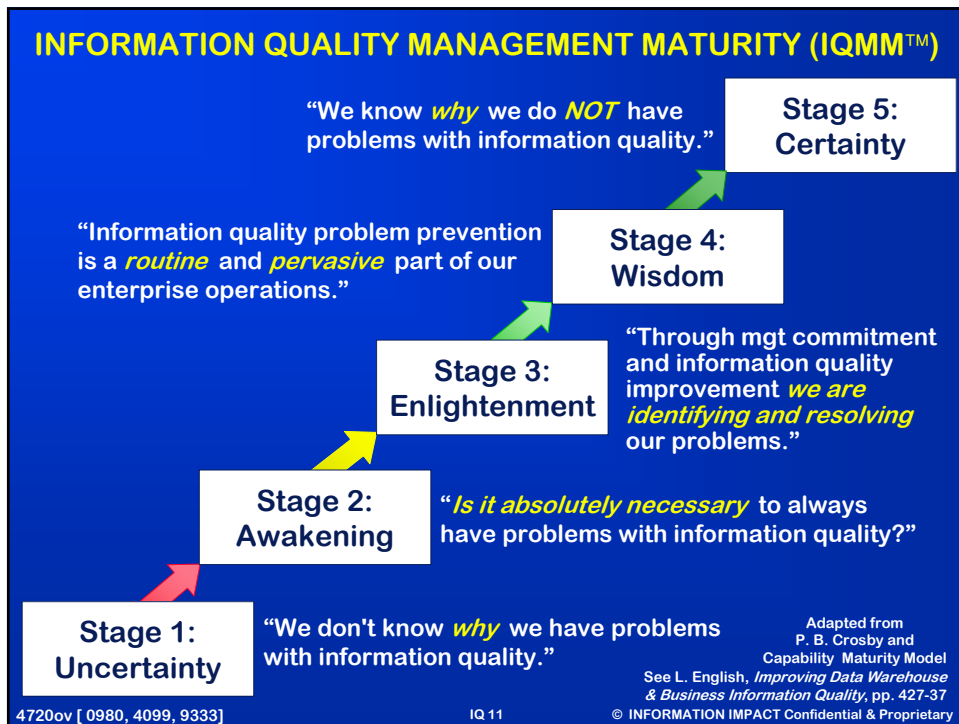




CPI = Continuous Process Improvement  
BPR = Business Process Reengineering

0807 [0870, 0896, 8818]      IQ 9      © INFORMATION IMPACT Confidential & Proprietary





P6 INFORMATION QUALITY MANAGEMENT MATURITY GRID					
Measurement Categories	Stage 1: Uncertainty (Ad hoc)	Stage 2: Awakening (Repeatable)	Stage 3: Enlightenment (Defined)	Stage 4: Wisdom (Managed)	Stage 5: Certainty (Optimizing)
1. Management understanding and attitude	No comprehension of information quality as a management tool. Tend to blame data administration or I/S org for "information quality problems" or vice versa.	Recognizing that information quality management may be of value but not willing to provide money or time to make it all happen.	While going through information quality improvement program learn more about quality management; becoming supportive and helpful.	Participating. Understand absolutes of information quality management. Recognize their personal role in continuing emphasis.	Consider information quality management an essential part of company system.
2. Information quality organization status	"Data" quality is hidden in application development departments. Data audits probably not part of organization. Emphasis on correcting bad data.	A stronger information quality role is "appointed" but main emphasis is still on correcting bad data.	Information quality organization exists, all assessment is incorporated and manager has role in development of applications.	Information quality manager reports to CIO; effective status reporting and preventive action. Involved with business areas.	Information quality manager is part of management team. Prevention is main focus. Information quality is a thought leader.
3. Information quality problem handling	Problems are fought as they occur; no resolution; inadequate definition; lots of yelling and accusations.	Teams are set up to attack major problems. Long-range solutions are not solicited.	Corrective action communication established. Problems are faced openly and resolved in orderly way.	Problems are identified early in their development. All functions are open to suggestion and improvement.	Except in the most unusual cases, information quality problems are prevented.
4. Cost of information quality as percent of revenue	Reported: unknown Actual: 20%	Reported: 5% Actual: 18%	Reported: 10% Actual: 15%	Reported: 8% Actual: 10%	Reported: 5% Actual: 5%
5. Information quality improvement actions	No organized activities. No understanding of such activities.	Trying obvious "motivational" short-range efforts.	Implementation of the 14 point program with thorough understanding and establishment of each step.	Continuing the 14 point program and starting to optimize.	Information quality improvement is a normal and continued activity.
Summation of company information quality posture	"We don't know why we have problems with information quality."	"Is it absolutely necessary to always have problems with information quality?"	"Through management commitment and information quality improvement we are identifying and resolving our problems."	"Information quality problem prevention is a routine part of our operation."	"We know why we do not have problems with information quality."

Adapted from P. B. Crosby & Capability Maturity Model

IQMM™ is a trademark of Information Impact Int'l  
0980ov [4719-20, 0811, 0921, 4057] L. English, *Improving Data Warehouse and Business Information Quality*, pg. 428  
IQ 12 © INFORMATION IMPACT Confidential & Proprietary



P6  
Pt 7

## IQ 7. INSTITUTE LEADERSHIP FOR INFORMATION QUALITY

- ❑ Management is *Leadership*—not “supervision”
  - Leaders enable workers to improve their processes
  - Most supervisors are just the opposite, because they implement inappropriate measures and rewards
- ❑ Information Quality ramifications:
  - Take the *lead* in information quality improvement
  - Educate and *coach* executives
  - Implement management *accountability*
  - Learn how your customers use information
  - Measure and reward the right things:
    - ↓ Teamwork, customer satisfaction, waste reduction, total cost of ownership

Source: L. English, *Improving Data Warehouse and Business Information Quality*, p 367+

0865 [4832-45, 4714-16]
IQ 13
© INFORMATION IMPACT Confidential & Proprietary


P6  
Pt 14

## IQ 14. TAKE ACTION TO ACCOMPLISH THE TRANSFORMATION FOR INFORMATION QUALITY

- ❑ Management must put everyone to work to transform org.
  - Must organize itself to administer the other 13 points
  - Senior management must feel the pain of status quo
  - Senior management must communicate to a critical mass of people why change is necessary for all
  - Every activity is a process that can be improved
- ❑ Use the Shewhart Cycle
 

4. Roll the process out and study the results— what did we learn?

3. Observe the effects of the “improvement”



1. Study a defective process to identify root cause(s) and define improvement(s)

2. Implement the improvement in a controlled way

Source: L. English, *Improving Data Warehouse and Business Information Quality*, p 350+

0879 [5144, 4832-45, 4714-16, 4979-93, 0899, 5562]
IQ 14
© INFORMATION IMPACT Confidential & Proprietary

P6 →

## ENTERPRISE EXCELLENCE VISION

Customer-Centered,  
Shared Vision + Capable, Trained,  
Empowered People +

Defined, Improved,  
Controlled Processes  
that delivers =  
Quality Just-In-  
Time Information

**\$UCCESS !!!**

4767 [2089, 89211] IQ 15 © INFORMATION IMPACT Confidential & Proprietary

P6 →

## INFORMATION QUALITY MANAGEMENT

### Mission / Vision

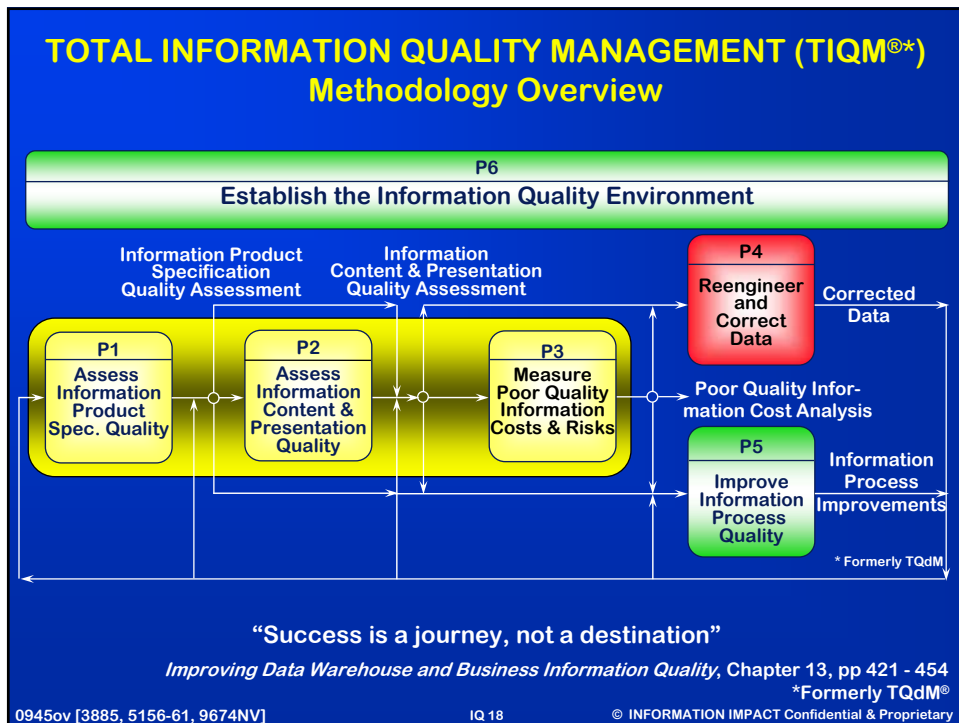
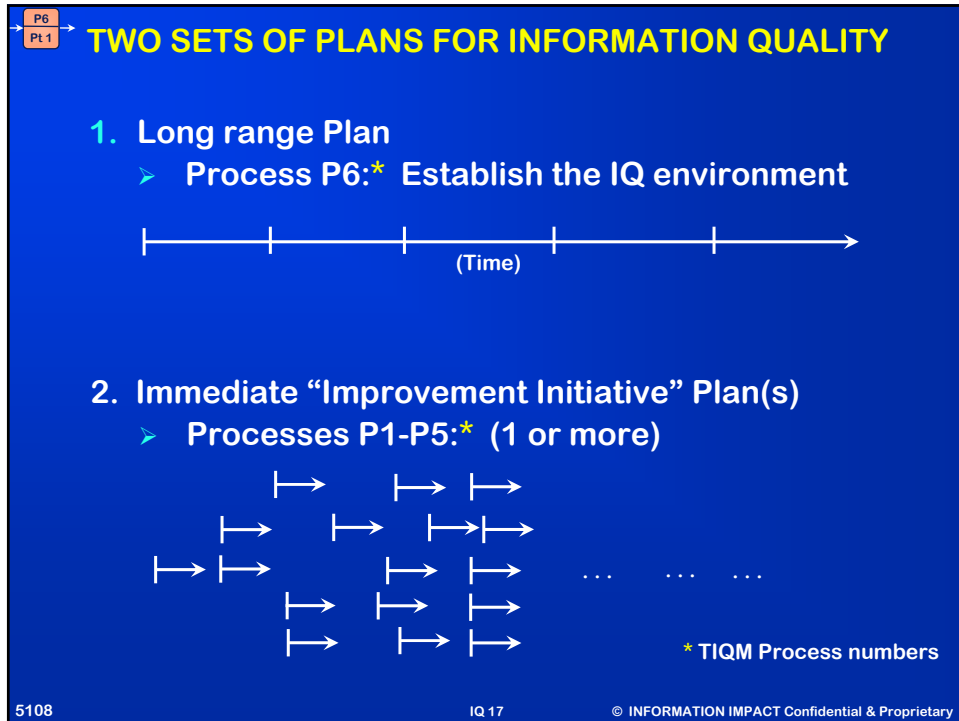
By implementing and performing sound and proven quality management principles and processes to our information processes, we enable the accomplishment of:

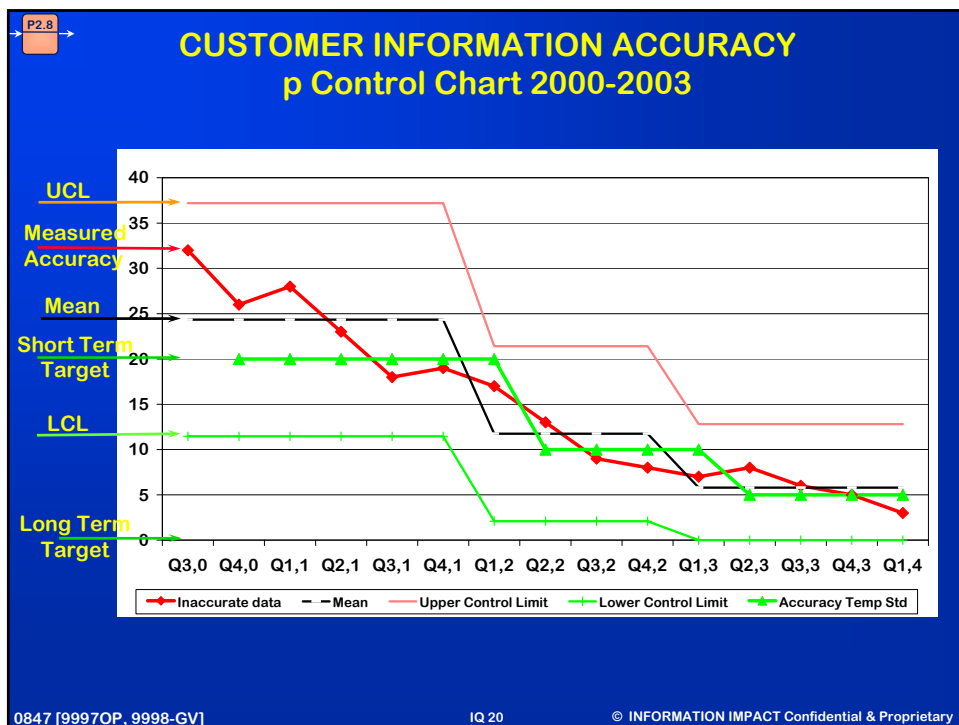
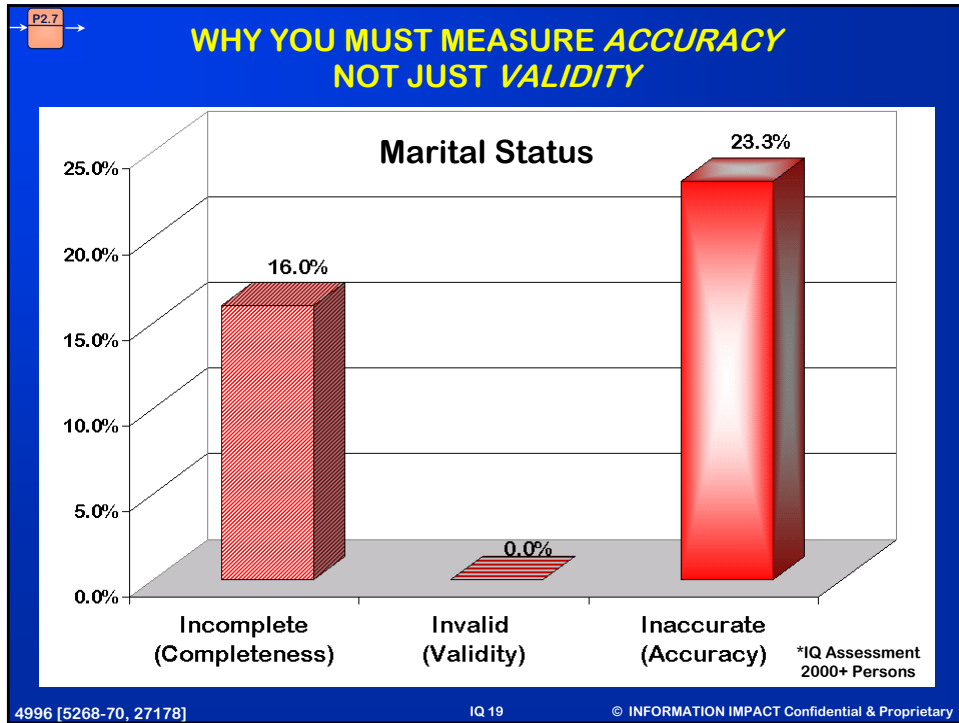
[ **ENTERPRISE MISSION HERE** ]

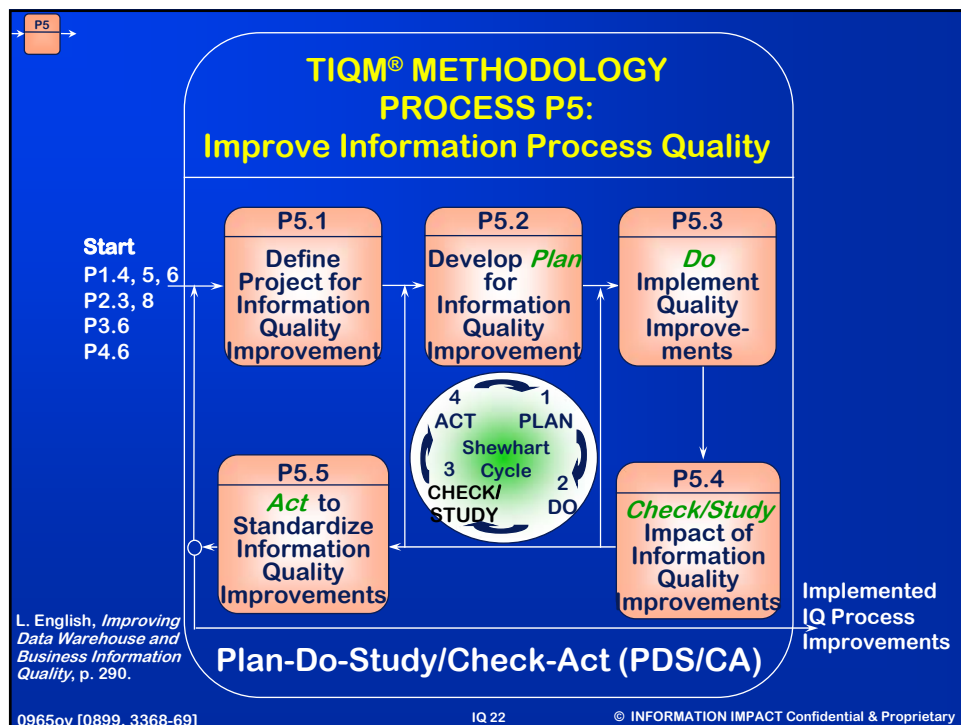
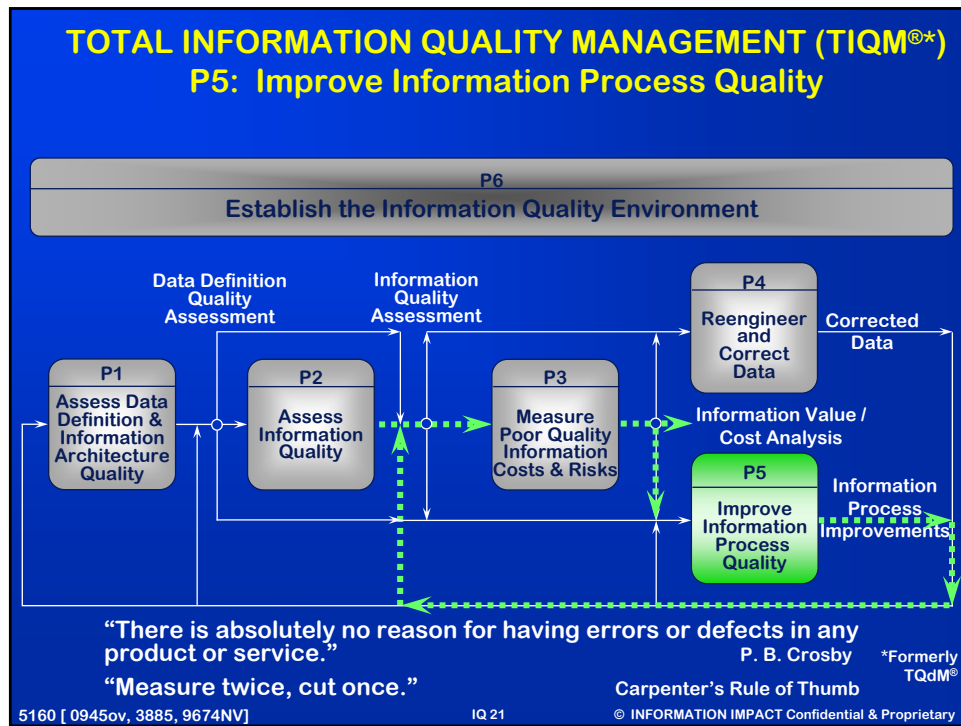
We do this by [e.g., “ ‘Increasing customer satisfaction’ by preventing errors in customer information, such as name misspelling, invoicing, sending wrong items.” or,

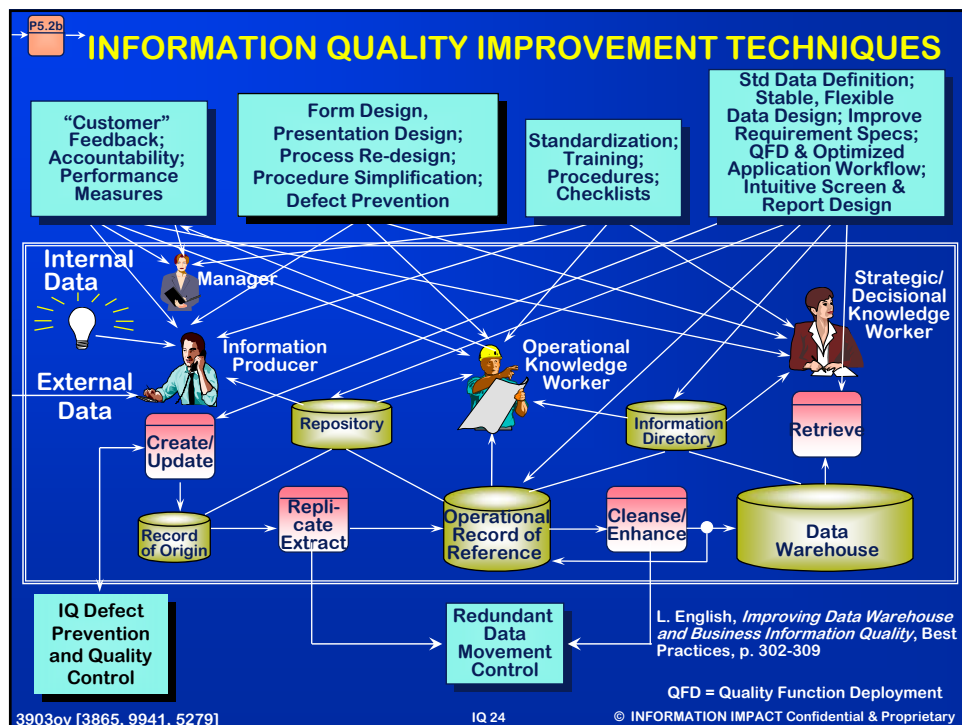
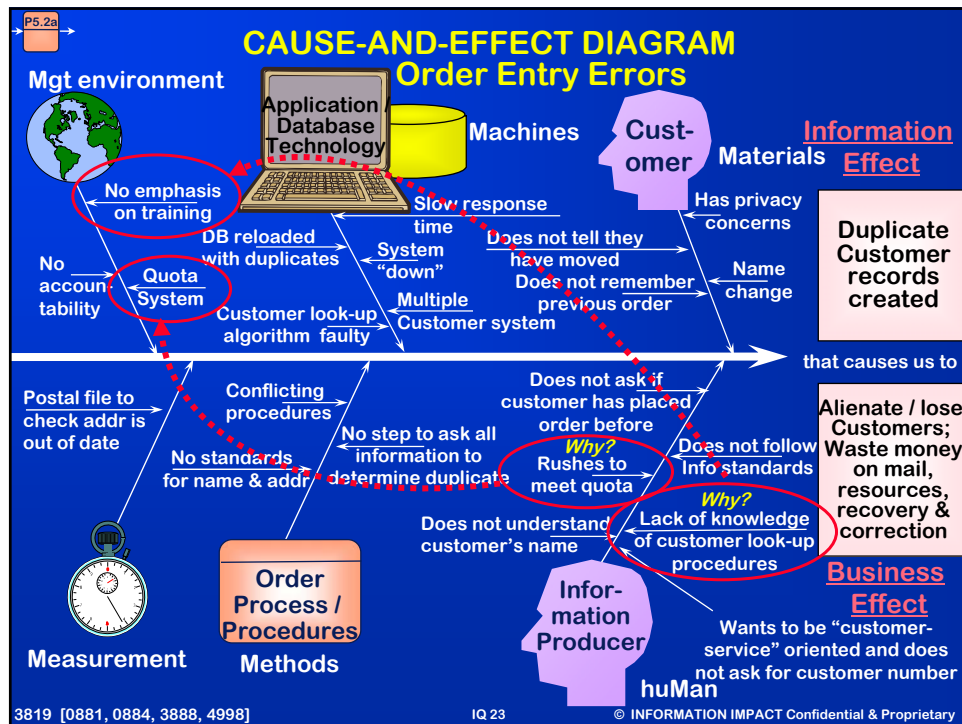
“ ‘Decreasing operating costs’ by decreasing costs of process failure, recovery and information ‘scrap and rework’ caused by poor quality information.”]

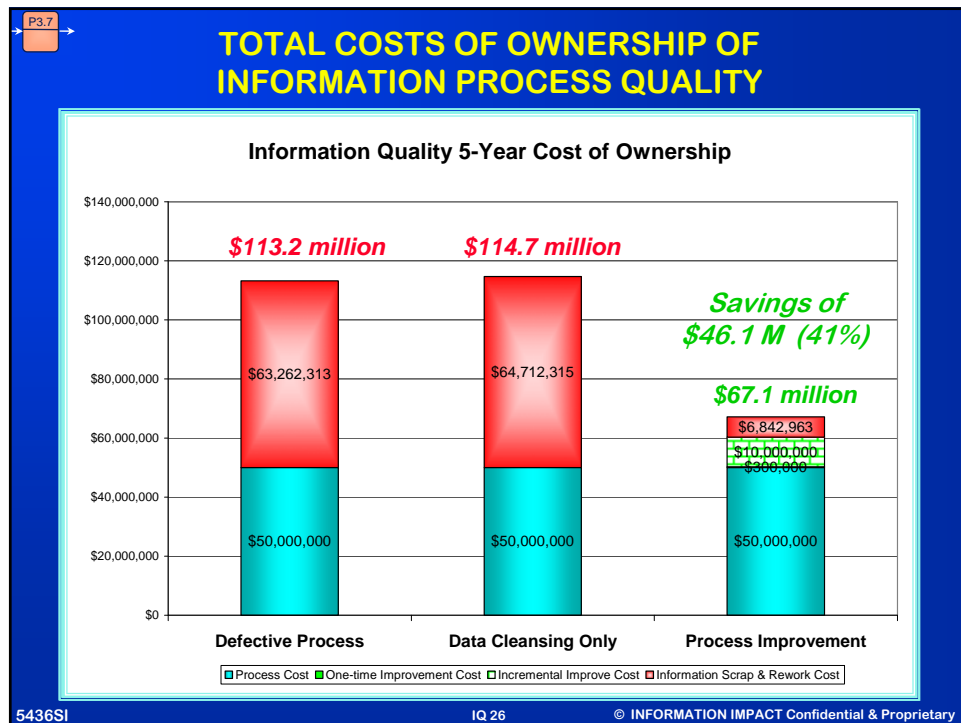
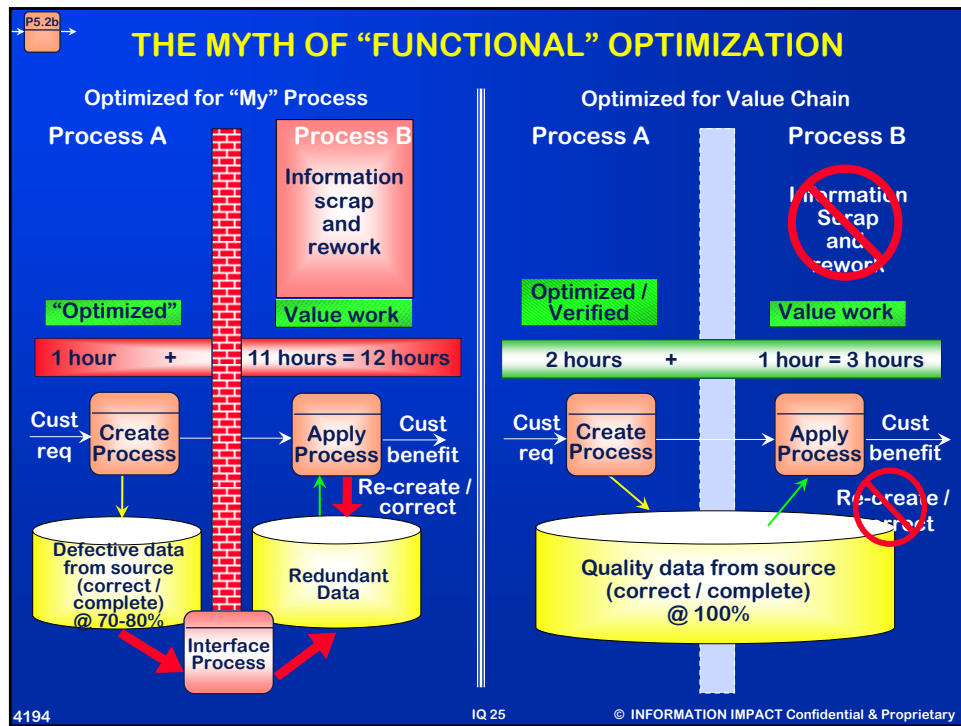
5020 IQ 16 © INFORMATION IMPACT Confidential & Proprietary

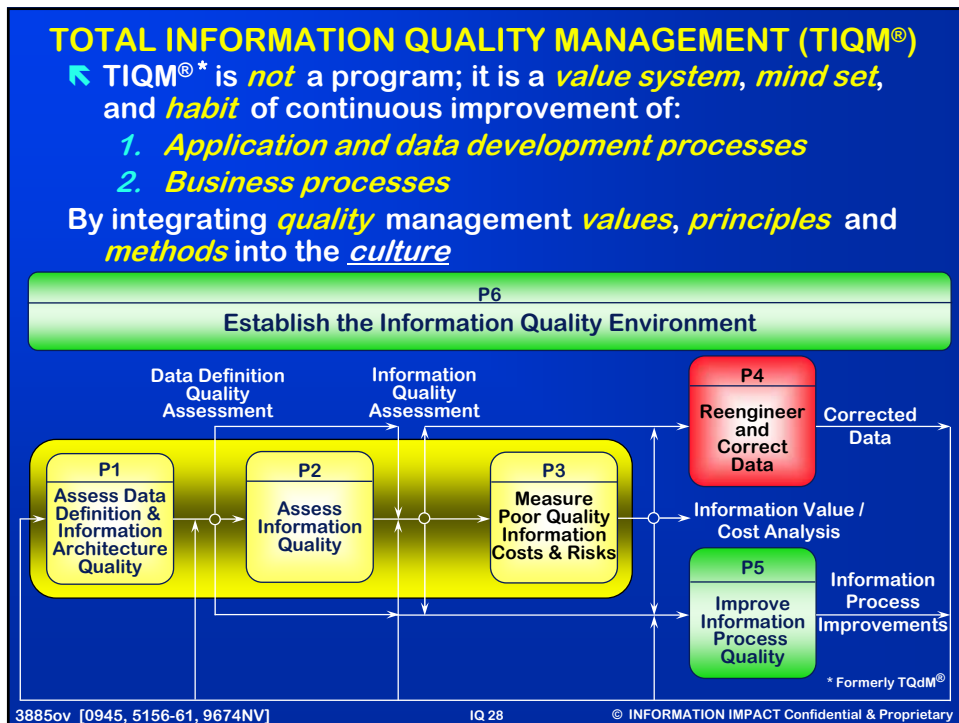
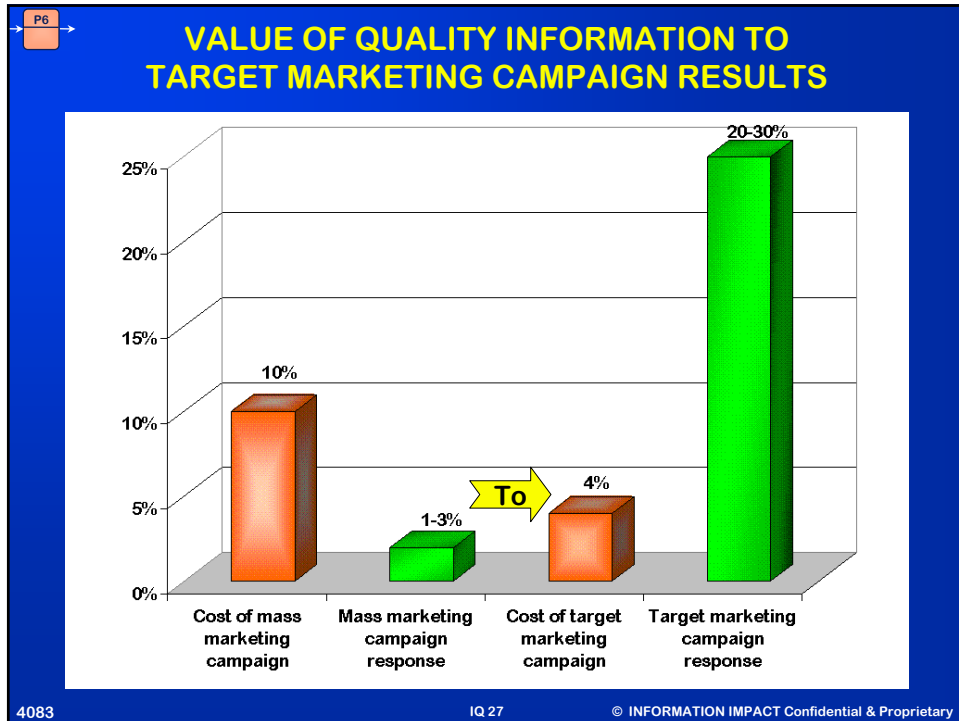














P6

## TOTAL QUALITY MANAGEMENT

### Deming's 14 Points

1. Create constancy of purpose toward improvement of product and service, with the aim to become competitive and to stay in business, and to provide jobs.
2. Adopt the new philosophy. We are in a new economic age. Western management must awaken to the challenge, must learn their responsibilities, and take on leadership for change.
3. Cease dependence on mass inspection to achieve quality. Eliminate the need for inspection on a mass basis by building quality into the product in the first place.
4. End the practice of awarding business on the basis of price tag. Instead, minimize total cost. Move toward a single supplier for any one item, on a long-term relationship of loyalty and trust.
5. Improve constantly and forever the system of production and service, to improve quality and productivity, and thus constantly decrease costs.
6. Institute training on the job.
7. Institute leadership. The aim of supervision should be to help people and machines and gadgets to do a better job. Supervision of management is in need of overhaul, as well as supervision of... workers.

Source: Deming, *Out of the Crisis*  
Larry English, *Improving Data Warehouse and Business Information Quality*, p338

0812 [0819, 0820, 4092-98]
IQ 29
© INFORMATION IMPACT Confidential & Proprietary

P6

## TOTAL QUALITY MANAGEMENT

### Deming's 14 Points (Cont.)

8. Drive out fear, so everyone may work effectively for the company
9. Break down barriers between departments. People in research, design, sales, and production must work as a team, to foresee problems of production and in use that may be encountered with the product or service.
10. Eliminate slogans, exhortations, and targets for the work force asking for zero defects and new levels of productivity. Such exhortations only create adversarial relationships, as the bulk of the causes of low quality and low productivity belong to the system and thus lie beyond the power of the work force.
11. a. Eliminate work standards (quotas) on the factory floor. Substitute leadership.  
b. Eliminate management by objective. Eliminate management by numbers, numerical goals. numerical goals. Substitute leadership.
12. a. Remove barriers that rob the hourly worker of his right to pride of workmanship. The responsibility of supervisors must be changed from sheer numbers to quality.  
b. Remove barriers that rob people in management and in engineering of their right to pride of workmanship. This means, *inter alia*, abolishment of the annual or merit rating and of management by objective.
13. Institute a vigorous program of education and self-improvement.
14. Put everyone to work to accomplish the transformation. The transformation is everybody's job. Management will explain by seminars and other means why change is necessary, and that the change will involve everybody. Deming, *Out of the Crisis*  
Larry English, *Improving Data Warehouse and Business Information Quality*, p338

0813 [0819, 0820, 4092-98]
IQ 30
© INFORMATION IMPACT Confidential & Proprietary

### TOTAL INFORMATION QUALITY MANAGEMENT: 14 Points

1. Create constancy of purpose for improvement of *information* product and service: Long term plan; the obligation to the *knowledge worker* never ceases
2. Adopt the new philosophy of quality *shared information* as a tool for business improvement: “Reliable (*quality*) shared information reduces costs”
  - Means *transformation* of I / S & business management
3. Cease reliance on data and application *inspections alone* to achieve information quality: *Design quality in* to the information design and production processes
4. End the practice of developing applications on the basis of “on-time,” “within budget” measures alone and capturing data at the lowest cost: *Develop single data creation programs and trust* in information producers\*

\*Note: Contract with your information suppliers

\* Adapted from Deming's 14 Points, See L. English, *Improving Data Warehouse & Business Information Quality*, ch 11

4714 [4092-93, 0854, 0858-9, 0861, 4832-45, 9675NV]

IQ 31

© INFORMATION IMPACT Confidential & Proprietary

### TOTAL INFORMATION QUALITY MANAGEMENT: 14 Points

5. *Improve constantly and forever the processes* of application and information development and service and of information production, through a *habit* of continuous “information defect prevention”
6. *Institute training* on information quality for all employees, especially management and producers
7. Institute leadership for information quality: appoint a full-time information quality leader; *management must assume accountability for* information quality
8. Drive out fear of data uncertainty or data correction: Implement incentive programs for finding / and correcting problem *causes; do not blame or punish*
9. *Break down barriers* between business areas: information management and application development; IT and business; business area and business area units

\* Adapted from Deming's 14 Points, See L. English, *Improving Data Warehouse & Business Information Quality*, ch 11

4715 [4093-94, 0863-67, 4832-45, 9676NV]

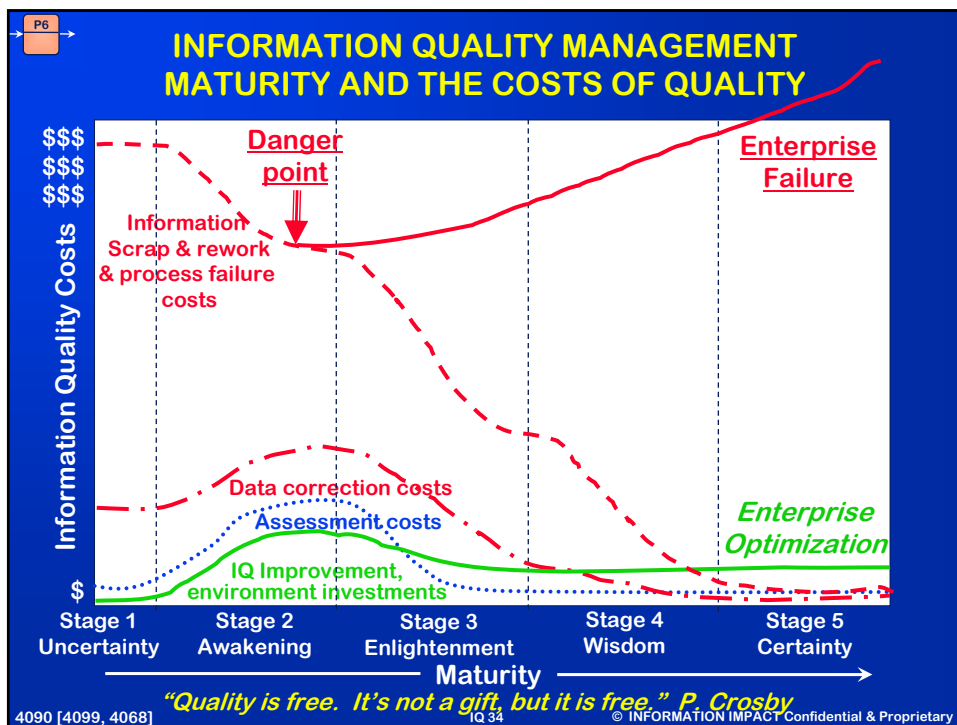
IQ 32

© INFORMATION IMPACT Confidential & Proprietary

## TOTAL INFORMATION QUALITY MANAGEMENT: 14 Points

10. Eliminate slogans and exhortations [only]; *replace with actions for information quality improvement*: Implement a Plan-Do-Check-Act process for information quality improvement
11. Eliminate quotas of “productivity” that increase errors and costs of scrap and rework: *Customer satisfaction*
12. Remove barriers to pride of workmanship; *empower information producers* to fix the broken processes
13. Institute a vigorous program of education and *self-improvement* for all people: understand the paradigm shift and learn tomorrow’s skills
14. Take action to accomplish the transformation for IQ: *Senior management must* feel the pain of the status quo, organize itself and communicate to a critical mass
  - Every process is a candidate for improvement

4716 [4095-96, 0872-73, 0877-79, 4832-45, 9677NV] \* Adapted from Deming's 14 Points, See L. English, Improving Data Warehouse & Business Information Quality, ch 11  
IQ 33 © INFORMATION IMPACT Confidential & Proprietary



27124

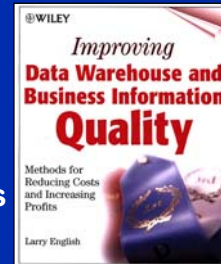
Thank you for your valuable time. Please share your feedback and comments as you apply your new knowledge (Larry.English@infoimpact.com)

*Larry P. English*

**www.infoimpact.com**

Your *Information Portal* for information quality and information management:

- See or share *IQ Best Practices*
- Review and link to *IQ Products*
- Links to *Other IQ Resources* & IQ web sites
- Recommended reading in the *Information Professional's Reference Library*
- And other information



ISBN: 0-471-25383-9  
John Wiley & Sons, 1999

Preview & see reviews at  
[www.infoimpact.com](http://www.infoimpact.com)

0688ov [ 0689-91, 27124]

IQ 35

© INFORMATION IMPACT Confidential & Proprietary

# Data Driven: Profiting from Your Most Important Business Asset



Thomas C. Redman, Ph.D.  
Navesink Consulting Group

At MIT Conference

Cambridge, MA

July 16, 2008

[www.dataqualitysolutions.com](http://www.dataqualitysolutions.com)

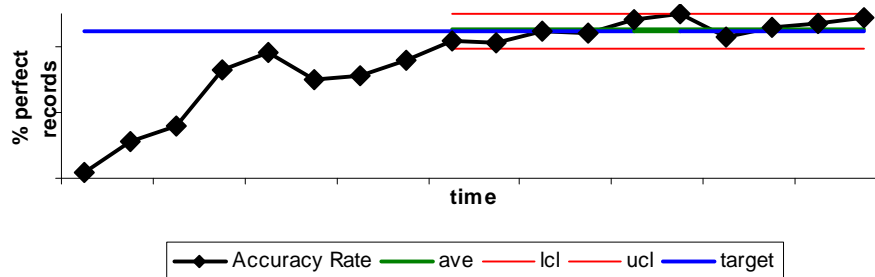
Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T.C. Redman, Page 1

## Those who apply diligent efforts (almost) always improve data quality. And benefit!

First-Time, On-Time Performance  
(actual results)



**Each error not made saves an average of \$500.  
This amounts to millions quickly!**

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 2

## Paradox, part 1

---

How do we reconcile the evident successes with the observation that data quality is so poor at so many companies?

Hypothesized Answer:

- They (usually) don't recognize data as assets (after all "out-of-sight, out-of-mind) and so worth the effort.

## Consultant's exercise: Fire!

---

You can save only one of the following:

- Antique French Desk.
- Brand new PC, with all the bells and whistles.
- Only copy of the organization's fifty biggest accounts.

The Data Doc's Response: Finally!

## The paradox, part 2

---

How do we reconcile the fact that everyone intuitively knows that data are critical assets with the fact that organizations don't manage them as such?

Hypothesized Answer(s):

- They don't understand what "manage data assets" really means.
- Specifically, they don't see how to make money with data.

## What Does "Manage Data Assets" Really Mean?

---

Generally recognized as business assets:

- Capital, in its various forms
- People, including the knowledge in their heads.

## Organizations naturally manage their assets...

- They take care of them.
- They put them to work, to make money.
- They adjust their management systems to account for the special properties of each asset.

## For data, “taking care” is mostly about quality

**Prescription 1:** Take steps to ensure that

- Possess and acquire the right kinds of data.
- People can access and understand them.
- People can trust that they are “good enough.”
- They are of high enough quality to withstand market scrutiny.
- They are kept safe from loss or theft.

It is highly significant that (almost) all organizations that diligently follow many of “the ten habits” make order-of-magnitude improvements.



## Putting data to work

**Prescription 2:** Use data to create new revenue

- Sell them directly in the market.
- Build them into other products and services.
- Use them to enhance other products/services.
- Use them to make better decisions.
- Use them to improve the day-in, day-out running of the business.

Critical point: Management must explicitly think through how they will put data to work in creating new value.

## Adjusting the management system

**Prescription 3:** Recognize that data have unique properties

- Example: Unlike other assets, data can be shared
- Most important: Data are the only asset that are uniquely an organization's own. The "ultimate proprietary technology."

**Prescription 3, cont:** Adjust the organizational structures, roles, and responsibilities as a result.

- Counterexample: Chief Information *Technology* Officer

## Outline:

---

- What does “manage data assets” mean?
- A bit of flavor for:
  - Putting data to work
  - The wondrous and perilous properties of data as an asset
  - Implications for the management system
  - The brutal (and growing) politics associated with data
- A new context for data quality
- The ten habits of those with the best data.

## Putting Data to Work

---

- The many ways to bring data to market
- A note on decision-making
- Your most important data

## A Note on Market Demands

- People and organizations have always wanted “more and better” data.
- Historically, the elite took steps to hoard data.
- Since the rise of democracy, some of their grip has been broken.
- Sheer demand continues to grow and is in little doubt:

“Inside IBM, we talk about 10 times more connected people, 100 times more network speed, 1,000 times more devices, and a million times more data.”\*

\*Lou Gerstner, quoted in McDougall, P., “More Work Ahead,” *Information Week*, December 18-25, 2000, p. 22.

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 13

## A Note on Market Demands-2

- To borrow from Twain,  
“the difference between the right data and the almost-right data is like the difference between lightning and a lightning bug.”
- People and organizations expect:  
“exactly the right data and information in exactly the right place at exactly the right time and in the right format to make a decision, complete an operation, or serve a customer.”

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 14

## So far, I've identified fifteen ways to fulfill these demands

### Provide Content

- ☐ New Content
- ☐ Re-package
- ☐ Informationalization
- ☐ Unbundling
- ☐ Exploiting Asymmetries
- ☐ Closing Asymmetries

### Facilitators

- ☐ Own the Identifiers
- ☐ Infomediation
- ☐ Data mining/Analytics
- ☐ Privacy and security
- ☐ Training
- ☐ New Marketplaces
- ☐ Infrastructure technologies
- ☐ Information appliances
- ☐ Tools

## Content Providers

**Basic Idea:** Provide newer, richer, better, etc. data to address customer needs

**Customer Need Fulfilled:** "Which diet will work best for me?/What can I realistically expect to achieve?"

**Industrial Age Examples:** Generic diet guides, Newspaper

**Information Age Examples:** Personal diet, Morningstar, Information Resources Inc.

## Informationalization

**Basic Idea:** Enhance existing products and services by building data and information in.

**Customer Needs Fulfilled:** Simplicity, integration, etc.

**Industrial Age example:** Product instructions

**Information Age examples:**

- Auto makers are now including GPS Navigation systems.
- NC State is re-designing the hospital gown, building a thermometer and other sensors in.

## Exploit Information Asymmetries

**Basic Idea:** Know more than the guy on the other side of the transaction.

**Customer Need Fulfilled:** Get the best possible “deal.”

**Industrial Age example:** Used car salesman

**Information Age examples:** Hedge Funds

## Infomediation

**Basic Idea:** Help people find the data and information they need

**Customer Need Fulfilled:** Waste less time.

**Industrial Age example:** Travel agents

**Information Age example:** Google

## Data Mining/Analytics

**Basic Idea:** Uncover hidden “nuggets” buried in the data.

**Customer Need Fulfilled (examples):**

- Deep insights into individual needs
- Exploit patterns of excellence/opportunities of improvement.

**Industrial Age example:** Statistical analysis

**Information Age example:** Harrah's, Amazon

## Other Ways to Bring Data and Information to Market - Content

- Repackage and filter to meet specific needs
  - FTID,
- Unbundle
  - Securities Research and trade processing
- Close Information Asymmetries
  - Consumer Reports

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 21

## Other Ways to Bring Data and Information to Market - Facilitators

- Own the identifiers
  - Standard & Poors
- Privacy and Security
  - Legal profession
- Define and Operate “data markets”
  - E-Bay
  - Facebook
- Training and Education
  - Internet-based training

Redman-MIT08-Data Driven

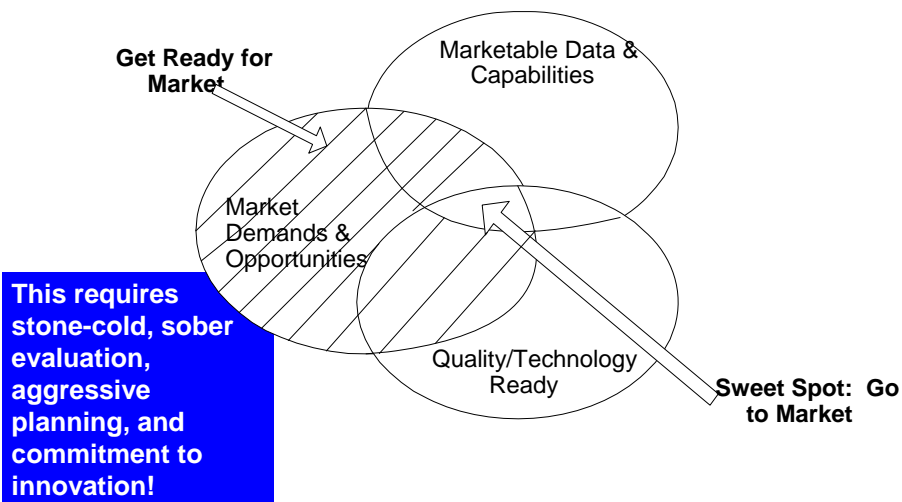
© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 22

## Make Better Decisions

Another good way to put data to work is systematically use them to make better decisions, align the organization to the tasks at hand, and execute.

## Implication: Organizations need to find and pursue their “data sweet spots”





## Implications, cont:

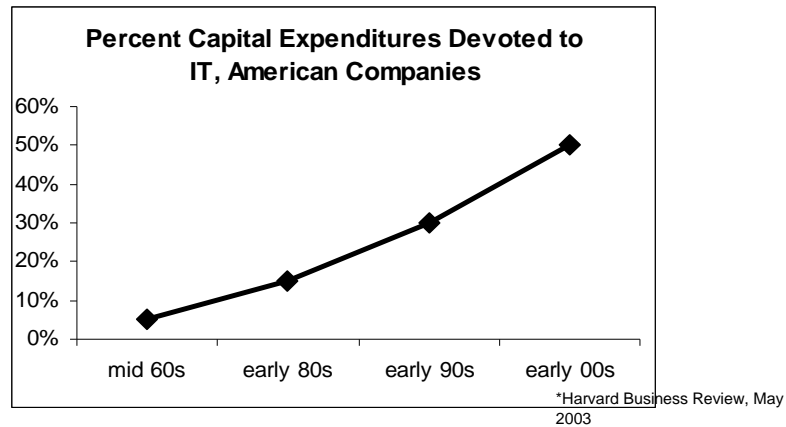
- ❑ Data Doc claim: “The organization’s most important data are those that help it make money.”
- ❑ Those used to create new revenue are especially important.
- ❑ Note that every organization exposes some data in its marketplaces.
- ❑ We data geeks should focus on these business opportunities and the required data.
- ❑ We should measure success by metrics like “new revenue from data.”
- ❑ Note: It is a lot easier to invest in revenue growth than cost reduction. Improved quality is a perfect example.

## The Wondrous, Perilous and Often Confounding Properties of Data In Organizations

- ❑ Most important: The ultimate proprietary technology.
- ❑ Data are “organic.”
- ❑ Note: About ten such properties really matter.

## “IT Doesn’t Matter,” Nicholas Carr\*

Information Technologies have penetrated every aspect of modern life.



Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 27

## Proprietary vs. Infrastructure Technologies

PROPRIETARY	INFRASTRUCTURE
Can be “owned” by a single organization	(Eventually) part of general business infrastructure
Patented drug, unique process	Railroads, electric grid
Protected	Become commoditized
Basis for sustained advantage	Not a basis for sustained advantage

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 28

## Advantage Stems from Scarcity...

- Carr argues that basic storage, processing, and transport technologies are now readily available to all.
- Carr does not argue that IT isn't important. Only that it is not strategic.
- He offers the following advice:
  - Spend less.
  - Follow, don't lead.
  - Focus on vulnerabilities, not opportunities.

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 29

## Finding Reasons to Attack Carr is Easy

- No proprietary technology/advantage lasts forever... or even very long.
- The pace of innovation in IT is only growing.
- Advantage can still be sustained by using IT in smarter ways.

**But many organizations seem to be following his advice!**

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 30

## Data are the Organization's Ultimate Proprietary Technology!

- No other organization has, or can have, the same data.
- Data are subtle and nuanced.
  - Model "customer" in unique ways that best suit it.
  - Capture and utilize unique "facts."
  - Processes to capture unique data are also difficult to copy.
- Eventually, of course, some data become standardized to facilitate communications.
- Data offer opportunity for sustained advantage—and everyone knows it!

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 31

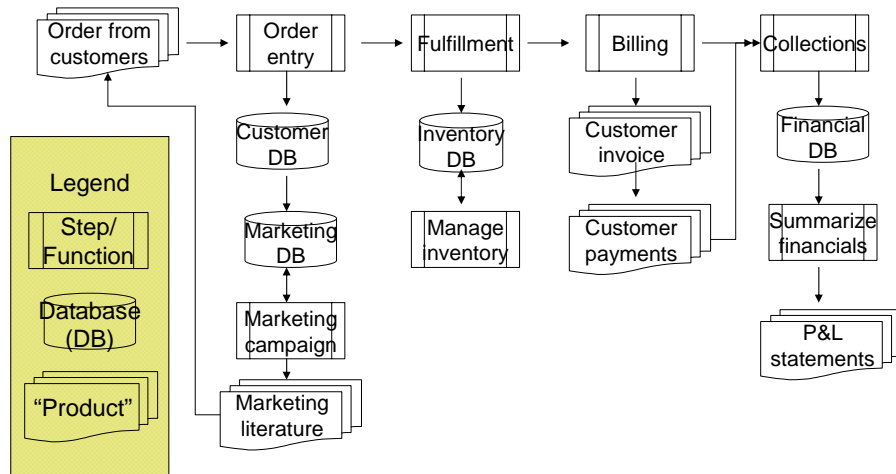
**Data are subtle  
and nuanced and  
have become the  
organization's  
*lingua franca***



Never underestimate the importance of local knowledge.

HSBC   
The world's local bank

## Data are “organic” and most useful “on the fly” (when they cross departmental boundaries)



Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 33

## Implications

- ❑ Must not confuse management of *technology* with management of *data*.
- ❑ Must be very careful about what data we standardize. Standard data has little marketplace value.
- ❑ Should strive for greater uniqueness, novelty, and depth in data put in the marketplace.
- ❑ Need to identify and explicitly manage the most important, end-to-end value-creating flows of data as “information chains” or Big-P processes.
- ❑ Need to improve quality, in its own right, but more especially to meet market demands.

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 34

## The Surprisingly Brutal Politics Associated with Data

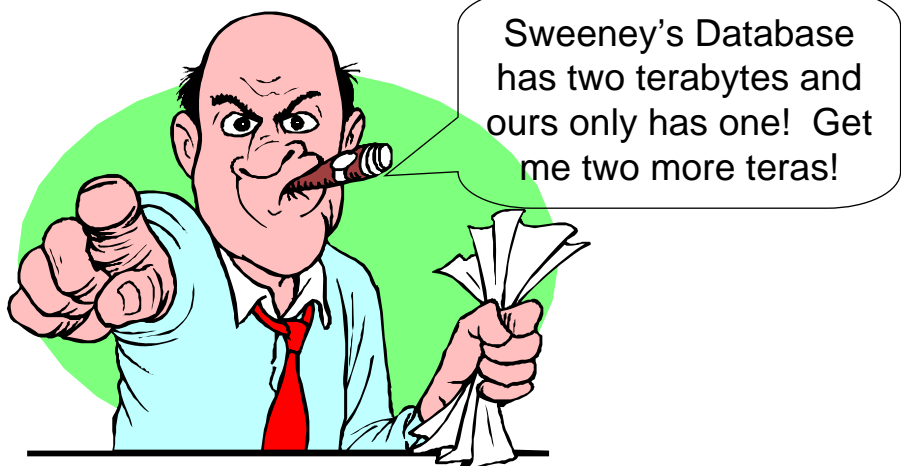
- Data Sharing
- Responsibility for quality
- About a dozen important, as they play out locally.

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T.C. Redman, Page 35

### 1. Power/Data Sharing/Ownership: In the Information Age, Possession of Data Conveys Power!

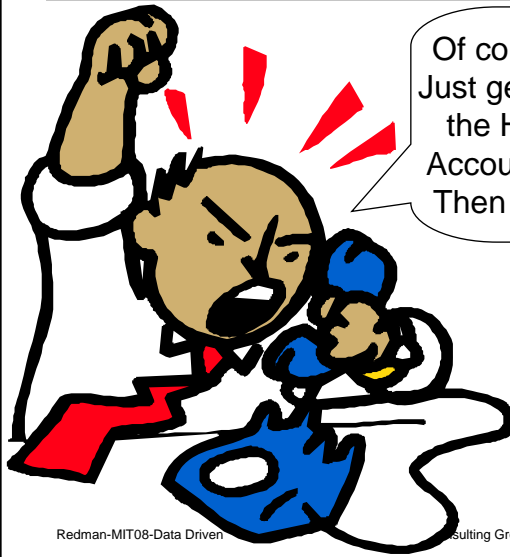


Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 36

## 1, cont. Though Universally Praised, Data Sharing is the Exception!



Of course you can have our data. Just get your 30-11 form signed by the Head of Legal, the Head of Accounting, and the Head of HR! Then we'll run it up the line here!!

NOTE: Many of *The 48 Laws of Power* (Greene and Elffers, Viking, 1998) seem to argue against sharing data.

Redman-MIT08-Data Driven

Consulting Group, 2000-2008

T. C. Redman, Page 37

## It is so easy for accountability to shift downstream!!!



$$\cos^2(x) + \sin^2(x) = 1$$

Here's how you do number 3, son

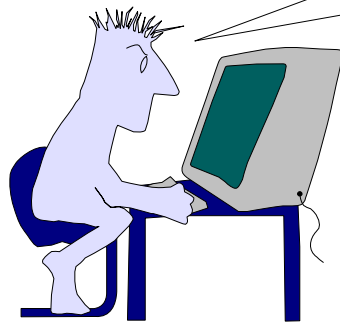


Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 38

## Who is responsible for data quality? Since the data are “in the warehouse,” it must be the CIO!



I've told that #\*%! CIO about these data problems a million times! Why can't they get them right?

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 39

## Landauer, *The Trouble with Computers* - 1993

Motivation: Roach, Strassman (early 90s): “Why aren’t computers improving productivity?”

- Computers are remarkably effective at “computing:”
  - Switching phone calls
  - Laser-guided weapons
  - Weather forecasting
- Computers are not so effective when automating poorly-defined processes:
  - Word Processing
  - Poorly-defined Business Processes

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 40



## Landauer, *The Trouble with Computers - 1993*, con't

In some cases, productivity doesn't improve, but there are other benefits:

- ATMs: Not cheaper, but always available.

Landauer's results are consistent with other results:

- Deming: "If you automate a factory that produces junk, you'll just produce junk faster."
- Data warehouses: Add little value unless decision processes are well defined.
- Enterprise Systems: Not accepted unless they match the way people work.
  - Example: \$170M Failure in FBI's "Virtual Case File."

## Implications

- You can't resolve the inter-related issues of ownership, management accountability, and quality through automation.
- Process management and improvement for quality and effectiveness.
- Automation for speed, efficiency, and scale.
- Need to explicitly get responsibility for data out of the CITO.
- New organization in "the business:" Chief Data Office.
- Those interested in data must learn how to build and use "political capital."
  - Politics is increasingly important.
  - Note: Politics is NOT inherently negative.

## Data Quality

A new and better context

The ten habits of those with the best data

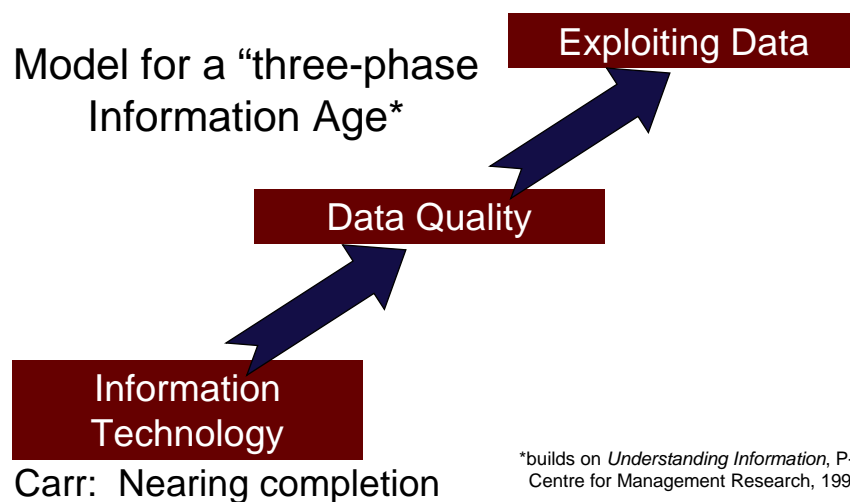
Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T.C. Redman, Page 43

## IT set the stage for, and is now giving way to, data

Model for a “three-phase  
Information Age”



\*builds on *Understanding Information*, P-E  
Centre for Management Research, 1994.

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 44

## Proper Context for Data Quality

### Existing

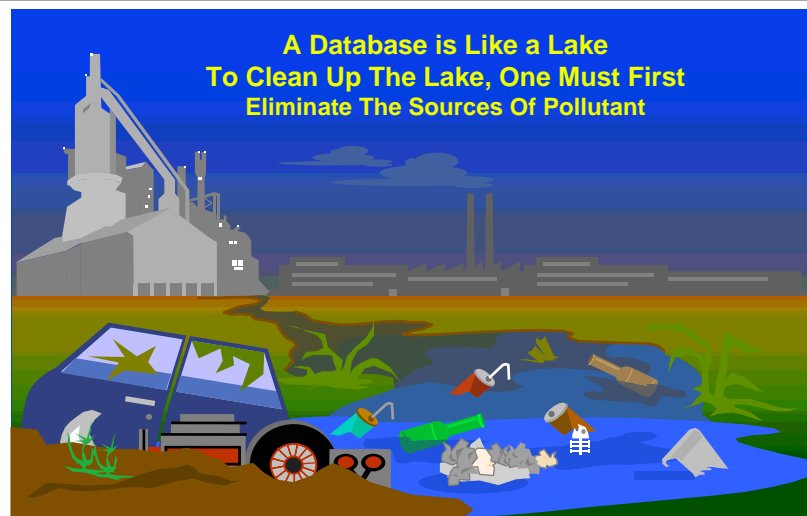
- ❑ Cost Reduction
- ❑ Internal customers
- ❑ “Master Data”
- ❑ CITO
- ❑ Standardization

### Proper

- ❑ Revenue growth
- ❑ Competitive advantage
- ❑ Data exposed in market
- ❑ “The business”
- ❑ Novelty, uniqueness

The case for data quality  
(anything for that matter) is a  
lot easier when it involves  
new revenue

## Those with the best data think “prevention”



## Approaches to Data Quality: Defect Prevention

*Most companies' current approach to data quality. Typical error rates are 1-5% and "cost of poor data quality" may be 20% of revenue.*

**FIRST-GENERATION:\***  
Inspection and Rework,  
to find and fix defects

**SECOND-GENERATION:\***  
Process/Supplier Management,  
to prevent defects

**THIRD-GENERATION:\***  
Design,  
defects "impossible"  
*Don't know  
of anyone here*

*To accomplish this, original sources of data are held accountable. Typical error rates are 1-2 orders of magnitude better and the cost of poor data quality is reduced about two-thirds.*

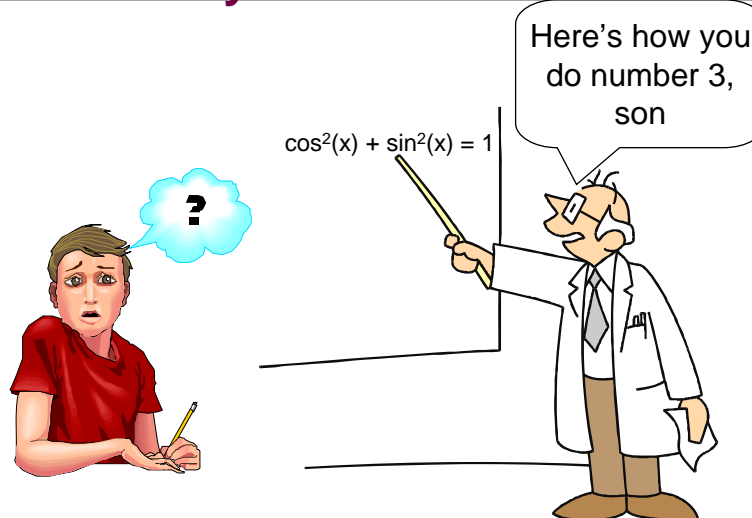
\*terms after Ishikawa

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 47

## They recognize that, left alone, accountability shifts downstream!!!



Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 48

## The (nearly-certain) results

Approach	Management Focus	Typical Error Rate	Cost of Poor Data Quality
Find and Fix (First-Gen)	The Past	1-5% (at the field level)	20% of revenue
Prevent Future Errors (Sec-Gen)	The Future	Two orders of magnitude better	Reduced by two-thirds

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 49

## Habit 1: Focus on the most important needs of the most important customers

Those with the best data adopt a customer-facing definition of quality.

In doing so, they recognize that:

- All data are not created equal. Similarly, customers, problems, and business opportunities are not created equal.
- Generally, the most important data are those needed to set and execute the company's most important business strategies.

And they focus as much of their energies on these customers, strategies, and data.

Said differently, their data quality programs are fully aligned with business strategy.

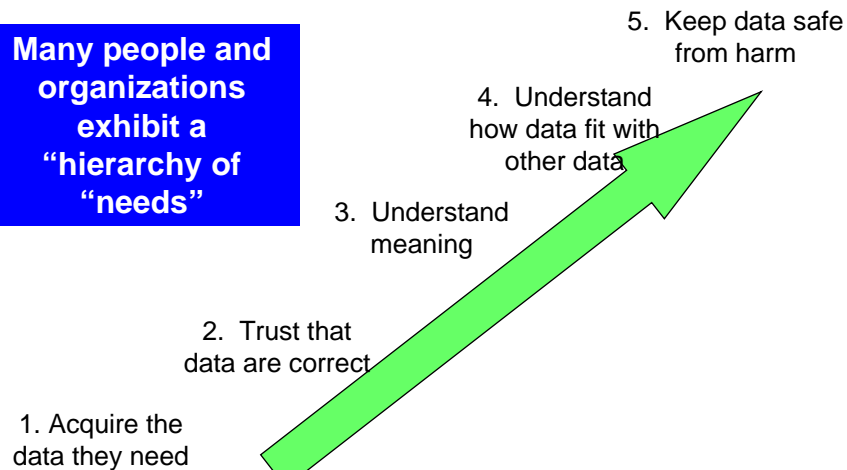
Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 50

## Data Doc's Hierarchy of Needs

Many people and organizations exhibit a "hierarchy of needs"



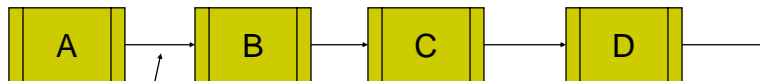
Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 51

## Habit 2. Process, process, process

They recognize that they create data via their cross-functional business processes



They recognize that most errors occur "in the white space"

They think "BIG-P"

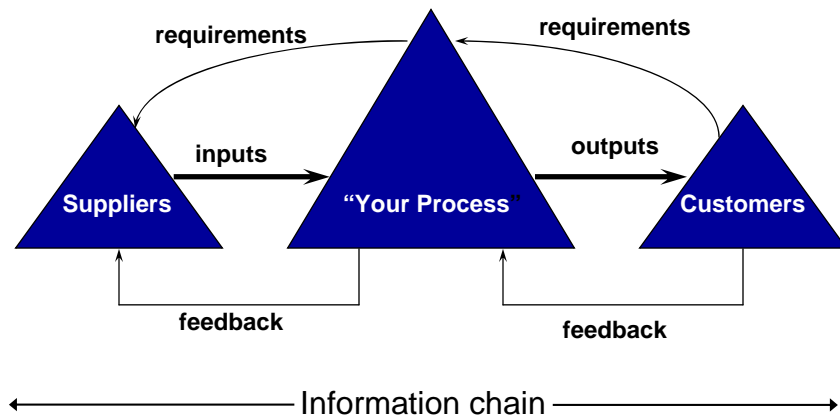
They recognize "the next guy" (serving the customer) as a customer

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 52

## They use the Customer-Supplier Model to establish requirements and feedback loops

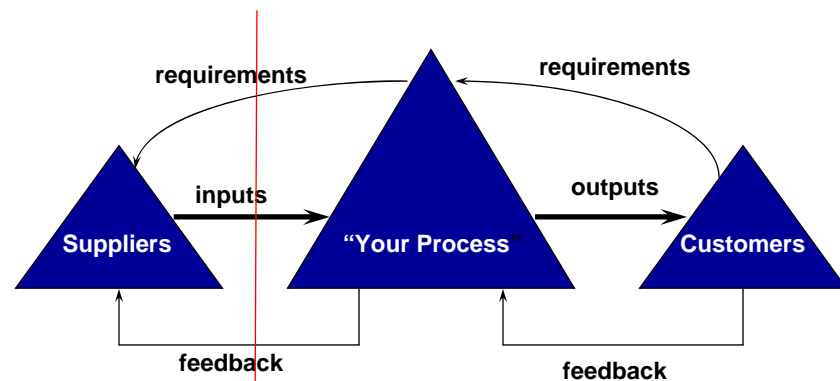


Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 53

## Habit 3: They employ supplier management for external sources of data



They expect high-quality data from outside. And invest (time) with their suppliers to get them

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 54

## Habit 4: They measure quality at the source in business terms

They define metrics with clear business implications.

Private Bank's Customer Data:

*Percent of statements with an error*

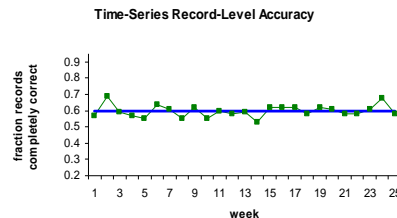
Telecom's Access Charges:

*Risk = Overbilling + Underbilling*

Many organizations:

*Fraction "perfect" records (interpreted as "work" done correctly)*

They measure continuously



They get good at interpreting results

They integrate top-line DQ metrics with other business results

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

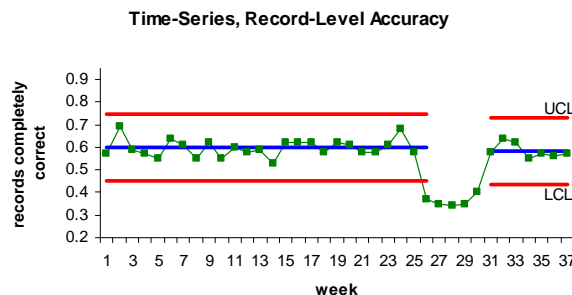
T. C. Redman, Page 55

## Habit 5: They employ controls at all levels to halt simple errors and establish a basis for moving forward

They employ simple edits to stop errors in their tracks:

Ex: (Title = Mrs., Sex = M) cannot be correct

They employ statistical control to identify process issues early and to look forward:

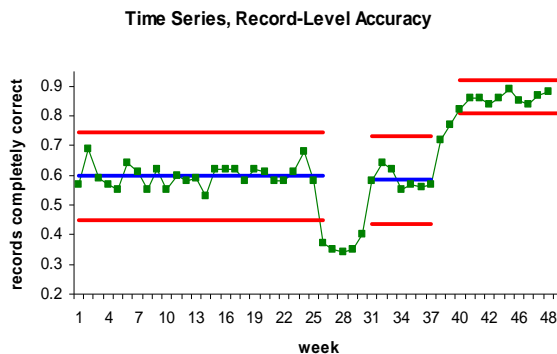


Redman-MIT08-Data

C. Redman, Page 56



## Habit 6: They have a knack for continuous improvement



They have a way of not just starting, but completing improvement projects, both to:

- eliminate root causes of error
- acquire new data

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 57

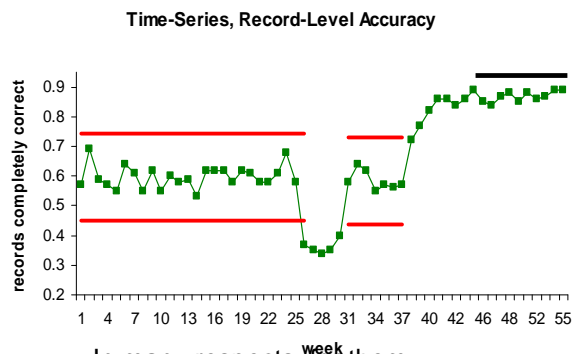
## Habit 7: Set and achieve aggressive targets

They focus not just on the level, but also on the rate of improvement

They set targets like:

- half the error rate every year
- add two significant new features every year

They decide to position themselves near the front with respect to quality in their industries



In many respects, for them planning for quality is no different than planning for revenue growth, new product development, etc.

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 58

## Habit 8: Formalize management accountabilities for data



Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 59

They recognize that responsibility for data lies with "the business," not IT.

Some codify responsibilities in policy.

My favorite (adopted for data):

"Don't take junk data from the guy upstream. And don't pass junk data on to the next guy!"

## Habit 9: A broad, senior group leads the effort

- They know that that quality programs go as far and fast as the senior person leading the effort demands.
- So a broad, committed, senior team leads the effort.

"They thought they could make the right speeches, establish broad goals, and leave everything else to subordinates... They didn't realize that fixing quality meant fixing whole companies, a task that can't be delegated."

*Dr. Juran, 1993*

Experience so far is that "data" is even tougher than the factory floor.

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 60

## Habit 10: Recognize that the “hard issues are soft” and actively manage change

They:

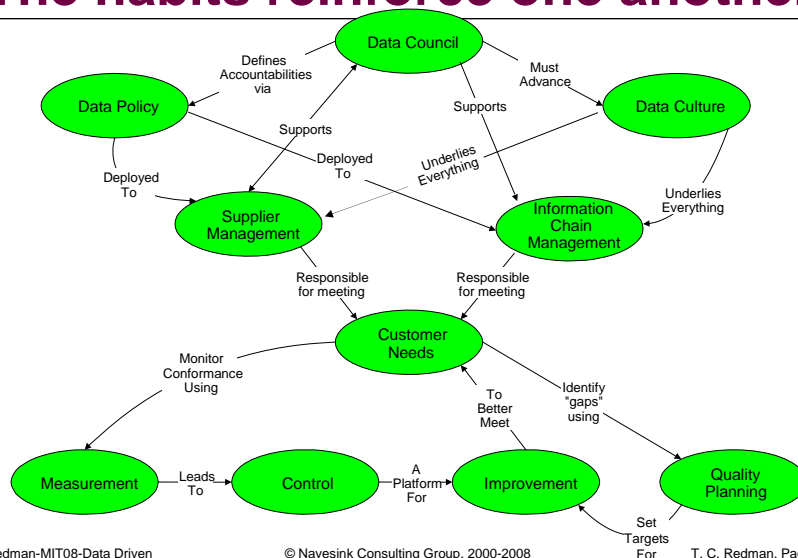
- Distinguish “I” from “IT.” They recognize that they can’t automate their way out of a quality issue.
- Start small. Create early wins.
- Actively manage change.
- Avoid unwinnable battles, especially early on.
- Build political capital.
- Over time, they build data quality into:
  - The organization
  - People’s psyche
  - To new systems

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 61

## The habits reinforce one another



Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 62

## The Ten Habits apply to all data, in all industries and government

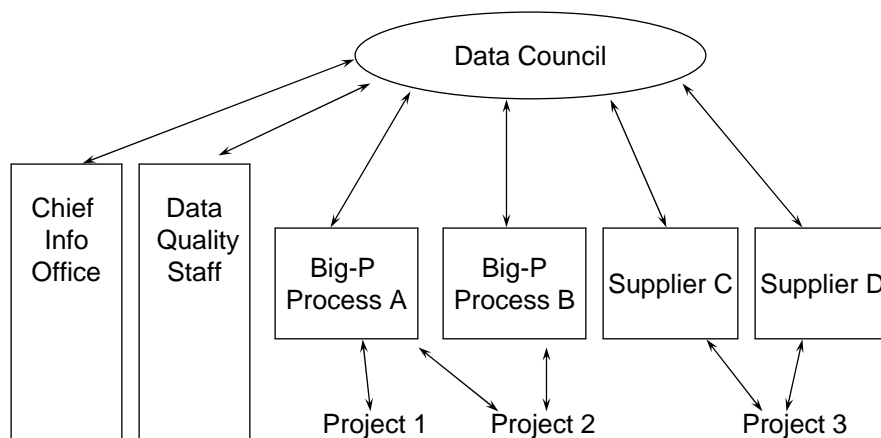
- ❑ Market, product, and people (customer and employee) data. Intelligence, scientific and logistics data. Health care data.
- ❑ Data created internally or gathered from external sources.
- ❑ Meta-data, master data, enterprise data.
- ❑ Data to be stored on paper, in operational systems, in warehouses, enterprise systems.
- ❑ Client statements, 10-Ks, prospectuses.
- ❑ Data only seen by computers and data that convince people to trust industries and companies (or not).

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 63

## Proposed Organizational Model for Data Quality\*



\*overlaid on current organization

Redman-MIT08-Data Driven

© Navesink Consulting Group, 2000-2008

T. C. Redman, Page 64

## Final Remarks:

“Data are assets” and they deserve to be managed as professionally and aggressively as other assets.

- Put them to work, especially in the market.
- Recognize that they are unlike other assets and advance the management system to account for, and leverage, these differences.
- From a quality perspective, the rigors of the marketplace should drive quality requirements.
- Follow the ten habits to meet marketplace requirements.

## What Did He Say?



**Questions?**

Thomas C. Redman, Ph.D.  
Navesink Consulting Group  
President  
732-933-4669  
[www.dataqualitysolutions.com](http://www.dataqualitysolutions.com)



The MIT 2008 Information Quality Industry Symposium



# The NATO Codification System as the Foundation of the ECCMA Open Technical Dictionary



## Overview

- History and basics of the NATO Codification System (NCS)
- The use and benefits of NCS data
- Case studies
- How the NCS laid the foundation for the ECCMA Open Technical Dictionary (eOTD)
- Video: The importance of using the right name!





## What Is the Purpose of Codification?

- To establish a common supply language throughout all logistic operations
- Language independence: All aspects of the item identification and description can be stored and exchanged in an encoded format
- To enable interoperability
- To optimize resource management by minimizing duplication in inventories

**Cataloging = Codification**



Prince Maurits  
of Nassau-  
Orange



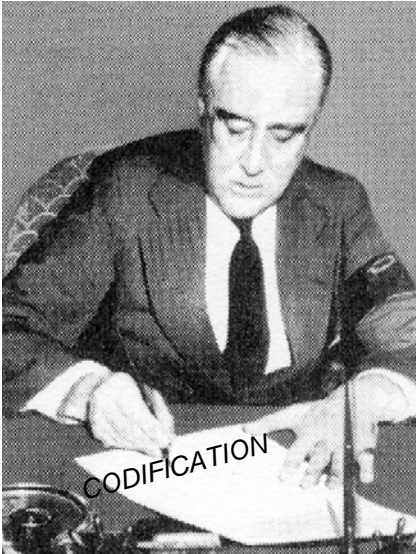


Simon Stevin  
Dutch Scientist



**Operational efficiency**

**Strict stock management**

**France  
1710  
King  
Louis XIV**




**Year 1945  
January the 18<sup>th</sup>**

**President ROOSEVELT :**

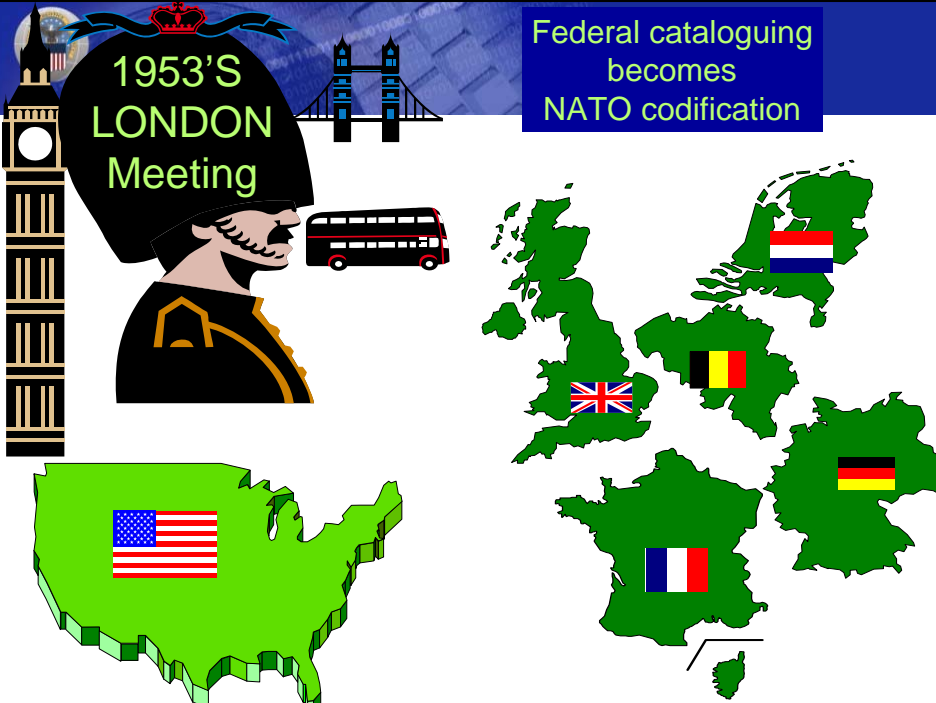
*“I request that procedures  
be examined  
to improve goods  
management  
for the efficient pursuit of  
war as well as  
for business in peacetime.”*





## Creation of NATO Alliance

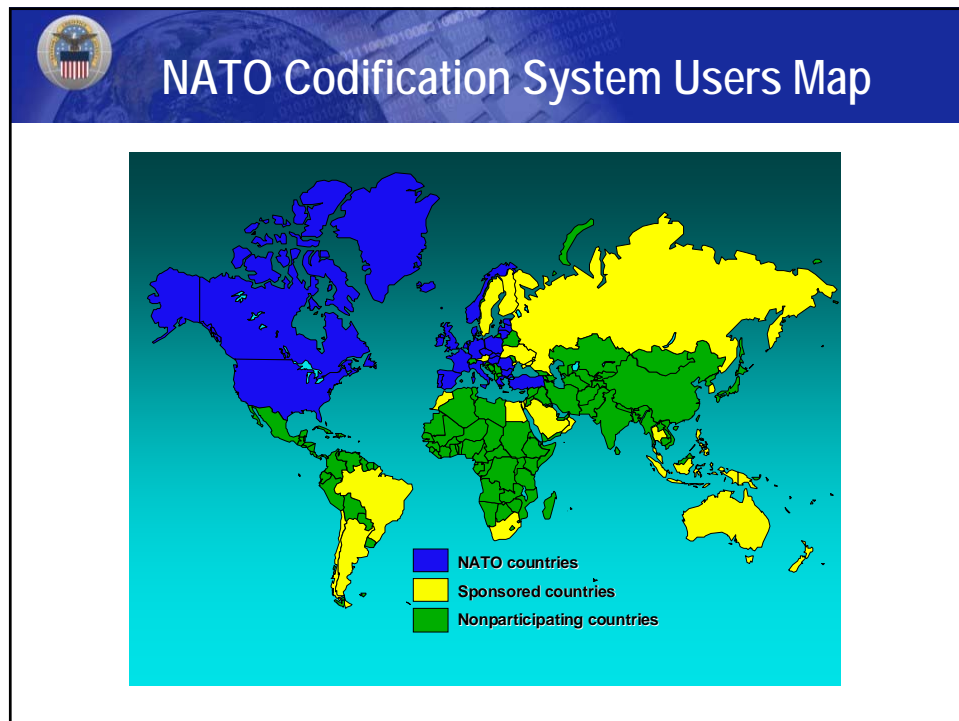
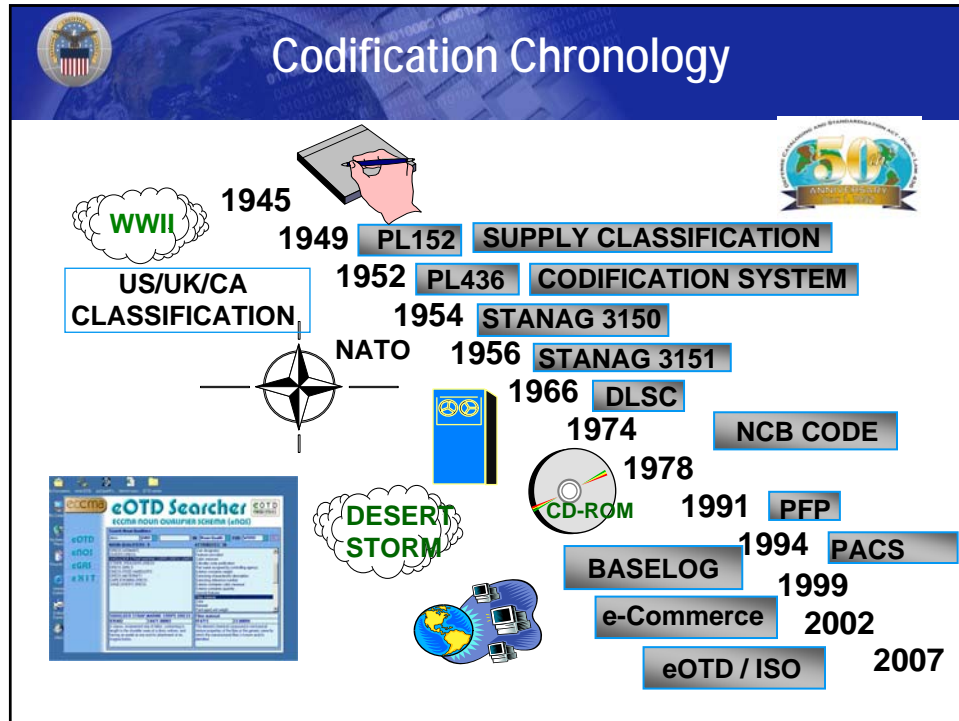
- **Political cooperation...**
  - Language problem!
- **Military cooperation...**
  - Language problem!
  - Language of logistics also a major problem
  - Each nation and even, each Armed Forces had some type of codification system



**1953'S  
LONDON  
Meeting**

Federal cataloguing  
becomes  
NATO codification

The illustration depicts the 1953 London Meeting. On the left, a stylized profile of Winston Churchill is shown next to a red double-decker bus. In the background, the Big Ben clock tower and the Tower Bridge are visible. To the right, a map of Europe highlights the founding member states of NATO with their respective flags: the United Kingdom (Union Jack), France (French tricolor), the Netherlands (Dutch tricolor), Belgium (Belgian tricolor), Germany (German tricolor), and the United States (US flag). The US is represented by a map of the United States in the bottom left corner.





## The NATO Codification System

- A set of rules and regulations that enable 26 NATO countries and 28 non-NATO nations to exchange logistic information about 16 million items of supply
- A flexible distributed information system that can be tailored to **national requirements**

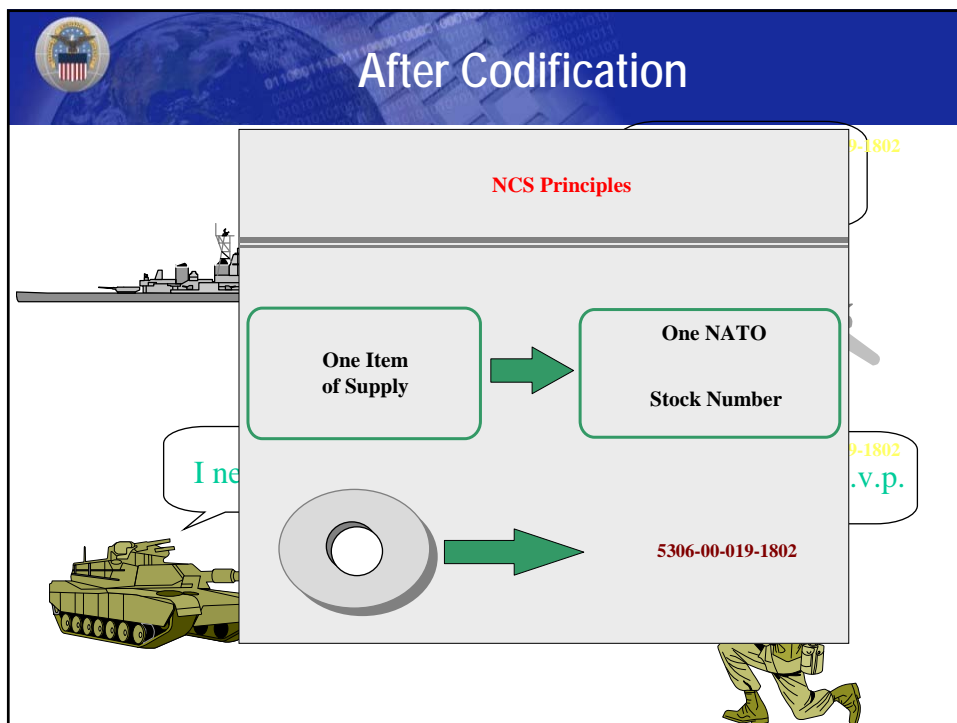
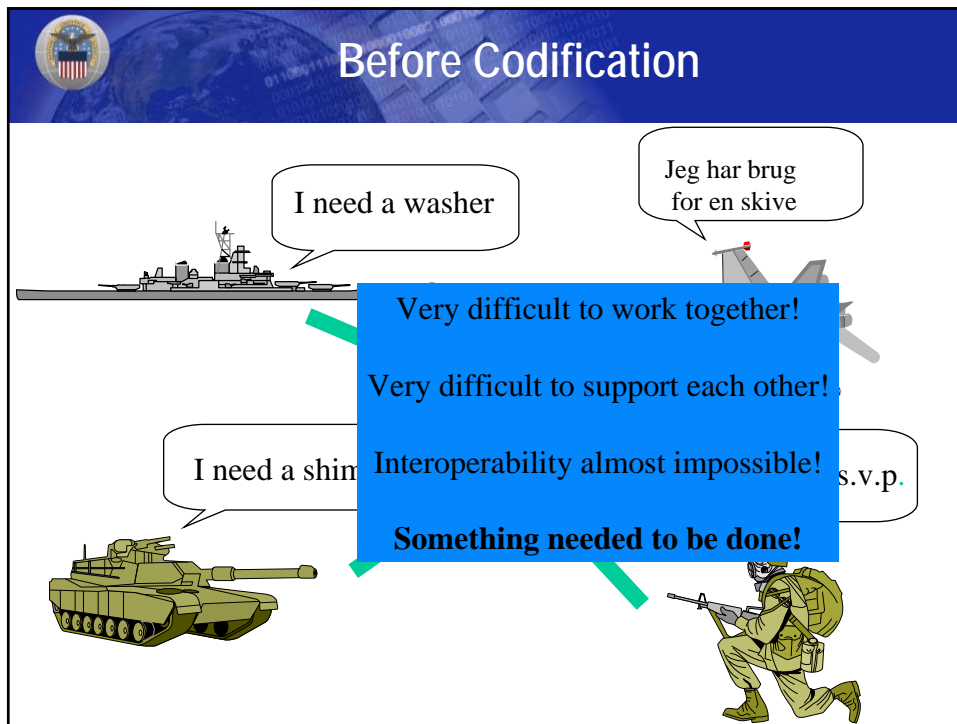


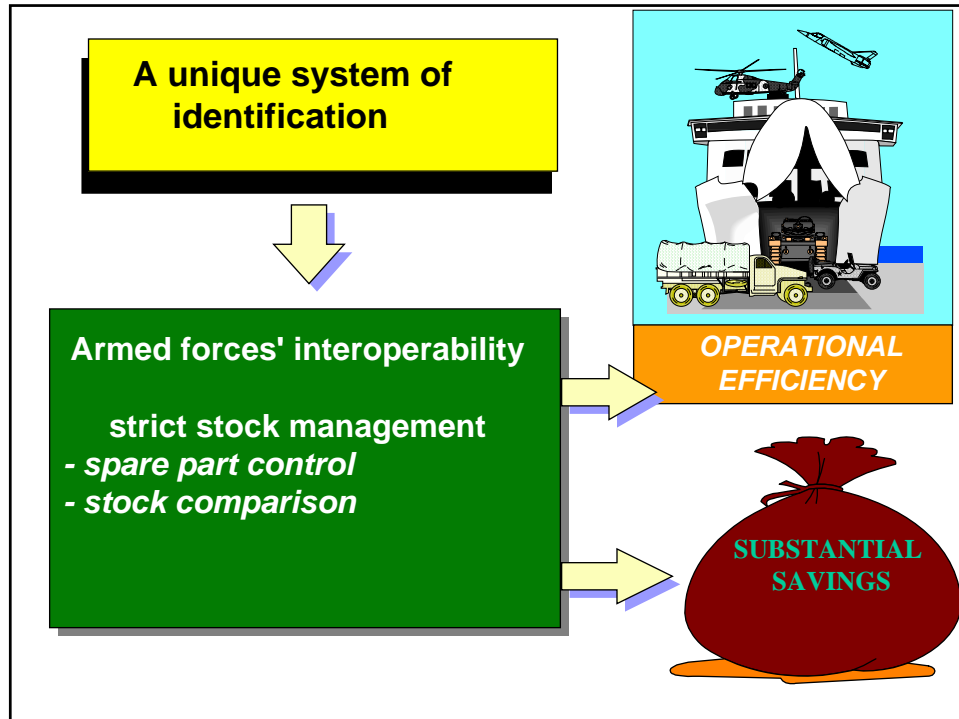
## The NATO Codification System


To facilitate logistic (co-)operation by using a uniform and common system for:

**Identification**  
**Classification**  
**Stock Numbering**

of items of supply







## The NATO Codification System

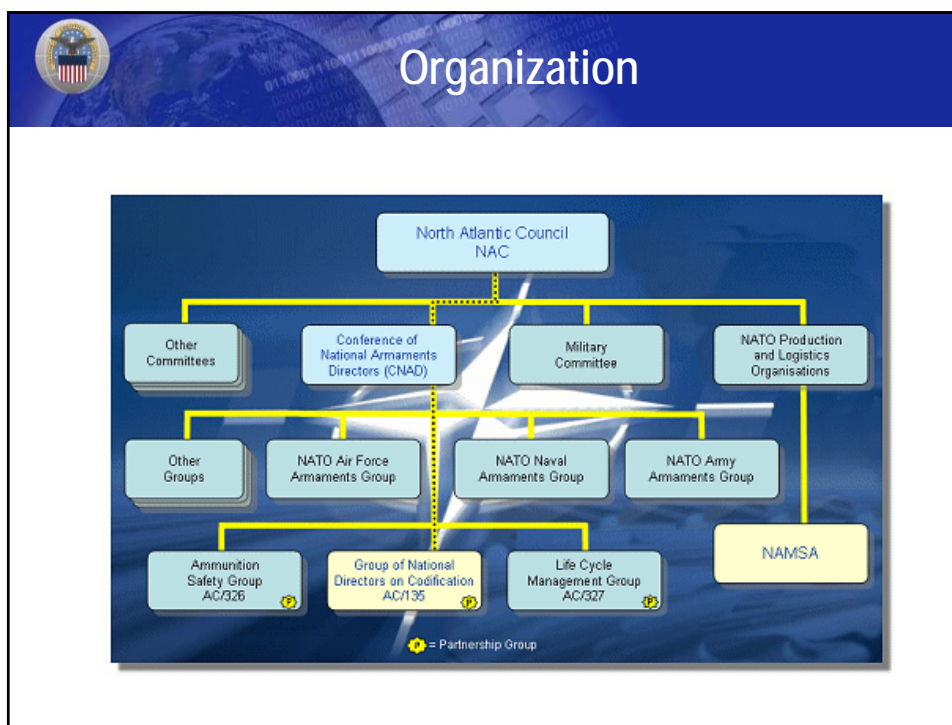
- Assignment of a uniform Item Name to an item of supply
- Use of a uniform system of classification of items of supply (**STANAG 3150**)
- Use of a uniform system of identification of items of supply (**STANAG 3151**)




## NATO Allied Committee 135 Mission

- **Provide a forum for discussion on policy matters concerning the NCS**
- **Review the progress in the implementation and application of the NCS**
- **Establish common regulations and procedures for NATO Codification**







## What Is a NATO Stock Number?

- NATO Stock Numbers represent item of supply concepts rather than an items of production
- An item of supply concept represents a cluster of characteristics related to form, fit, and function
- Many items of production may fit a single item of supply concept

**THE NATO STOCK NUMBER (NSN)**


**5905-00-7345199**

**GROUP**  
Electrical and electronic equipment and components

**CLASS**  
Resistors

**NGB Code**  
 00 = United States  
 12 = Germany  
 14 = France  
 99 = United Kingdom

**Non significant number**  
 which, with NGB Code, uniquely identifies the item



## NATO Stock Number (NSN)

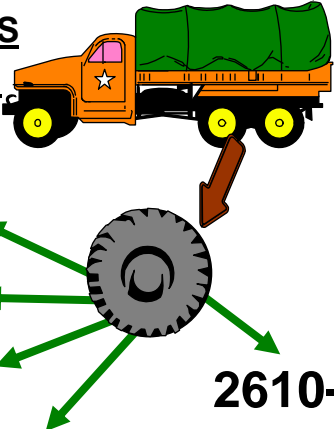
**MANUFACTURERS**  
**IDENTIFICATION**  
**SYSTEM**

DUNLOP  
11-00-20SPTGM

GOODYEAR TIRE CO  
11-00-20SRLR

GOODYEAR FRANCE  
11-00-20UNISRL

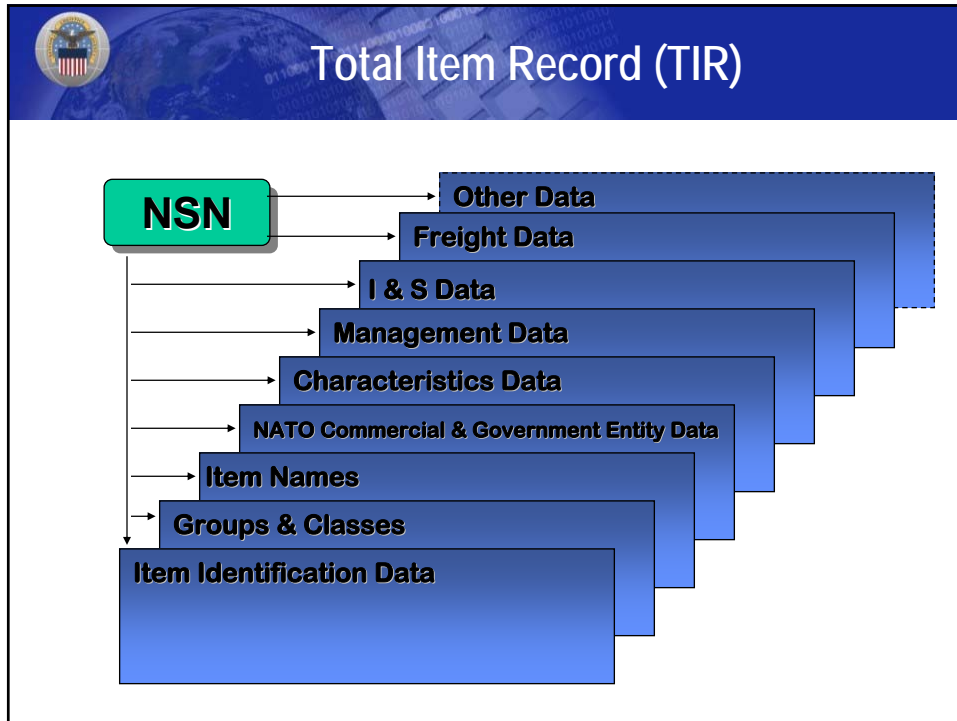
CUP SNC  
1100R20GSRT4-16PR



**USERS**  
**CODIFICATION**  
**SYSTEM**

NAVY  
ARMY  
AIR FORCE  
**OTHER COUNTRIES**

**2610-14-3224604**  
Single Stock Number



## NSNs and the Item of Supply Concept

NSN/NSSN	INC	AIN/NON AIN	TIIC	RPDMRC	FMSN
5905-00-734-5199	05311	RESISTOR 1 FIXED FI	4	9	030

MCRL 0005/0016		Seg B 0000/0014	
NCAGE Reference Number	--RM Codes-- F C U S J AA	User	Country
73168 06-250144-036	3 5 9 D KE	VB	Spain
30184 122200-001040	1 5 2 D 9Z	VJ	Singapore
05869 4171402-620	3 5 2 D ZX	VK	Kuwait
05869 4171402-801	3 5 2 D ZB	ZA	Australia
96214 418295-40	3 5 1 D KE	ZE	New Zealand
F8224 99004052	4 5 2 D ZB	ZF	France
F2663 C07H3-330UJ	3 5 2 D ZB	ZG	Germany
81349 M22684-01-0040	3 2 2 D KE	ZH	South Korea
81349 MILR22684-1	3 4 1 D KE	ZK	United Kingdom
F1621 R0M25-330UJ	3 5 2 D ZB	ZN	The Netherlands


- 1 Item Identification Data
- 2 Manufacturers' Part Numbers (References)
- 3 Users



**WebFLIS**  
Federal Logistics Information System

WebFLIS Home 4/9/2008 1:22:35 PM

**WebFLIS National Stock Number (NSN) Output Data**  
[Search again?](#)

 **NSN:** 8040001449774  
[\(Warfighter Search\)](#)  
**Item Name:** ADHESIVE  
**Query Type:** PUBLIC  
**Date of query:** 4/9/2008 1:22:35 PM

**Note:** This is a representative picture only, of this item.

**Identification** [Back to Top](#)

FIIG	INC	CRIT CD	II	RPD MRC	DMIL	DMIL INT CD	NIIN ASGMT	PMIC	ADP	ESD EMI	HMIC	HCC
A535P0	11297	X	1		A	1	1969043	A			Y	

**SCHEDULE B:**

**ENAC:**

**Reference/Part Number** [Back to Top](#)

REF/PN	CAGE-CD	STAT	RNCC	RNVC	DAC	RNAAC	RNFC	RNSC	RNJC	SADC	HCC	MSDS
985310-3	00752	A	5	1	4	ZZ						
RTV189	01139	A	5	9	4	ZZ		B			NI	808ZH
SM10192-02	07690	A	5	9	6	ZZ		B				
15567-002	10138	A	5	1	4	75						
P15-3145-...	14304	A	5	2	5	9Z	4	D				

**Management** [Back to Top](#)

EFF-DT	MOE	AAC	SOS	UI	UI PRICE	OUP	CIIC	SLC	REP	USC
2008092	DA	G	GSA	CA	\$60.63	B	U	4	Z	A
2008092	DE	G	GSA	CA	\$60.63	B	U	4	N	E
2008092	DM	G	GSA	CA	\$60.63	B	U	4	Z	M
2007274	DN	G	GSA	CA	\$66.60	B	U	4		N
2008092	IG	G	GSA	CA	\$60.63	B	U	4		I

**Characteristics (Decoded)** [Back to Top](#)

MRC	REQUIREMENT STATEMENT	CLEAR TEXT REPLY
NAME	ITEM NAME	ADHESIVE
AGXW	PHYSICAL FORM	PASTE
AKKF	QUANTITY WITHIN EACH UNIT PACKAGE	12.000 OUNCES
ALXZ	SPECIFIC USAGE DESIGN	FOR HIGH STRENGTH BONDING OF SENSITIVE ELECTRONIC COMPONENTS AND SUBASSEMBLIES
HUES	COLOR	GRAY
MATT	MATERIAL	SILICONE
SUPP	SUPPLEMENTARY FEATURES	10.3 OUNCE IN 12 OUNCE CARTRIDGE

[Search Again?](#)

**WebFLIS** Rev 3.9

Customer Service 1-877-352-2255 or DSN 661-7766 Email: [DLIS-Support@dla.mil](mailto:DLIS-Support@dla.mil)  
Privacy/Security Accessibility Contact Webmaster

This Site Reviewed Quarterly  
Last Updated: Thursday, January 27, 2005



## NCS Statistics

- **Approximately 16 million NATO Stock Numbers have been assigned**
- **Approximately 7 million by the U.S. and 9 million by the other NATO countries**
- **Approximately 30 million reference numbers have been registered on these NSNs**
- **Approximately 1.5 million manufacturers and other organizations are registered**
- **These NSNs contain more than 31 million user registrations**



## The Scope of the NCS

- **More than 47,000 structured and defined Item Names**
- **More than 27,000 properties to describe items**
- **More than 150,000 property values to discriminate one item from another**





## NSNs Are Assigned to Spare Parts




**A-10**  
**32,254 NSNs**



**M1A1 21,415 NSNs**



**HUMMV**  
**14,655 NSNs**

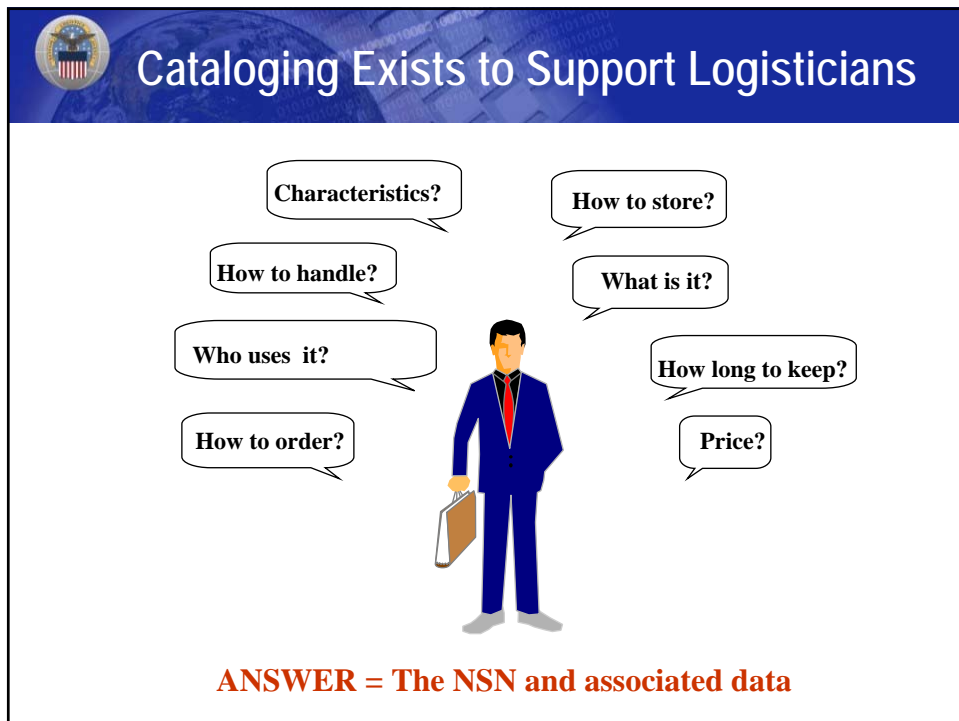
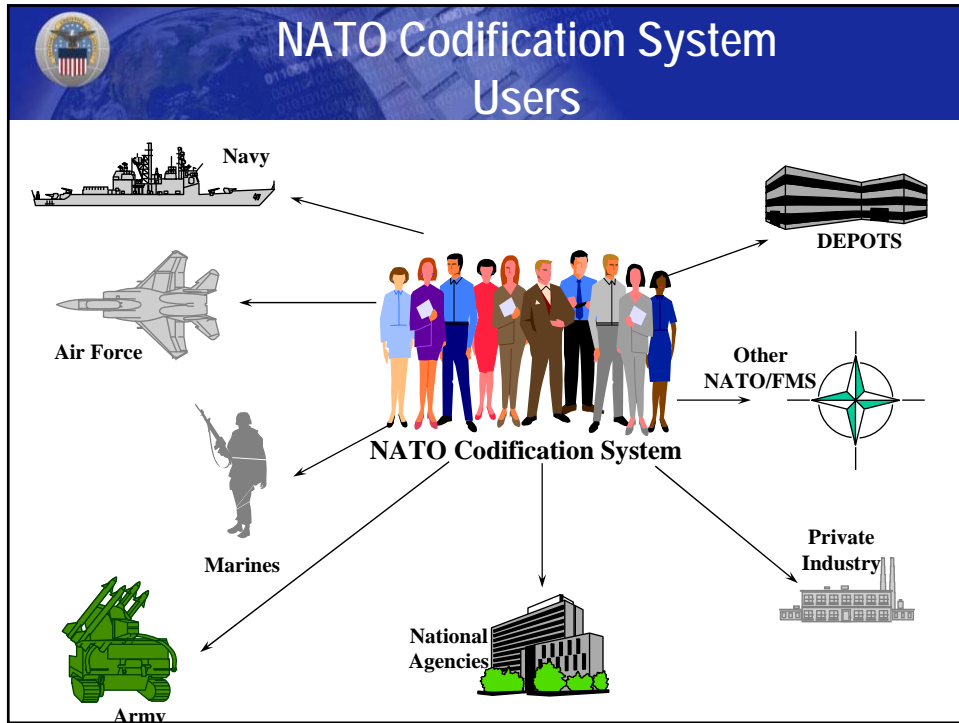


## What Is Codified?



- Bayonet Knife 1095-01-496-5107
- Night Vision 5855-01-040-0107
- Receiver-Transmitter, Radio 5895-00-009-0119
- Cover, Helmet Camouflage Pattern 8415-00-000-0168
- Eyepiece, Microscope 6650-00-001-5401
- Fragmentation Vest 8470-00-122-1299
- Cup, Helmet Chin Strap 8470-01-217-5632
- Pocket, Ammunition Magazine 8465-00-254-2771
- Carrier, Grenade 8465-00-261-5001
- Canteen, Water 8465-00-082-3054
- Belt, Cartridge 8465-00-038-5050
- Trousers, Camouflage Pattern 8415-01-422-4730
- Combat Boots 8430-00-186-6954
- Headset Microphone 5965-00-009-8679
- Battery Pack 6130-00-004-3752
- Wiring Harness Cable 5825-00-574-3712
- Magazine Cartridge 1005-00-921-5004

- Everything needed by soldiers, sailors, airmen, and marines
- Everything needed by all sectors of government, including, office supplies, parts for space vehicles, toiletries, computer equipment, and fuels





## Cataloging Link With Logistics

**Q:**

### ACQUISITION



What is the part number of known items?

"What should we call this?"

"Which NATO supply class?"

"Who manufactures this item?"

"Is it already stock listed?"

**A:**

### CATALOGING



- NATO Stock Number (NSN) records provide:
  - past sources of procurement
  - identification of the item
  - cost of the item
  - record of key logistics decisions
- NSN is key to other procurement information



## Cataloging Link With Logistics

**Q:**

### SUPPLY MANAGEMENT



"What's the last recorded price?"

What is the unit of issue?

Can another item be substituted?


What is the acquisition advice code??

**A:**


### CATALOGING



- Records initial logistics support decisions
- Records changes to those decisions throughout life cycle
- Provides means to notify all users of changes
- Offers flexibility
- Is a single, comprehensive source of information needed to manage items

 **Cataloging Link  
With Logistics**

**Q:** **MAINTENANCE**


 "What is it?"

What is the NSN so I can order it?


What is the part number?

What is the CAGE?


**A:** **CATALOGING**



- Takes the “wrench turner” from repair manual to the supply system
- Provides information on alternate sources, substitutable parts, interchangeability, and so forth
- Shows who manages the spares, how they’re managed, how much they cost, unit of issue, and so forth

 **Cataloging Link  
With Logistics**

**Q:** **STORAGE & DISTRIBUTION**

 Does it contain hazardous material?

Who is the manufacturer?

What are the packaging requirements?

What is the unit of issue?


What are the freight requirements?

Is there risk of theft?


Should it be demilitarized?

What is its value (\$)?

**A:** **CATALOGING**




- Indicates hazardous material content, precious metals content, physical security requirements, other characteristics
- Is flexible to meet national requirements for storage & distribution



## Cataloging Link With Logistics

**Q: DISPOSAL**




"What is it?"      How can I identify this item?

What is the unit of issue?      *Is it hazardous?*

Should it be demilitarized?      How should it be stored?


**A: CATALOGING**



- Helps identify unknown items.
- Provides Demilitarization information
- Provides information about hazardous material
- Helps ensure environmentally sound disposal
- Provides storage information

## Benefits of the NATO Codification System

**BEFORE**




More Inventory + Multiple Procurements + No Visibility of Assets Across Services = Wasted Resources


■Standard Data Elements

■Item of Supply Concept

■Single Manager

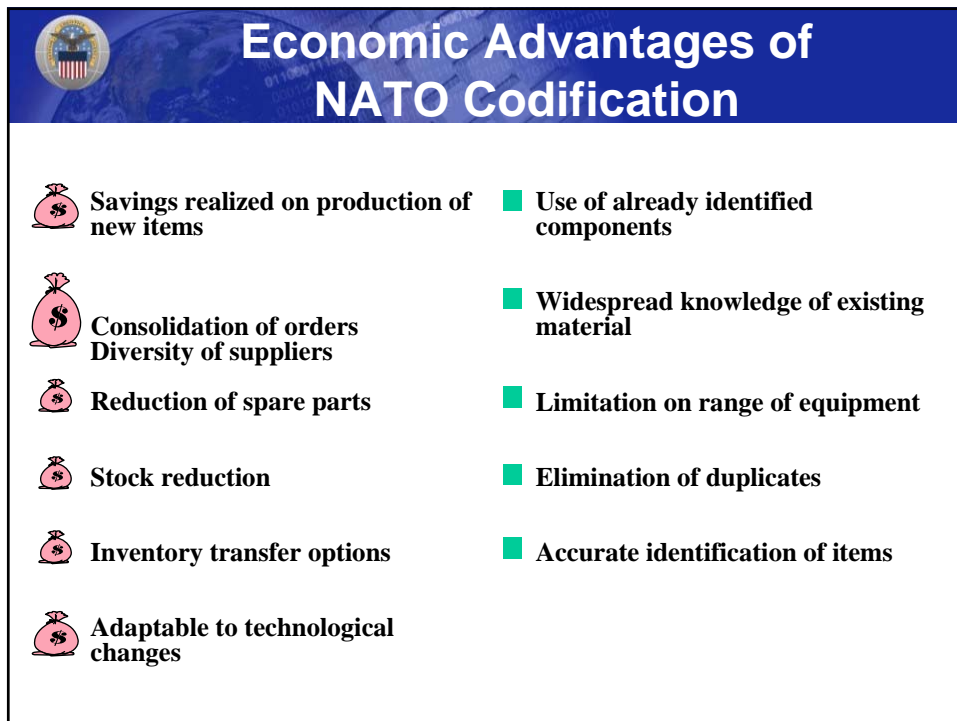
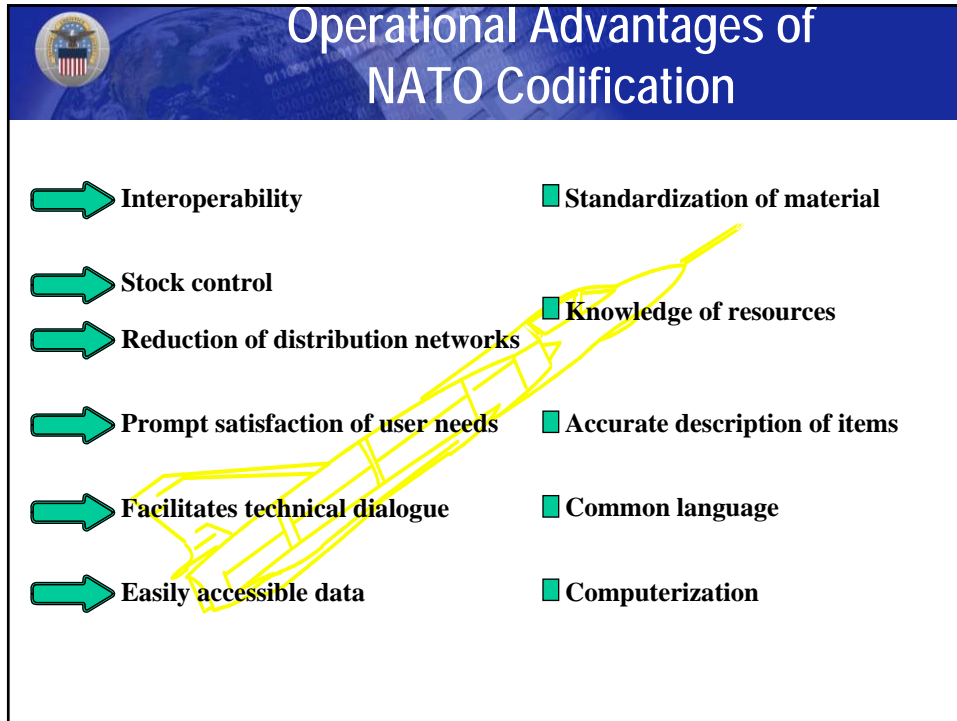


**AFTER**



Less Inventory + Consolidated Buys + Sharing Of Assets = More Effective Force










## Benefits of the NCS

■ Operational







■ Flexibility





Relations With Industry


■ Economic

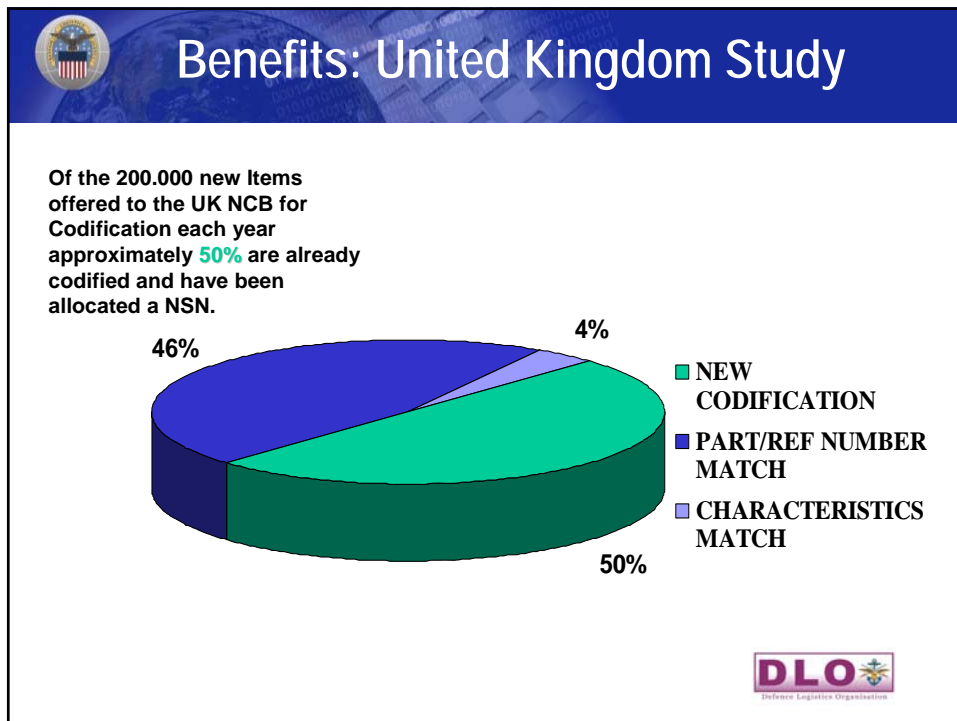




■ Interoperability

Environment







## Singapore Study: The NCS Adds Value to Supply Chain Management

### Item Entry Control

- Prevents unnecessary inventory growth and item duplication. Items managed by NSNs instead of manufacturers part numbers
- Savings in warehousing and item management fees (supply bases managed by contractors, charge \$150/line item/yr)
- About 25% of new items detected with existing NSNs each year (5-7,000 items or savings of >\$750K/yr)



## Singapore Study: The NCS Adds Value to Supply Chain Management

### OEM Part Number Breakout

- Seek out true manufacturers and their part numbers
- Promotes competitive bidding during procurement, leading to lower prices. OEM prices can be 30% lower than prime contractors prices





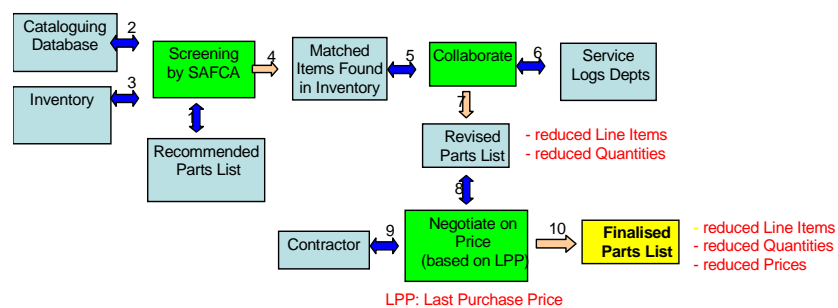
## Singapore Study: The NCS Adds Value to Supply Chain Management

### Pre-Provisioning Screening

- Items procurement intends to purchase screened against existing inventory for dead/surplus stocks to be released.
- Comparison of last purchase price to help negotiate for better pricing. Recent aircraft project, high value items >US\$10K per item screened against NSN file. Difference >US\$4M (for 156 items) between contractor & NSN prices.



## Pre-Provisioning Screening




**Can Surplus Stocks Be Released ?**  
**What is the Last Purchase Price?**



## Singapore Study: The NCS Adds Value to Supply Chain Management

- **Million Dollar Project Award - International Exposition of Innovation and Quality Circles 2005**





## Case Study: Codification In Bosnia Peacekeeping

- **Logistics operations under UN deficient because of a lack of a common technical language**
- **Many unneeded spare parts delivered under perilous conditions**
- **NATO forces found NCS of great benefit after NATO takeover**



An international non-profit membership association of industry and government master data managers and their application or service providers

### Our Mission

To increase the quality and lower the cost of descriptions through developing and promoting the implementation of International Standards for Master Data Quality



## Aims and Objectives

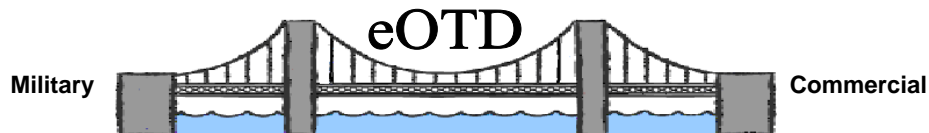
- **DLIS and AC/135 undertook the partnership with ECCMA and involvement with ISO for the following reasons:**
  - To automate the codification process
  - To improve the quality and availability of data
  - To help align the NCS with international standards
  - To increase cooperation with industry





## Military/Commercial Bridge

**“There is and always has been a philosophical gulf between the application of cataloging for military purposes and ... for commercial. ...commercial practices are not precise enough to support cost-effective military inventory management and military cataloging is far too detailed and costly for commercial purposes ...ECCMA offers a way to bridge the gulf” - Mr. Alan Williams, Assistant Deputy Minister, Canadian Department of National Defence**

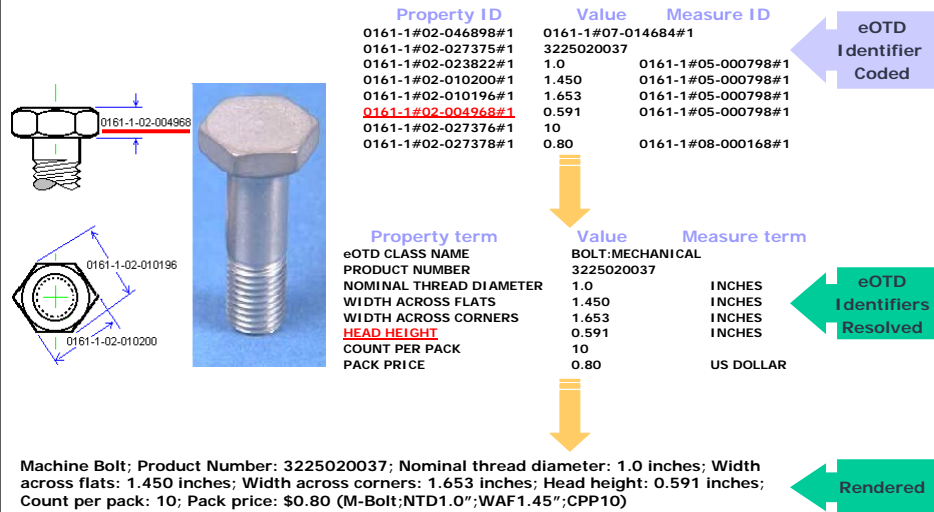


## The ECCMA Open Technical Dictionary

- The ECCMA Open Technical Dictionary (eOTD) is an open technical dictionary of cataloging concepts used to create unambiguous language independent descriptions of *individuals, organizations, locations, goods and services*
- The ECCMA Open Technical Dictionary (eOTD) is based on the NATO Codification System (NCS) with a more modern database architecture oriented toward the commercial world



## Common Terminology = Common Mapping

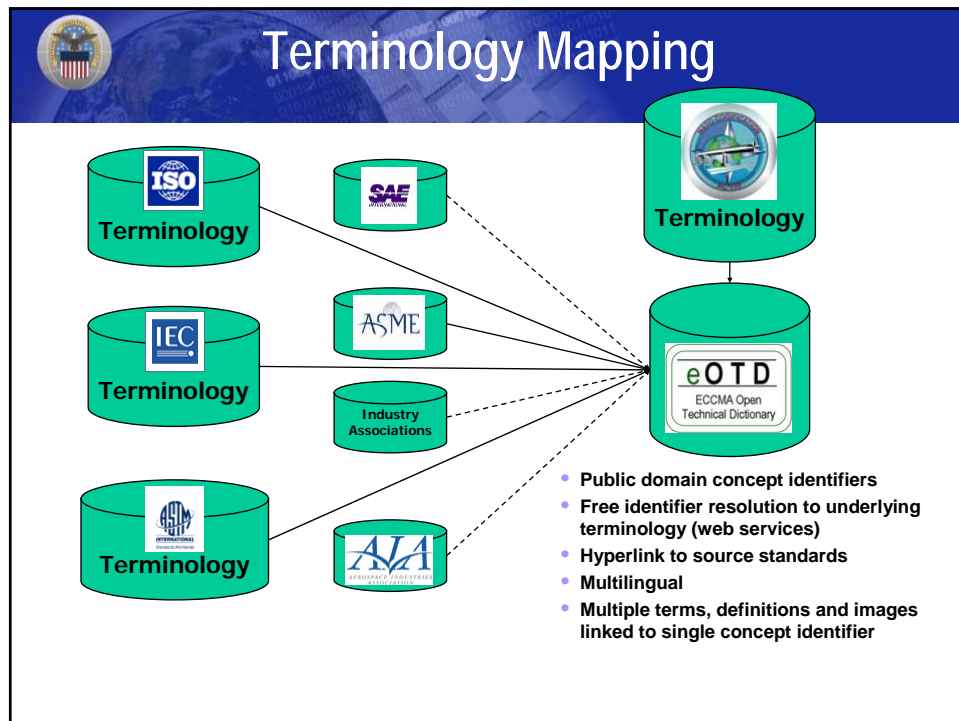
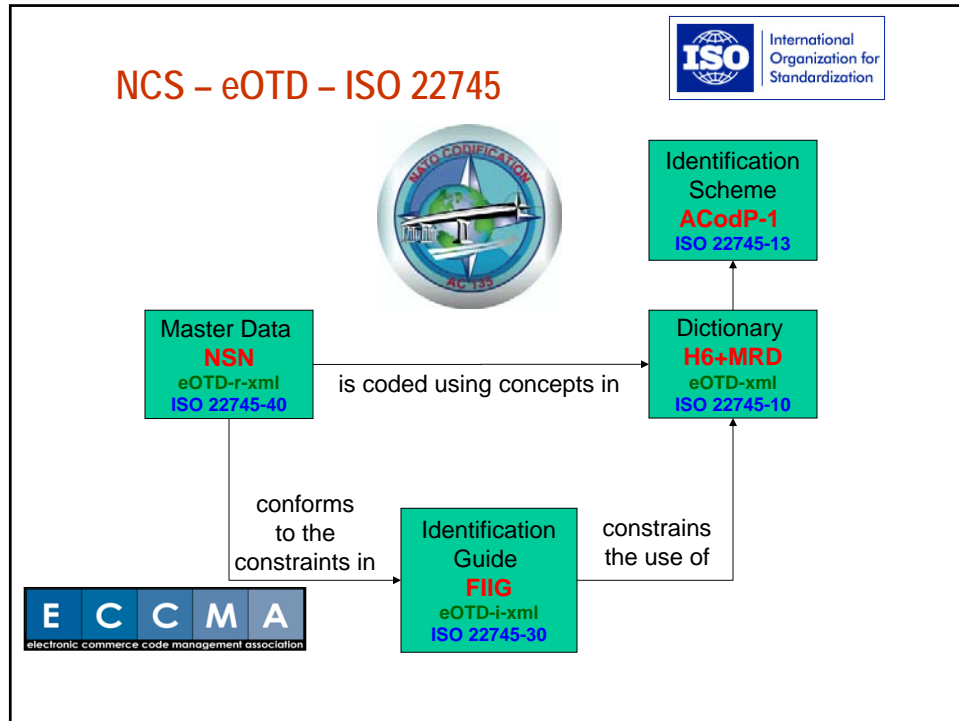


## Master Data


data held by an organization that describes the entities that are both independent and fundamental for an enterprise, that it needs to reference in order to perform its transactions

*Master data describes individuals, organizations, locations, goods, services, rules and regulations.*

- Customers
- Suppliers
- Materials
- Services
- Assets
- Locations
- Employees
- MSDS







## Transformation Through Automation

<b>Before</b>	<b>After</b>
<ul style="list-style-type: none"><li>• lack of clarity on data requirements</li><li>• disparate data format</li><li>• disparate data content</li><li>• disparate metadata</li><li>• potentially subjective human judgment</li><li>• operate as an additional process</li></ul>	<ul style="list-style-type: none"><li>• application processable data requirement statements</li><li>• consistently mapped metadata</li><li>• standard characteristic data exchange format</li></ul>

**impact: faster, better, cheaper**



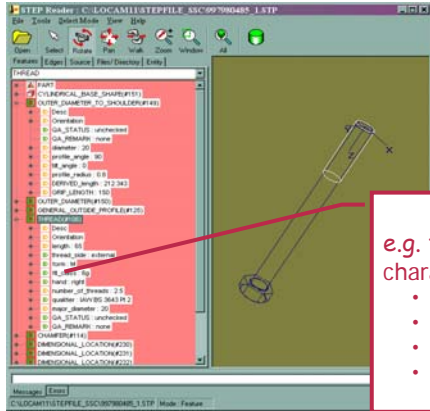
## ISO 22745: eOTD as International Standard

- Fundamentals of ISO Standard 22745:
  - Embodies eOTD metadata into international standard
  - Creates a standard data requirements statement (Identification Guide)
  - Creates a standard request for characteristic data that can be processed by manufacture's applications (PDM, ERP)
  - Creates a standard characteristic data exchange format
  - Describes how characteristic data can be tagged in STEP design files (ISO 10303)



## ISO 22745: Automation of Cataloging

- Mapping Catalog Data from Source Data

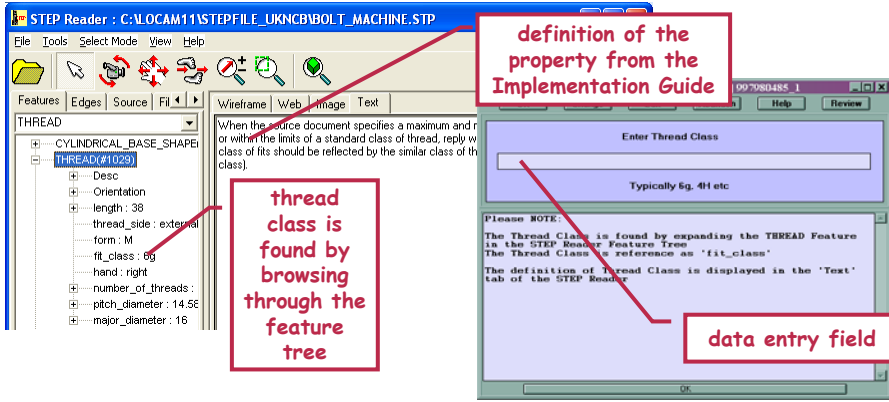


**features**  
e.g. thread characteristics including

- length (65 mm)
- form (ISO M)
- class (6G)
- diameter (20 mm)

## ISO 22745: Automation of Cataloging

- Create data one time and use throughout life cycle



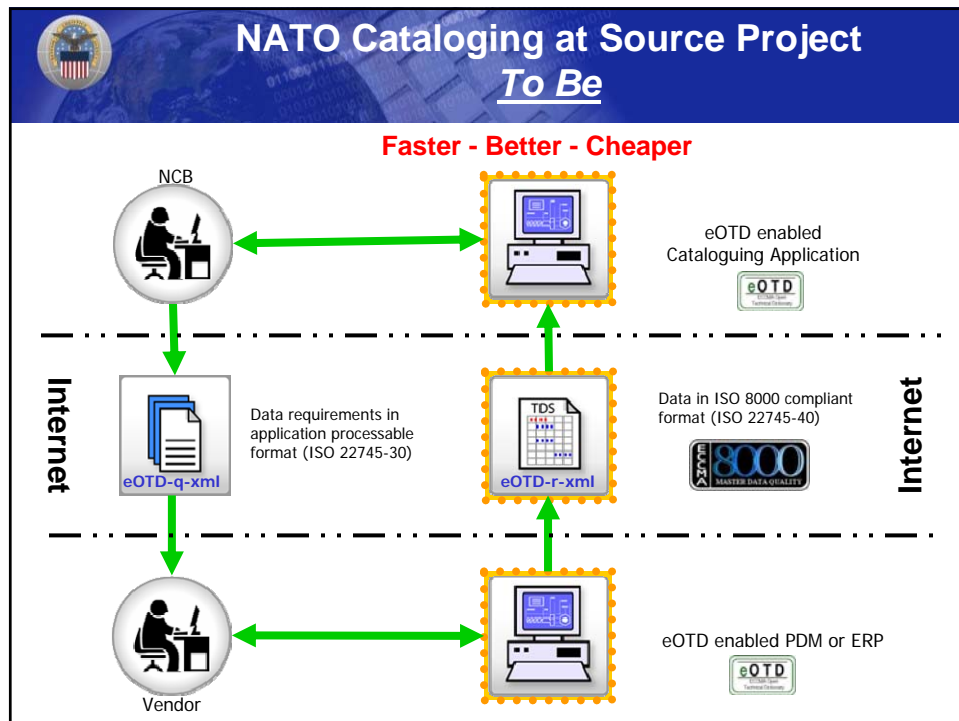
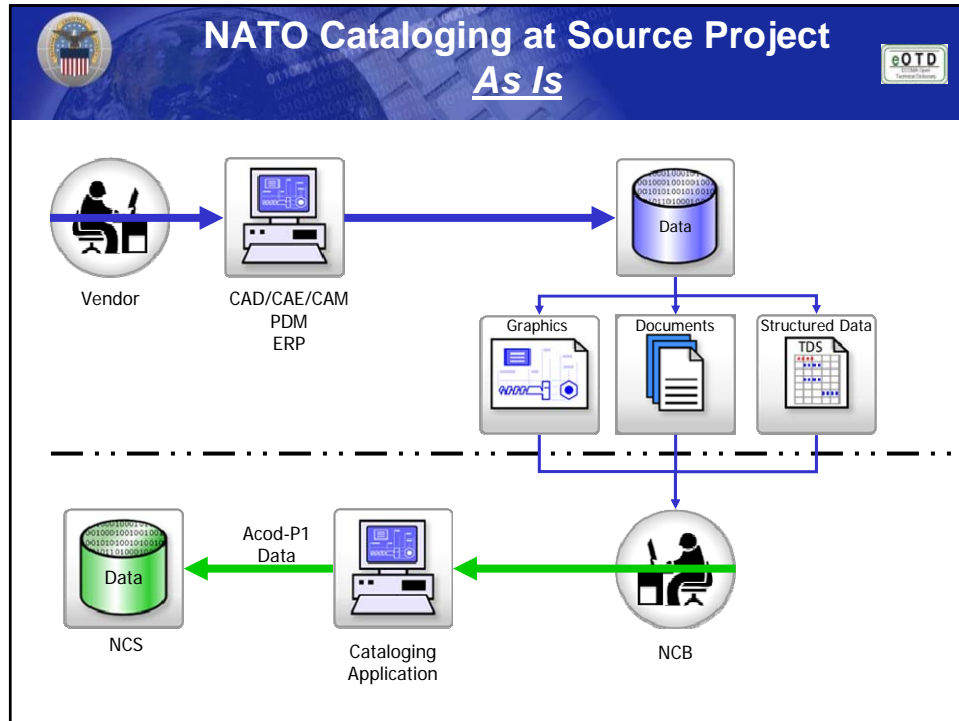
**thread class is found by browsing through the feature tree**

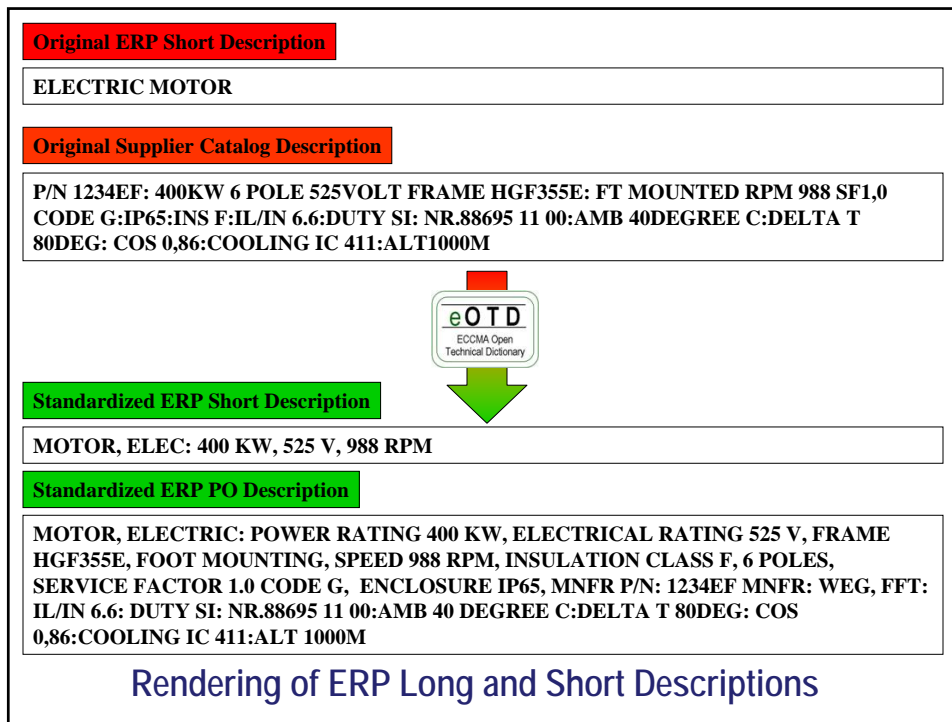
**definition of the property from the Implementation Guide**

**data entry field**

When the source document specifies a maximum and/or within the limits of a standard class of thread, reply with the class of fits should be reflected by the similar class of the class.

Please NOTE:  
The Thread Class is found by expanding the THREAD Feature in the STEP Reader Feature Tree.  
The Thread Class is referenced as 'fit\_class'.  
The definition of Thread Class is displayed in the 'Text' tab of the STEP Reader.







## ISO 8000: A Standard for Master Data Quality

- **Fundamentals of ISO Standard 8000:**
  - Set standards and a certification process for master data quality
  - Encompasses master data quality but can easily be extended into all types of data quality
  - Defines different areas of data quality
    - Provenance
    - Traceability
    - Currency
    - Completeness





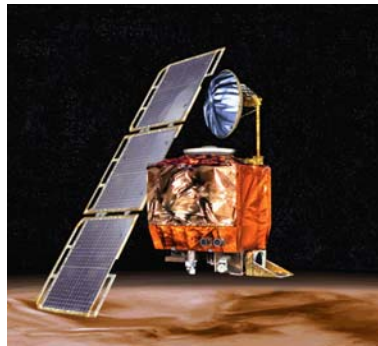

## Benefits of ISO 8000

- **Providing faster access to better quality characteristic data**
  - Faster NSN assignment
  - More complete records
  - Better search resolution
  - Fewer duplicates
  - Fewer item reduction studies
- **Benefits**
  - Higher customer satisfaction
  - Savings in design and life cycle costs
  - Reduced acquisition lead time
  - Increased supportability and safety of systems and equipment



YOUR POINTING AT IT WON'T HELP - THE COMPUTER RECORDS SHOWS NONE IN STOCK.

**What is the cost of bad data  
quality?  
In this case \$125,000,000  
The price of a Mars Climate  
Orbiter!**



Mars Orbit Insertion Burn	M/D/Y HH:MM:SS PDT (Earth Receive Time, 10 min. 49 sec. Delay)	Distance (miles)	Speed (miles/hr)	Force (Pounds)
Begin	9/23/99 02:01:00	121,900,000	12,300	143,878
End	9/23/99 02:17:23		9,840	


  

Mars Orbit Insertion Burn	YYYYMMDD EDT (Earth Receive Time, 10 min. 49 sec. Delay)	Distance (km)	Speed (km/sec)	Force (Newtons)
Start	19990923 05:01:00	196,200,000	5.5	640
Finish	19990923 05:17:23		4.4	


## International Organization for Standardization

- 156 National standard organization members (one per country)  
(AFNOR, ANSI, BIS, BSI CNIS, DIN, PKN .....)
- 192 Technical Committees
  - 3 000 Technical bodies
  - 50 000 domain experts
- Central Secretariat in Geneva
  - 150 staff
- ISO TC 184 Industrial automation systems and integration
  - ISO TC184 SC4 Industrial data (STEP)
    - ISO 22745 (open technical dictionaries and their application to Master Data)
    - ISO 8000 (Data Quality)







## ISO TC 184/SC 4 voting members




International  
Organization for  
Standardization



- Australia, [SAI](#); Standards Australia International, Ltd
- Austria, [ON](#); Österreichisches Normungsinstitut
- Brazil, [ABNT](#); Associação Brasileira de Normas Técnicas
- Bulgaria, [BDS](#); State Agency for Standardization and Metrology
- China, [SAC](#); Standardization Administration of China
- Czech Republic, [CNI](#); Czech Standards Institute
- France, [AFNOR](#); Association française de normalisation
- Germany, [DIN](#); Deutsches Institut für Normung
- Italy, [UNI](#); Italian National Standards Body
- Japan, [JISC](#); Japanese Industrial Standards Committee
- Korea, [KATS](#); Korean Agency for Technology and Standards
- Netherlands, [NEN](#); Nederlands Normalisatie-instituut
- Norway, [SN](#); Standards Norway




- Portugal, [IPQ](#); Instituto Português da Qualidade
- Russia, [GOST](#); Federal Agency on Technical Regulating and Metrology
- South Africa, [SABS](#); South African Bureau of Standards
- Spain, [AENOR](#); Asociación Española de Normalización y Certificación
- Sweden, [SIS](#); Swedish Standards Institute
- Switzerland, [SNV](#); Swiss Association for Standardization
- United Kingdom, [BSI](#); British Standards Institution
- United States, [ANSI](#); American National Standards Institute




## Justification for ISO 22745/8000

- Item reduction studies (identification of duplicates)**
  - *Save up to 15% of total inventory cost*
- Better sourcing and contracting**
  - *Save up to 20%*
- Substitution and interoperability**
  - *Part standardization during design and manufacture*
    - Increases equipment availability*
    - Can be mission critical*








## Codification: the Center of eCommerce

- Online catalogs are a critical success factor for eCommerce initiatives
- An electronic representation of goods and services: facilitates global buy/sell activities



Boeing Technology  
Phantom Works

## Phantom


**The eOTD is a foundation for design collaboration and industry standards.**

ISO 22745 and the eOTD are the foundational enablers for the breakthrough our industry needs in the next generation of direct, accurate, and effective collaboration across the supply chain at meaningful and granular levels of data exchange never before imagined.

Alton Sanders  
Senior Manager,  
IDS Engineering Standards Control Function  
PW Knowledge and Reuse Management (KARMA)



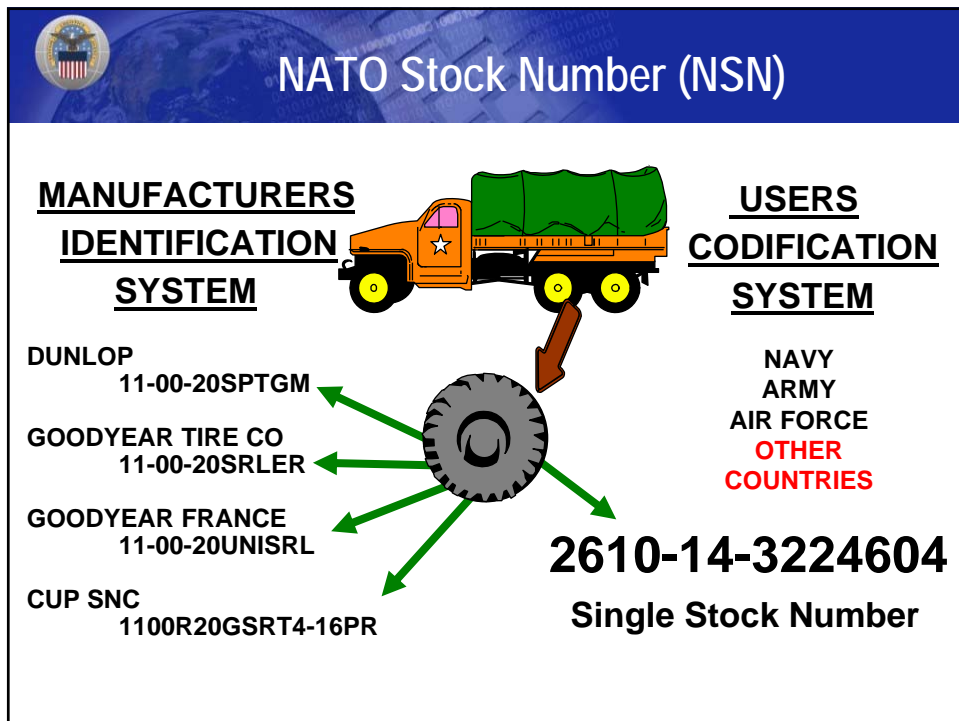


**Boeing Technology**  
Phantom Works

Phantom

*"Boeing currently buys 200 different kinds of safety glasses and 80 different shades of white paper. The defense and commercial aircraft divisions each negotiate for their own aluminum and titanium. Why can't we buy two or three kinds of safety glasses? Why can't we have standard part numbers that go across the enterprise?"*

James F. Albaugh, CEO Boeing Integrated Defense Systems,  
Business Week March 13, 2006





# The Value

**ANGLO COAL**  
52368965412 – Tyre Bridgestone 435/95 R25

**ANGLO BASE METALS**  
56329845 – Tyre BS 435/R25 Standard Purpose E3 2 Star Radial

**ANGLO PLATINUM**  
125435 – Bridge Stone 25inch 435/95

**KUMBA RESOURCES**  
965123465 – Tyre Bridgestone Part Number 12345



### One Common Anglo Number

**Standardised Long Description:**  
Tire: Pneumatic, Vehicular: Service Type for Which Designed: Loader Tire Rim Nominal Diameter: 25" Tire Width: 445mm Aspect Ratio: 0.95 Tire Ply Arrangement: Radial Ply Rating: 2" Tire & Rim Association Number: E3 Tread Material: Standard Tire Air Retention Method: Tubeless Tire Load Index and Speed Symbol: NA Tread Pattern: VHB TKPH Rating: 80

**Standardised Short Description:**  
Tire Pneumatic: Loader 25" 445mm 0.95 2"

# masterpiece

(sparesFinder)

**Provisional Corporate Match**

Original Data				Local Catalogue Matches			
Manufacturer	Part No.	Description	Price(USD)	Working Part No.	Manufacturer	Master Part No.	Score
WARMAN INTERNATIONAL LTD	B217	SEAL O-RING	0	B217	WARMAN INTERNATIONAL LTD	B217	1.094IN; 0.094IN; RBR 0
WARMAN INTERNATIONAL LTD	B109	SEAL O-RING	0	B109	WARMAN INTERNATIONAL LTD	B109	1.361IN; 0.111IN; RBR 0
JAMES WALKER	0B034107	SEAL O-RING	0	0B034107	JAMES WALKER	0B034107	2-3/8IN; 1/8IN; RBR; BLK 0
LIGHTNIN MIXERS PTY LTD	115763VIT	SEAL O-RING	0	115763VIT	LIGHTNIN MIXERS PTY LTD	115763VIT	Location: Peru Stock Code: 000391219 2-3/8 IN 1/8 IN RUBBER DEG F 0
LIGHTNIN MIXERS PTY LTD	115861PSP	SEAL O-RING	0	115861PSP	LIGHTNIN MIXERS PTY LTD	115861PSP	INSIDE DIAMETER 2-3/8 IN 1/8 IN RUBBER DEG F 0
SEW EURODRIVE	32303AV	RING	0	32303AV	SEW EURODRIVE	32303AV	CROSS-SECTIONAL HEIGHT MATERIAL TEMPERATURE RATING HARDNESS RATING COLOR SPECIFICATION/STANDARD DATA 0
STERLING FLUID SYSTEMS	45.8 - 0410B	RING	0	45.8 - 0410B	STERLING FLUID SYSTEMS	45.8 - 0410B	BLACK 0
FRANKLIN ELECTRIC	275743133	SEAL O-RING	0	275743133	FRANKLIN ELECTRIC	275743133	BLK 0
MOYNO CO	3207905210	SEAL O-RING	0	3207905210	MOYNO CO	3207905210	BLK 0
LECO CORPORATION	611-476	SEAL O-RING	0	611-476	LECO CORPORATION	611-476	1.611IN; 0.3740IN; RBR 0
LECO CORPORATION	611-477	SEAL O-RING	0	611-477	LECO CORPORATION	611-477	2-1/16IN; 0.2IN; RBR; BLK 0
INGERSOLL DRESSER PUMP COMPANY	20A11CM268	SEAL O-RING	0	20A11CM268	INGERSOLL DRESSER PUMP COMPANY	20A11CM268	8.50MM; 8.75MM; RBR 0
LIGHTNIN MIXERS PTY LTD	11581PSP	SEAL O-RING	0	11581PSP	LIGHTNIN MIXERS PTY LTD	11581PSP	11.8IN; 1/4IN; RBR; BLK 0
MARATHON PUMPS	560020360	SEAL O-RING	0	560020360	MARATHON PUMPS	560020360	1.19IN; 0.094IN; RBR 0
MARATHON PUMPS	560022360	SEAL O-RING	0	560022360	MARATHON PUMPS	560022360	1.47IN; 0.094IN; RBR; BLK 0

[Activity History / Assign User / Add Note](#)  
[View Long Description](#)

[Reject Match](#)  
[Rematch](#)  
[Approve Match](#)

Results: 1 to 15 of 159

## Catalog Compose: Cleansing Productivity Tool

The screenshot displays the 'Catalog Compose' software interface. The main window is titled 'Cataloging (TRAININGOP)'. It features a 'Criteria' section at the top with 'Results: 22'. Below this is a 'Task Criteria' section with 'Results: 15'. The central area shows a list of items with columns for 'Code', 'Description', 'Legacy ID', 'Legacy Description', 'INC', 'Name', and 'Item'. The list includes various bearing types such as 'BALL BEARING', 'BEARING BALL AND ROLLER', and 'BEARING BALL AND ROLLER, THRUST'. A detailed view of a selected item is shown on the right, including its 'Legacy Description' and 'Feedback' options. The bottom status bar indicates 'Selected: 3' and 'Naming: 1'.

### Stock Code Catalogue Data Sheet

<b>Stock Code</b>	000408187
<b>Corporate Stock Code</b>	PIGO-028721
<b>Unit of Issue</b>	EA
<b>Object</b>	VALVE
<b>Qualifier</b>	BALL
<b>Status</b>	NOT DONE

**Short Description**  
VALVE, BALL: 32MM, PUSH ON, PVC BODY, BALL & SEAT EPDM, EPDM, HANDLEVER OPERATED

**Purchase Description**  
VALVE, BALL: SIZE 32MM, CONNECTION PUSH ON, PVC BODY MATERIAL, TRIM BALL & SEAT EPDM, SOFTGOODS EPDM, HANDLEVER OPERATED

Attribute	Value
<b>BODY MATERIAL</b>	PVC
<b>CONNECTION</b>	PUSH ON
<b>DESIGN RATING</b>	*****
<b>OPERATED</b>	HANDLEVER
<b>SIZE</b>	32MM
<b>SOFTGOODS</b>	EPDM
<b>SPECIFICATION</b>	*****
<b>STYLE</b>	*****
<b>TEMPERATURE RATING</b>	*****
<b>TRIM</b>	BALL & SEAT EPDM

## Generate New Descriptions

The screenshot shows the DPMIS RAMIS software interface. The main window displays a table of material descriptions with columns for Material Number, Organization, and Description. The 'Generate Descriptions' button is visible in the top right. Below the table, there are several steps in the process: Step 1 - Gen Descr Config, Step 2 - Template Attr Config, Step 3 - Unique Reptr Config, Step 4 - Attribute Values Addr, Step 5 - Attribute Values, Step 6 - Part Numbers, Step 7 - Documentation, Step 8 - Reference Data, and Step 9 - Descriptions. The bottom of the screen shows a record count: Record: 23 to 2611.

## A Vision Realized

- The NATO Codification System is the foundation of an international standard for creating standard descriptions
- The eOTD is an open standard for encoding Master Data through the life cycle of a product – from design through disposal


The diagram illustrates the product lifecycle from design to disposal. It shows a sequence of icons: a beaker and test tubes (design), a factory (production), a warehouse (distribution), and a person at a computer (disposal). Arrows indicate the flow from design to production, production to distribution, and distribution to disposal.



## Vision of the Future

- **Logisticians will manage information more than material**
- **NCS data is the foundation for logistics information**
- **Benefits will grow**



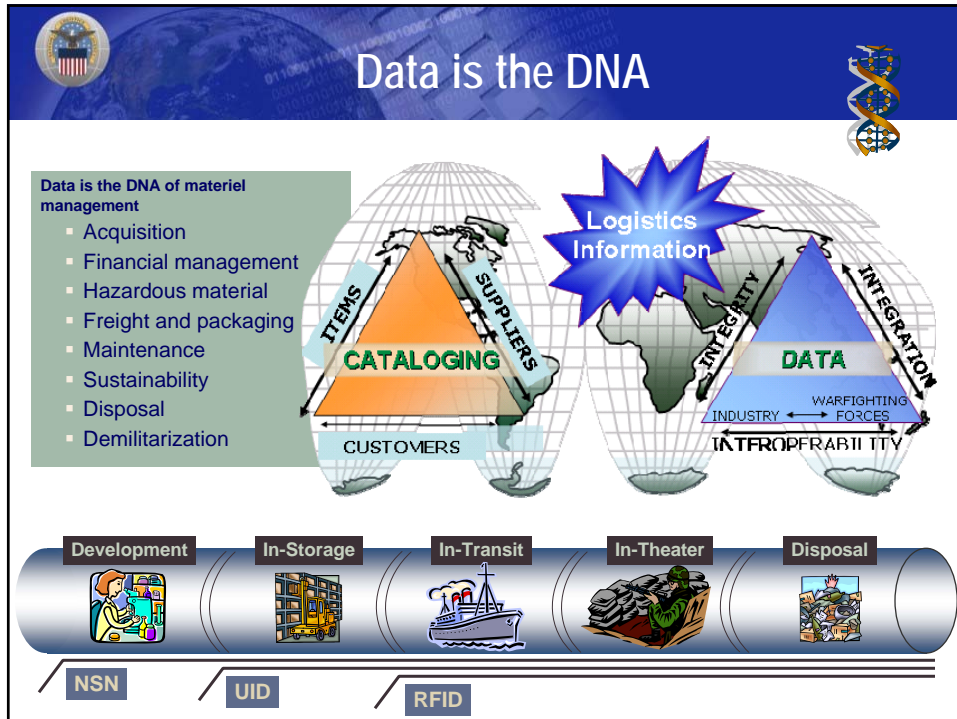


## Vision for the Future

What is impossible to do right now, but, if you *could* do it, would fundamentally change your business?

1990 Joel Arthur Barker

- **Cataloging at source (vendor supplied data)**
  - Common metadata
    - an end to data mapping
  - Requirement specifications
    - an end to incomplete data
  - Data provenance
    - an end to inaccurate information
- **Lowers the cost of cataloging and increases long term data reliability!**



**The NATO Codification System**

**The Bridge to Interoperability**

**Web Sites**

NATO Allied Committee 135: [www.nato.int/codification](http://www.nato.int/codification)  
ECCMA: [www.eccma.org](http://www.eccma.org)



The MIT 2008 Information Quality Industry Symposium



## Meeting the requirements of ISO 8000-110:2008 Master Data Quality

Peter Benson  
ISO 8000 Project Leader  
Executive director and chief technical officer  
Electronic Commerce Code Management Association (ECCMA)



The MIT 2008 Information Quality Industry Symposium



### ISO 8000 - Data Quality – Parts under development

Part 1: Overview, principles and general requirements

Part 2: Terminology

Part 100: Master data: Overview

**Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification**

Part 120: Master data: Provenance

Part 130: Master data: Accuracy

Part 140: Master data: Completeness





The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – ISO Definitions

### **information**

knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning [ISO/IEC 2382-1:1993]

meaningful data [ISO 9000:2005]

### **data**

re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing [ISO/IEC 2382-1:1993]

### **quality**

degree to which a set of inherent characteristics fulfils requirements [ISO 9000:2005]

### **characteristic**

distinguishing feature [ISO 9000:2005]

### **requirement**

need or expectation that is stated, generally implied or obligatory [ISO 9000:2005]



The MIT 2008 Information Quality Industry Symposium



## ISO 8000-110: Data Quality – Master Data

### **master data**

data held by an organization that describes the entities that are both independent and fundamental for an enterprise, that it needs to reference in order to perform its transactions

*Master data describes individuals, organizations, locations, goods, services, rules and regulations.*

- Customers
- Suppliers
- Materials
- Services
- Assets
- Locations
- Employees
- MSDS
- .....





The MIT 2008 Information Quality Industry Symposium



## ISO 8000-110: Master Data Quality - Motivation

### Supplier and Manufacturers recognize that:

- data integration is one of the keys to a long term relationship
- the ability to provide their customers with quality data is a significant differentiating factor.



### Suppliers and Manufacturers are:

- publishing the specifications of their products, capabilities and services on their web sites.
- looking to increase their visibility and understand that the best way to do this is to improve the quality of their data.

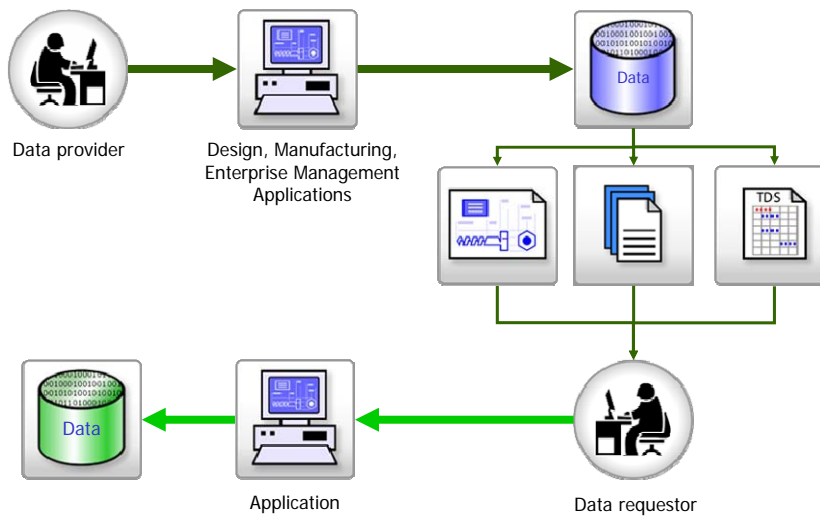
**Suppliers and manufacturers are looking for a Standard that they can use to identify the quality of their data.**

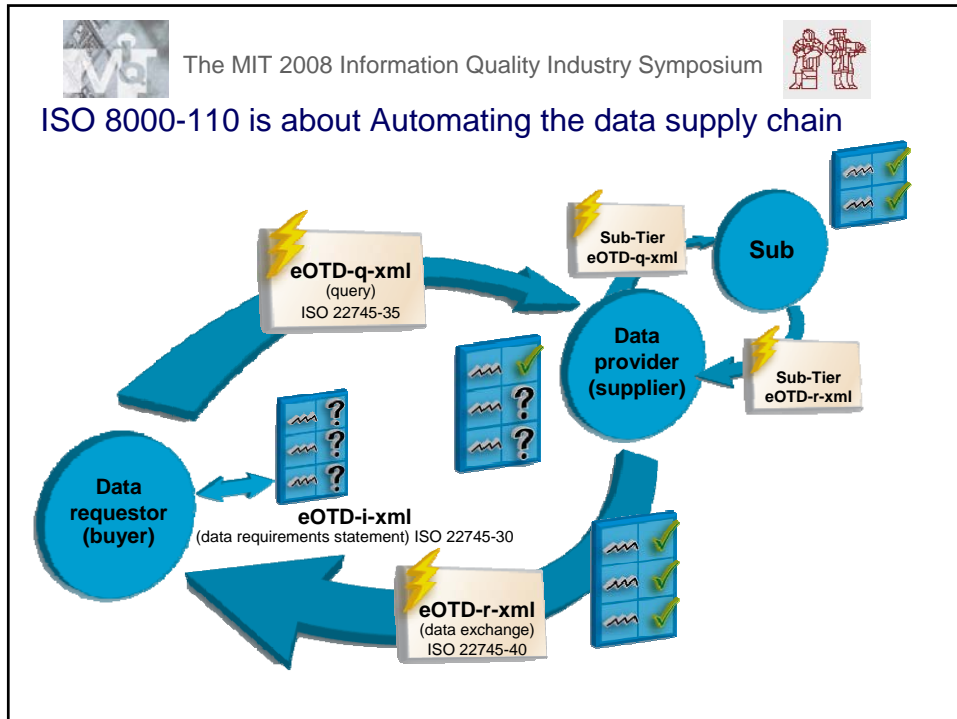


The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – Traditional data transfer





The MIT 2008 Information Quality Industry Symposium

### Data quality and Intellectual Property (IP)

All identifiers are copyright. They belong to the organization that issued them and their use is subject to the terms and conditions imposed by the issuer.

*Unless identifiers have been declared available for public use without a licence, they should never be used to retrieve data that was not supplied by the owner of the identifier unless you have specific permission to do so.*

*In order to protect your data from claims of "joint work" you should not use proprietary identifiers as metadata.*



The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – Buyer contract clause

**The contractor, sub-contractor or supplier shall supply technical data in electronic format on any of the items covered in this contract as follows:**

- a. **The data shall comply with applicable ISO 22745-30 compliant Identification Guides.**
- b. **The data shall be encoded using concept identifiers from the ECCMA\* Open Technical Dictionary (eOTD), an ISO 22745 compliant open technical dictionary.**
- c. **The data shall be provided in eOTD-r-xml, an ISO 22745-40 compliant Extensible Markup Language (xml) format published by ECCMA\*.**
- d. **The data shall be certified as ISO 8000-110 compliant.**

\* The Electronic Commerce Code Management Association (ECCMA) ([www.eccma.org](http://www.eccma.org)) is the Dictionary Maintenance Organization for the eOTD, a compliant open technical dictionary as defined by ISO 22745 and can provide technical assistance in meeting this requirement.



The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – Part 110

**This part of ISO 8000 specifies requirements that can be checked by computer for the exchange, between organizations and systems, of master data that consists of characteristic data.**

**The following are within the scope of this part of ISO 8000:**

requirements regarding conformance to a formal syntax for master data messages;  
semantic encoding requirements for master data messages;  
requirements regarding conformance to data specifications for master data messages;

**The following are outside the scope of this part of ISO 8000:**

requirements regarding the management of master data internally within an organization;  
requirements regarding recording the history of master data; records of the history of the origination, modification, and transfer of custody or ownership of data are commonly referred to as the data provenance and are defined in ISO 8000-120.  
requirements regarding accuracy of master data defined in ISO 8000-130;



The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – Part 110

### Syntax

Each master data message shall contain in its header a reference to the formal syntax to which the master data message complies. The reference shall be an unambiguous identifier for the specific version of the formal syntax that was used to encode the master data message.

### Semantic encoding

Semantic encoding is the technique of replacing natural language terms in a message with identifiers that reference data dictionary entries... Each reference shall be to a data dictionary entry contained in a data dictionary. The reference shall preserve the integrity of the recipient's data in that the reference to the data dictionary entry may be integrated with the recipient's own data without the creation of a joint work.

**Syntax and semantic resolution shall be available at no charge unless the data carries a "fee based encoding" warning label.**

### Conformance to requirements

Each master data message shall contain in its header a reference to the data specification to which the master data message complies. The reference shall be an unambiguous identifier for the specific version of the data specification that was used to encode the master data message.



The MIT 2008 Information Quality Industry Symposium



## ECCMA ISO 8000-110 Certification

- ✓ **Organization name:**
- ✓ **NCAGE** (required for NATO Codification Bureau certification only):
- ✓ **Point of contact name:**
- ✓ **Point of contact email address:**
- ✓ **Point of contact telephone number:**
- ✓ **eOTD-q-xml email address:**





The MIT 2008 Information Quality Industry Symposium



## Software application and data service provider

Certifies that a specific version and release of a software application is ISO 8000-110 compliant.

The examination reviews the ability of the application to:

- ✓ access the eOTD using web services,
- ✓ import and export Identification Guides in eOTD-i-xml,
- ✓ import and export master data in eOTD-r-xml,
- ✓ generate queries in eOTD-q-xml (where appropriate)

For a software application the certificate is valid indefinitely but only applies to a specific version of the software application;

For a data service provider the certificate is valid for one year and must be renewed annually.



The MIT 2008 Information Quality Industry Symposium



## Data requestors (Master Data Managers):

Certifies that an Organization requesting data is ISO 8000-110 compliant.

The examination reviews the ability of the Organization to:

- ✓ create Identification Guides in eOTD-i-xml,
- ✓ generate queries in eOTD-q-xml,
- ✓ read data in eOTD-r-xml.

The certificate is valid for one year and must be renewed annually.





The MIT 2008 Information Quality Industry Symposium



## Data providers (Suppliers)

Certifies a data provider as ISO 8000-110 compliant.

The examination reviews the ability of the Organization to:

- ✓ receive and process a query in eOTD-q-xml
- ✓ send data in eOTD-r-xml.

The certificate is valid for a single email address for a period of one year.



The MIT 2008 Information Quality Industry Symposium



## Software application and data service provider certification

Test 1: Accessing the eOTD using web services and reading eOTD-r-xml

1. Obtain a copy of the eOTD web services implementation guide
  - [www.eccma.org/eOTD/web\\_services\\_imp](http://www.eccma.org/eOTD/web_services_imp)
2. Obtain current eOTD-xml schemas and WSDL files
  - [www.eccma.org/eOTD/XML\\_schemas](http://www.eccma.org/eOTD/XML_schemas)
  - [www.eccma.org/eOTD/web\\_services\\_def](http://www.eccma.org/eOTD/web_services_def)
3. Implement eOTD web services
  - Open source client is available at [www.eccma.org/eOTD/web\\_services\\_client](http://www.eccma.org/eOTD/web_services_client)





The MIT 2008 Information Quality Industry Symposium



## Software application and data service provider certification

### Test 1/3: Accessing the eOTD using web services and reading eOTD-r-xml

#### 4. Test web services

- Download eOTD-r-xml test encoded and expanded files from [www.eccma.org/eOTD/test\\_files/eOTD-r-xml](http://www.eccma.org/eOTD/test_files/eOTD-r-xml)
- Import encoded eOTD-r-xml file into client application
- Run client application to resolve identifiers using web services
- Verify results against expanded file

#### 5. Demonstrate web services capability

- Request test encoded eOTD-r-xml from ECCMA certification services by sending email to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: Request eOTD web services certification test message (note encoded eOTD-r-xml will be sent as a reply to the originating email address)
- Send test expanded eOTD-r-xml to ECCMA certification services by sending email to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: eOTD web services certification expanded file verification



The MIT 2008 Information Quality Industry Symposium



## Software application and data service provider certification

### Test 2/3: importing and exporting eOTD-i-xml and writing eOTD-r-xml

1. Obtain a copy of the eOTD Identification Guide (IG) implementation guide
  - [www.eccma.org/eOTD/IG\\_imp](http://www.eccma.org/eOTD/IG_imp)
2. Obtain current eOTD-i-xml schemas
  - [www.eccma.org/eOTD/XML\\_schemas](http://www.eccma.org/eOTD/XML_schemas)
3. Implement eOTD-i-xml import and export in application under test





The MIT 2008 Information Quality Industry Symposium



## Software application and data service provider certification

### Test 2/3: importing and exporting eOTD-i-xml and writing eOTD-r-xml

#### 4. Test eOTD-i-xml import and writing eOTD-r-xml

- Download eOTD-i-xml test suite from [www.eccma.org/eOTD/test/eOTD-i-xml](http://www.eccma.org/eOTD/test/eOTD-i-xml)
- Load eOTD-i-xml test file
- Verify correctness of import using a method appropriate to the application
  - Compare eOTD-i-xml against application-specific screens, data files, etc.
  - Create and export eOTD-r-xml using imported eOTD-i-xml

#### 5. Test eOTD-i-xml export

- Download eOTD-i-xml test suite from [www.eccma.org/eOTD/test/eOTD-i-xml](http://www.eccma.org/eOTD/test/eOTD-i-xml)
- Upload and modify test eOTD-i-xml
- Save from application's native representation to eOTD-i-xml
- Verify correctness of export using a method appropriate to the application
  - Compare native representation against eOTD-i-xml file



The MIT 2008 Information Quality Industry Symposium



## Software application and data service provider certification

### Test 2/3: importing and exporting eOTD-i-xml and writing eOTD-r-xml

#### 6. Demonstrate eOTD-i-xml import and export and writing eOTD-r-xml capability

- Request test eOTD-i-xml from ECCMA certification services by sending email to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: Request eOTD-i-xml certification test message (note encoded eOTD-i-xml will be sent as a reply to the originating email address)
- Send matching eOTD-r-xml to ECCMA certification services by sending email to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: Verification eOTD-i-xml import
- Modify Identification Guide and send modified eOTD-i-xml to ECCMA certification services by sending email to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: Verification modified eOTD-i-xml.







The MIT 2008 Information Quality Industry Symposium



## Software application and data service provider certification

### Test 3/3: Generating eOTD-q-xml queries

1. Obtain current eOTD-xml schemas and examples
  - [www.eccma.org/eOTD/XML\\_schemas](http://www.eccma.org/eOTD/XML_schemas)
2. Implement eOTD-q-xml
3. Test eOTD-q-xml generation
  - Download eOTD-q-xml test encoded and expanded files from [www.eccma.org/eOTD/test\\_files/eOTD-q-xml](http://www.eccma.org/eOTD/test_files/eOTD-q-xml)
  - Manually read expanded file and generate validation, missing characteristic data and missing reference data queries in eOTD-q-xml using application
  - Verify results against encoded files
4. Demonstrate eOTD-q-xml capability
  - Request test eOTD-q-xml from ECCMA certification services by sending email to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: Request eOTD-q-xml certification test message (note encoded eOTD-r-xml will be sent as a reply to the originating email address)
  - Send eOTD-q-xml to ECCMA certification services by sending email to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: Verification eOTD-q-xml (note: email must contain all three forms)



The MIT 2008 Information Quality Industry Symposium



## Data requestors (Master Data Managers):

### Test 1/1: Creating eOTD-i-xml, generating eOTD-q-xml and reading eOTD-r-xml

1. Demonstrate ability to register an Identification Guide
  - Logon to ECCMA Identification Guide Registry
  - Create or modify an identification guide
  - Export eOTD-i-xml
  - Send eOTD-i-xml to ECCMA certification services by sending email to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: DMD verification eOTD-i-xml registration.
2. Demonstrate ability to create eOTD-q-xml
  - Send eOTD-q-xml to ECCMA certification services by sending email to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: MDM verification eOTD-q-xml (note: email must contain all three forms)
3. Demonstrate ability to read eOTD-r-xml
  - Request test encoded eOTD-r-xml from ECCMA certification services by sending email to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: MDM request eOTD-r-xml (note encoded eOTD-r-xml will be sent as a reply to the originating email address)
  - Send test expanded eOTD-r-xml to ECCMA certification services by sending expanded eOTD-r-xml to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: MDM verification eOTD-r-xml expanded file verification





The MIT 2008 Information Quality Industry Symposium



### Data providers (Suppliers):

Test 1/1: Responding to eOTD-q-xml and writing eOTD-r-xml

#### 1. Demonstrate ability to respond to an eOTD-q-xml

- Request test eOTD-q-xml from ECCMA certification services by sending email to [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: Request eOTD-q-xml certification test message PNxxxx 0161-1#01#yyyyyy (xxxx should be the part number and yyyyyy the class identifier of the item that you wish to have queried, the eOTD-q-xml will be sent as a reply to the originating email address)
- Send eOTD-r-xml to ECCMA certification services at [ECCMA\\_8000-110\\_certification@eccma.org](mailto:ECCMA_8000-110_certification@eccma.org) with subject line: Verification eOTD-q-xml response





University HealthSystem Consortium

## Data Quality in Healthcare Comparative Databases

**Steve Meurer PhD, MBA/MHS**  
**Vice President, Clinical Data & Informatics**

THE POWER OF COLLABORATION

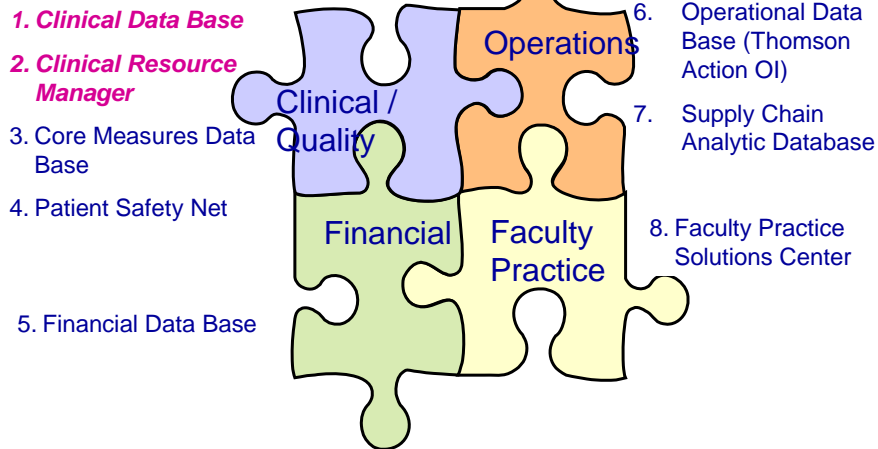
1© 2007 University HealthSystem Consortium

## University HealthSystem Consortium

- A member owned and governed consortium of academic medical centers
  - This relationship is what makes us unique
  - Approximately 90% of all major not for profit academic medical centers are UHC members
  - Affiliate hospitals are welcome and increasing in numbers (we currently have over 150 associate member hospitals)
  - Nearly 140 members and affiliates subscribe to the CDB
- UHC began in 1984, and has had only 2 CEOs
- UHC provides comparative databases, associated services, a Group Purchasing Organization, and networking opportunities

©2007 University HealthSystem Consortium 2

## 8 Comparative Databases



***“Healthcare’s single most important issue is its inability to improve”***

***Don Berwick***

***Reasons for this are many, but a major hurdle is that very little quality data is perfect***

***HOWEVER, Imperfect data can be very useful in providing direction for improvement efforts ... only if you understand the imperfections***

## R2 x I3 = Change

- Relationships
- Resources
- Information
- Incentives
- Innovation

*Using data to tell as story /  
motivate improvement*

1. Is the data **accurate**?
2. Do you have **appropriate** comparisons / targets?
3. Is the data **adjusted** properly?
4. Do you have the **necessary** data?
5. Is the data **analyzed** correctly?
6. Is the data **presented** correctly (both in print and word)?

©2007 University HealthSystem Consortium 5

## Source/Scope of CDB Data

### **Scope**

- Inpatient Discharges
- Outpatient (Currently in R&D) will include ED, observation, chemo/rad therapies, and selected ambulatory procedures
- Three years of rolling data available online

### **Source**

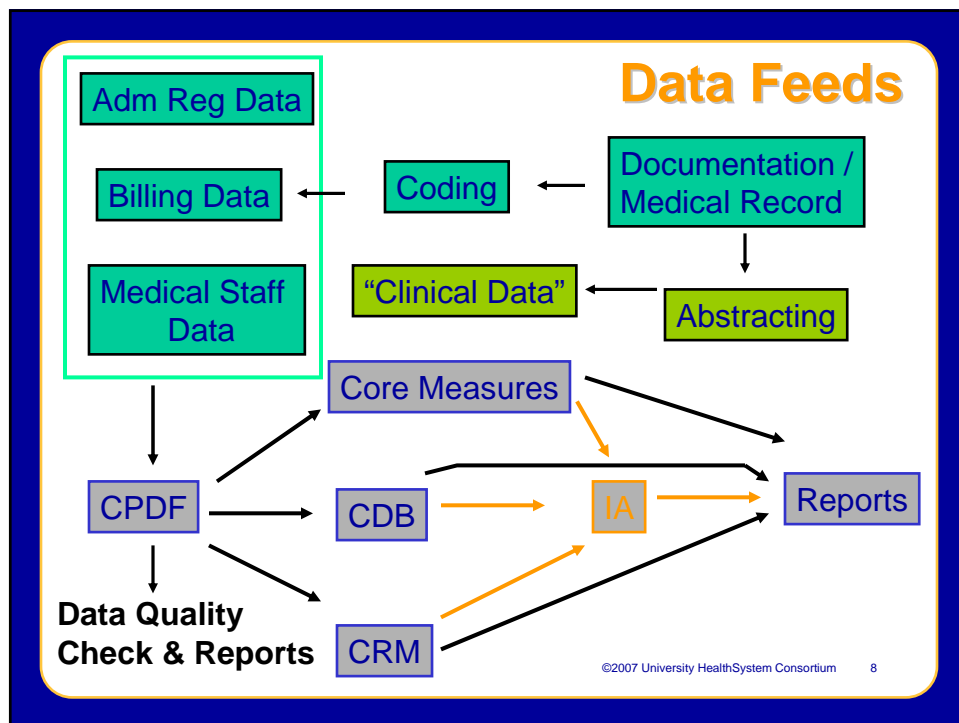
- CPDF – data feed for both CDB and CRM (line item detail)
- Monthly submission

©2007 University HealthSystem Consortium 6

## Data Quality

1. Does the data smell or look *fishy*?
  1. UHC has developed an automated process that examines member data and spits out data quality reports
    1. These reports will look at all variables and ask whether they are within a target range
    2. If a variable is not within the target and does not effect overall statistics, the data still passes
    3. If a variable is not within the target and effects overall statistics, the data is returned to the member to be fixed
2. Is the data an accurate reflection of clinical practice?

©2007 University HealthSystem Consortium 7



CDP Data Quality Reports - Microsoft Internet Explorer

## Q1 2008 Data Quality Report for XYZ Hospital

Address: <https://cimprod.uhc.edu/ODU/reports/MainUser.aspx?enddt=080331&ver=A>

FOR PERIOD 01/01/2008 - 03/31/2008 VERSION: A

### SUMMARY LISTING BY EXCEPTION TYPE

FIELD NAME	EXCEPTION MESSAGE	NUMBER OF EXCEPTIONS
ADMIT DATE	DOB = ADMIT DATE -- PDX CD NOT NB	2
ADMIT DATE	AD-DT AFTER PRC-DT -- PRC-DT=DEFAULT	7
ADMIT DATE	PRC-DT (1-4) BEFORE AD-DT, MAY BE VALID	22
ADMIT DATE	LOS > 400 DAYS, MAY BE VALID	1
ADMIT SOURCE	NB ADMIT SOURCE -- PDX CD NOT NB	2
ADMIT STATUS	DX1=NB, ADMIT STAT ASSIGNED NB CODE	28
BIRTH WEIGHT	BIRTHWEIGHT IMPUTED BY UHC	19
BIRTH WEIGHT	BIRTHWEIGHT CONFLICT BW SENT AND IMPUTE	2
BIRTH WEIGHT	ACTUAL BIRTHWEIGHT MISSING	57
DISCHARGE DATE	LOS > 400 DAYS, MAY BE VALID	1
DISCHARGE STATUS	INVALID -- DEFAULT ASSIGNED	1
DX	BLANK FIELD	12
DX	AGE INVALID FOR DX/PROC	2
DX	AGE INVALID FOR DX/PROC	7
DX	INVALID PRINCIPAL DIAGNOSIS	12
DX	CHARGES = ZERO, MAY BE VALID	1
ICU BEGIN TIME	FIELD CONTAINS THE DEFAULT VALUE	713
ICU END TIME	DEFAULT VALUE ENCOUNTERED	970
ICUDAYS	ICU DAYS != CALCULATED DAYS	321

2180

Local intranet

## Is the data an accurate reflection of clinical practice?

### Administrative vs. Clinical Data

- Debate on the usefulness of administrative data
- Clinical data requires analysis of the chart and can be very expensive
- Administrative data also comes from analysis of the chart
- The chart is a result of the clinician's (mainly physicians) documentation

## Similarities & Differences

### Administrative Data

- From medical record of discharged patient
- Began as a financial process
- Completed by educated coders
- Uses a standardized methodology
- Does not include values or test results

### Clinical Data

- From a medical record & other IT systems
- Individualized by the nature of the project
- Usually completed by clinicians
- Individualized by the nature of the project
- Could include values or test results

*The medical record is the place where clinicians take the results of tests and document the patient's condition*

©2007 University HealthSystem Consortium 11

## Literature Review

- 'Administrative data outperformed single-day chart review for comorbidity measure'.
  - Luthi et al. *International Journal for Quality in Health Care*. Vol 19. No. 4 Aug 2007. pges 225-231.
- 'Enhancement of claims data to improve risk adjustment of hospital mortality'
  - Pine et al. *JAMA*. Vol. 297. No.1 Jan 3, 2007. pges 71-6.
- 'Developing data production maps: meeting patient discharge data submission requirements'
  - Davidson, Lee and Wang. *Int. J. Healthcare Technology and Management*. Vol. 6 No. 2, 2004. pges 223-240.
- 'Comparison of administrative data and medical records to measure the quality of medical care provided to vulnerable older patients'
  - MacLean et. al. *Medical Care*. Vol 44. No. 2, Feb 2006. pges 141-8.

©2007 University HealthSystem Consortium 12



## What Variables Can be Investigated

- ✓ Risk Adjusted Outcomes – Observed and Expected LOS, Mortality and Cost
- ✓ Other variables include: Complications, Readmissions, AHRQ PSIs, Charge, CMI

### Performance based on:

- ✓ Hospitals
- ✓ Product Lines
- ✓ DRGs & MS-DRGs
- ✓ Diagnoses / Procedures
- ✓ Physicians
- ✓ Discharge Date/Month/Year
- ✓ Patient Demographics

### Resource Utilization\*:

- ✓ Blood Products
- ✓ Drugs
- ✓ Imaging Tests
- ✓ ICU
- ✓ Med/Surg Supplies
- ✓ Pharmacy
- \* CRM

Items that may be different between administrative and clinical data

©2007 University HealthSystem Consortium

13

## Uses of CDB / CRM Data

1. Ongoing consistent reports for meetings
  - Scorecards
  - Examining a DRG per meeting
  - Standard agenda items on Medical Staff Meetings, Leadership Meetings, Board Meetings
2. Improvement Initiatives
  - Drill down from scorecards
  - Answering a question
  - Improvement Priorities
3. Research
4. Improve accuracy of documentation & coding

©2007 University HealthSystem Consortium

14

## 2008 Data Quality Related Projects

- MS DRGs (complete)
  - Developed for resource use and are derived from a grouper
- Present on Admission
  - Must be consistently documented
- Bringing in 'clinical data' (e.g. lab results)
  - Infection Control Tool
- Shortening time frame for submission & return of data
- Download re-architecture
- Adding nursing units and physician names
- Post hospital mortality
  - Currently use phone follow up &/or master death file

©2007 University HealthSystem Consortium 15

## 3 Forms of Expression

- Management Reports
- Quality & Accountability Study
- CDB Online Data Tools

em Consortium 16

## Quality & Accountability Study

- Three years
- Beginning to get traction as the most statistically based ranking on quality
- Measures include: mortality (aggregate and by product line), core measure (did each patient receive all measures), AHRQ patient safety indicators with the highest signal ratios, & equity (core measures by race, gender & SES)

©2007 University HealthSystem Consortium 17

Excellent improvement seen from 2006 to 2007

**UHC University HealthSystem Consortium**

Score (Rank)	Overall	Mortality	Effectiveness	Safety	Equity	Efficiency	Pt. Center
2006	56.4 (68)	52.2% (61)	46.9% (56)	58.8% (55)	92.6% (53)	46.9% (69)	62.5% (5)
2007	64.2 (36)	58.2% (42)	65.0% (32)	65.4% (55)	100.0% (1)	43.8% (75)	No data

**Mortality: O/E Ratio** 3's or below in no domains! 8 on 2 **Domain Weight: 35%**

*Individual Product Line*

	1	2	3	4	7	8	9 / 99	10	11	12	14	15	16	17
2006	1.48 (3)		1.20 (3)	1.01 (4)	0.88 (4)		1.83 (4)	0.35 (6)		1.26 (4)	0.83 (5)	1.03 (3)	No data	0.88 (4)
2007	0.90 (5)	5.76 (LV)	0.88 (6)	1.02 (5)	1.23 (4)	0.00 (8)	0.47 (5)	0.48 (LV)	0.00 (LV)	0.79 (5)	0.96 (5)	1.11 (4)	No data	1.00 (5)

**Kid/pan tx and plastic surg**

*Hybrid Domain Scoring*

	PL Avg. Score	PL Composite	Agg. Rate (Score)	Agg. Composite	Domain Score
2006	(4.18)	52.2%			52.2%
2007	(5.30)	66.3%	0.99 (4)	50.0%	58.2%

In 2006, Mortality Domain score made up of just the PL Composite. In 2007, Mortality Domain score is the average of the PL Composite and the Agg. Composite.

**Effectiveness: Rate (Score)** **Domain Weight: 35%**

	AMI	HF	PN	SIP	READM	Metric Avg. Score
2006	76.2% (5)	62.5% (4)	46.7% (2)		6.1% (4)	(3.75)
2007	85.0% (7)	71.4% (6)	62.5% (5)	54.0% (4)	5.5% (4)	(5.20)

Note: N/A denotes no rate available and score was imputed.

**Safety: Rate per 1000 (Score)** **ht: 20%**

*Individual PSI Scoring\*\**

Note: LV denotes not scored due to low volume.

Adobe Reader - [QA2007\_Details\_Stanford.pdf]

### 3's or below in the following PSIs: Death in low mortality DRGs

Safety: Rate per 1000 (Score)										Domain Weight: 20%
Individual PSI Scoring**										
	2	4	5	6	7	8	9	10	11	
2006	1.87 (2)	126.80 (4)	0.14 (5)	2.77 (3)	4.28 (5)	0.35 (6)	3.56 (4)	2.23 (5)	3.97 (5)	
2007	1.10 (3)	128.55 (4)		1.44 (4)	3.24 (5)	0.25 (5)				PSI Avg. Score (4.64)
	12	13	14	15	16	17	18	19	20	
2006	18.69 (4)	9.08 (5)	0.94 (5)	7.17 (4)	0.00 (8)		0.00 (LV)	0.00 (LV)	0.00 (LV)	
2007	14.91 (5)			6.26 (5)		No data	No data	0.00 (LV)		(4.43)

Note: LV denotes not scored due to low volume.

Equity:					Domain Weight: 10%
2006	AMI	HF	PN		
Gender	3	2	3		
Race	2	3	3		
SES	3	3	3		
Score	8	8	9		

For each cell, score of 1 was given for p<0.01, 2 for p<0.05, and 3 for all others, including those with no data (i.e. score imputed).

Efficiency: Cost or O/E Ratio (Score)					Domain Weight: 0%
	LOS	Total Cost	Labor Cost	Supply Cost	
2006	0.97 (5)	\$12404 (3)	\$4555 (4)	\$2875 (3)	
2007	1.02 (4)	\$12645 (3)	\$4644 (4)	\$2814 (3)	

Note: N/A denotes no rate available and score was imputed.

Patient Centeredness					Domain Weight: 0%
2006: Press Ganey overall					
2007: HCAHPS q21 (% 9 or 10)					

Note: N/A denotes no rate available and score was imputed. In 2007, data was not imputed; if no data was available, no score was given.

**\* Mortality Product Line Codes**

1 - BMT	10 - HIV	18 - NeuroSurg	30 - SurgOnc
2 - Burns	11 - Kidney/PancreasTx	20 - OB	31 - GenSurg
3 - Cardio	12 - LiverTx	22 - Ortho	32 - Trauma
4 - CT/Surg	14 - MedOnc	23 - ENT	34 - Urology
7 - GI	15 - GenMed	24 - Peds	35 - VascSurg
8 - GYN	16 - Neonate	25 - PlasticSurg	36 - Vent
9/96 - Heart/LungTx	17 - Neuro	28 - Rheum	37 - SpinalSurg

**\*\* Safety PSI Codes**

2 - Death in low mortality DRGs	10 - Postop phys/meta derange	17 - Birth trauma
4 - Failure to resusc	11 - Postop respiratory failure	18 - OB trauma (vag w/ instr)
5 - Foreign body left during proc	12 - Postop DVT/PE	19 - OB trauma (vag w/o instr)
6 - Iatrogenic pneumothorax	13 - Postop sepsis	20 - OB trauma (cesarean)
7 - Selected infections	14 - Postop wound dehiscence	
8 - Postop hip fracture	15 - Accident/puncture	
9 - Postop hemorrhage	16 - Transfusion reaction	

## Management Reports

Key Indicator Report (KIR)

Clinical Outcomes Report (COR)

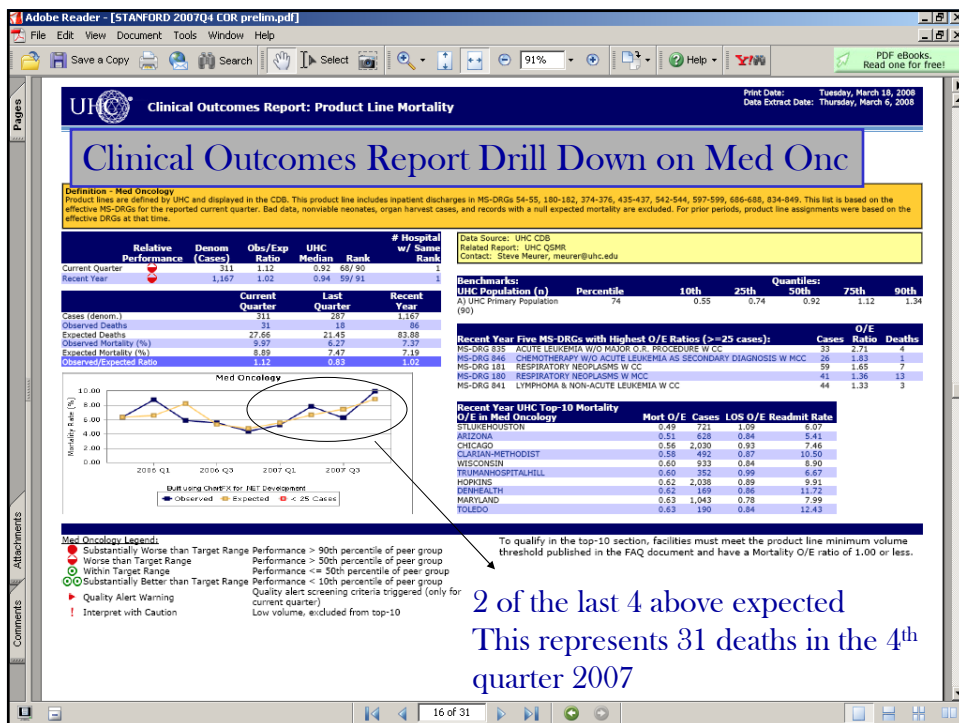
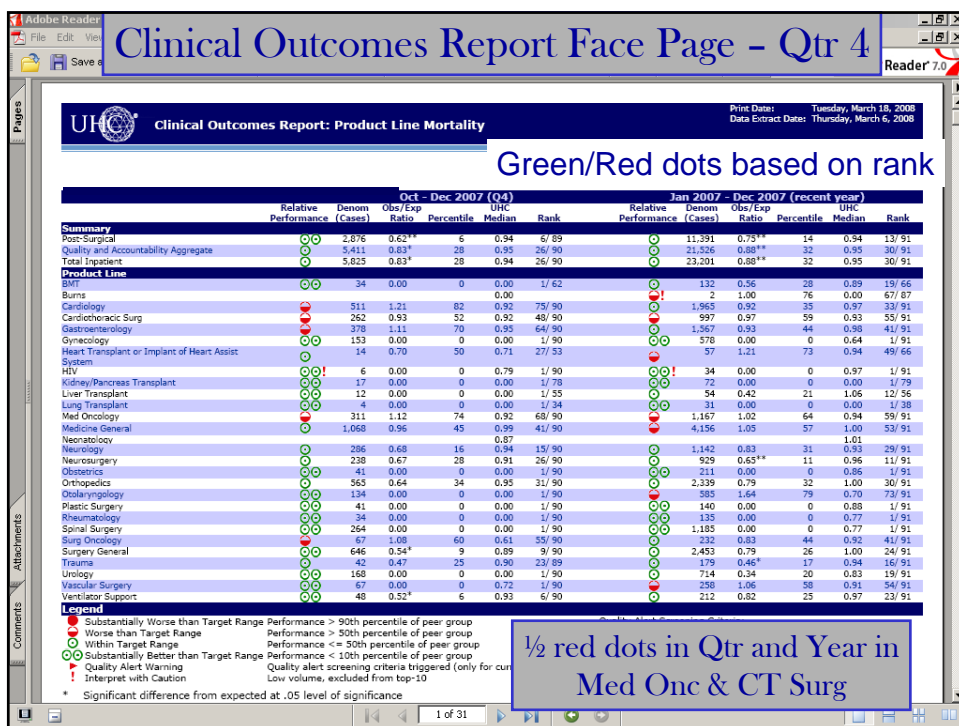
Hospital Quality Measures Report (HQMR)

Quality & Safety Management Report (QSMR)

Efficiency Management Report (EMR)

Supply Chain Report (SCR)

- Semi-static reports you receive quarterly
- KIR can be thought of as a balanced scorecard
- Widely dispersed among the membership
- The more databases you are in, the more data you will receive



**CDB Interface Default Report**  
*Volume, LOS, ICU, Complications, Mortality*

**LOS Summary By HCO**

HCO	Cases	LOS Outlie	Mean LOS	StDev LOS	Mean LOS	LOS Index	Variance (Days)	% ICU Cases	Mean ICU
	22,231	145	5.03	**	8.46	5.50	0.91	-10,411	25.76
	23,203	107	5.75	**	7.92	5.54	1.04	4,880	19.84
	30,074	366	6.11	**	10.30	5.42	1.13	20,947	16.39
	32,618	121	5.21	**	8.08	5.69	0.91	-15,818	15.59
	55,317	220	4.61		6.81	4.57	1.01	2,064	10.82
	30,673	198	5.59	**	8.07	5.78	0.97	-5,823	25.15
	21,774	174	5.71	**	8.57	5.44	1.05	5,996	21.82
	52,817	254	5.00	**	7.21	4.81	1.04	10,442	11.15
	44,004	269	4.89		7.52	4.87	1.00	984	14.42
	44,586	307	5.84	**	9.42	5.23	1.12	27,368	15.65
	59,435	286	5.08	*	8.21	5.15	0.99	-4,225	36.96
	36,363	223	5.76		10.26	5.79	0.99	-1,067	9.86
	23,589	142	5.61	*	8.14	5.49	1.02	2,842	18.83

*Time Frame is CY 2007*

Sum=28235.44

**CDB Interface Default Report (cont.)**  
*Volume, LOS, ICU, Complications, Mortality*

**LOS Summary By HCO**

HCO	Cases	Risk Pool Cases	% With Comp's(2)	% Deaths (Obs)	% Deaths (Exp)	Mortality Index	% Early Deaths
	22,231	8,008	18.32	2.07	**	2.70	0.77
	23,203	13,116	19.27	2.74	**	3.11	0.88
	30,074	14,816	19.5	2.10	**	2.51	0.84
	32,618	13,715	14.33	1.65	**	1.87	0.88
	55,317	18,062	13.24	1.34	**	2.17	0.62
	30,673	13,146	19.41	2.63	**	3.45	0.76
	21,774	9,421	16.98	1.97	**	2.74	0.72
	52,817	24,244	15.53	1.82	**	1.99	0.92
	44,004	15,757	14.53	2.25		2.15	1.05
	44,586	18,589	19.14	1.73		1.83	0.95
	59,435	32,159	20.06	1.62	**	1.93	0.84
	36,363	13,039	17.97	1.38	**	1.59	0.87
	23,589	12,586	17.4	1.94	**	2.58	0.75

Sum=41378.13

MSDRG 163 – Chest Px w/ MCC exp of 36%  
DRG exp of 24%  
20 day LOS exp of 20 days  
SOI and ROM of extreme

High c-value of .924, close to 20,000 cases in the model

***Although administrative data has no results, it will include all conditions that are diagnosed from notes and results***

UHC CDP Online - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Close Report  
Print Report  
Save Report

Related Links  
[Zyns Health](#)

Clinical Pathway  
Constructor:  
[Lumb Laminect. Presp](#)  
[Eral](#)

Evidence-Based  
Forecasters:  
[Summary](#)

Case 2 of 4

Patient ID	Encounter Number	Admit Date	Admit Day	Admit Source	Admit Status
0707760	12151569	3/6/2007	Tuesday	Non-Facility Point of Origin	Urgent

Discharge Date	Discharge Day	Discharge Status	Age	Norm IIB	Sex	Race
3/20/2007	Tuesday	Expired (all in-hospital deaths except for Medicare or CHAMPUS hospice patients)	65	No	Male	White

ICU Days	Early Death	Base MS-DRG	MS-DRG	DRG	APR-DRG	Product Line (MS-DRG)	Product Line (DRG)	Severity of Illness	Risk of Mortality
1	No	147	457	546	304	Spinal Surgery	Spinal Surgery	Major	Moderate

Princ Proc MD	Attesting MD	MD Specialty	Primary Payer	Secondary Payer
120055	120055	Orthopedic Surg	Medicare Traditional/Indemnity	Commercial/Private Preferred Provider Organization (PPO)

LOS Observed	LOS Expected (MS-DRG)	LOS Expected (DRG)	LOS Outlier	Mortality Expected (MS-DRG)	Mortality Expected (DRG)	Cost Observed	Cost Expected (MS-DRG)	Cost Expected (DRG)	Charges Observed
14	6.83	6.77	No	0.00076	0.00642	30,737	46,042	44,218	116,315

Diagnoses

(1) 1985 - secondary bone ca  
(2) 4019 - hypertension nos  
(3) 1890 - kidney ca nec  
(4) 73313 - path fx vertebrae  
(5) 99709 - nerv syst surg comp nec  
(6) 72402 - spinal stenosis-lumbar  
(7) e8768 - abn rrv-surgical px nec

Procedures

(1) 8108 - posterior lumbar fusion  
(2) 9604 - insert endotracheal tube  
(3) 8162 - fusion/refusion 2-3 vert

Complications

other complications of procedures

14 day LOS, very few diagnoses

Pediatric Indicators

none indicated

Quality Indicators

none indicated

Patient Safety Indicators

none indicated

CPM Category	Resource	Total Cost	CPM Category	Resource	Total Cost
Imaging & Diagnostics	x-ray chest		Imaging & Diagnostics	x-ray (other)	
Imaging & Diagnostics	mri body w/wo contrast		Lab	basic metabolic panel	
Lab	comp metabolic panel		Lab	calcium	
Lab	abo rh		Lab	antibody screen	
Lab	complete blood count		Lab	pt/iaqt	
Lab	pt/iaqt		Lab	hematocrit	
Lab	arterial blood gas		Lab	ortho components (screws)	
Med/Surg Supplies	mech comp devices		Pharmacy	amlodipine	
Pharmacy	cefazolin		Pharmacy	dexamethasone (systemic)	
Pharmacy	diazepam		Pharmacy	diphenhydramine	
Pharmacy	famotidine		Pharmacy	terbaryl	
Pharmacy	glycopyrrate		Pharmacy	heparin sodium	
Pharmacy	hydromorphone		Pharmacy	isoflurane	
Pharmacy	ketamine		Pharmacy	lidocaine (inj. anest)	
Pharmacy	megestrol acetate		Pharmacy	metoclopramide	
Pharmacy	midazolam		Pharmacy	morphine	
Pharmacy	neostigmine		Pharmacy	phytonadione	

resources

Done

Local intranet

## Data Quality Study

- Goal is to evaluate whether the data in the CDB is an accurate reflection of clinical practice
- Used the 5 Chicago area academic medical centers
- Studied the data quality reports as well as global reports from the CDB
- 5 variables for each organization were chosen and contact with the member determined if the variance was real, an artifact of coding or documentation or something other



## Study Summary

- UHC found the data discrepancies were mostly an effect of documentation and coding practices. In particular, they resulted from:
  - institutional emphasis on particular product lines,
  - documentation/coding of secondary diagnoses based on impact on reimbursement,
  - patient population, and
  - institutional patient safety/quality programs.

©2007 University HealthSystem Consortium 29

4. MS-DRG	a	b	c	d	e	Total
781 Other antepartum diagnoses w medical complications	90.0 4%	85.4 0%	92.0 4%	90.52 %	<b>79.3</b> <b>8%</b>	86.4 8%
782 Other antepartum diagnoses w/o medical complications	9.96 %	14.6 0%	7.96 %	9.48 %	<b>20.6</b> <b>2%</b>	13.5 2%
2. Fluid and Electro Disorders						
	a	b	c	d	e	Total
	N =	N =	<b>N =</b>	N =	N =	N =
Comorbidity	22,374	26,969	<b>15,008</b>	22,056	32,380	118,787
Fluid and electr disorders (n)	4,070	4,718	<b>2,038</b>	4,985	6,094	21,905
Percent of All Cases	18.2%	17.5%	<b>13.6%</b>	22.6%	18.8%	18.4%

7 University HealthSystem Consortium 30

## 2. Characteristics of Tobacco Use

ICD-9 Code	a	b	c	d	e
	N =	N =	N =	N =	N =
All Cases	22,374	26,969	15,008	22,056	32,380
v1582 - hx tobacco use (n)	3611	2612	152	4998	<b>82</b>
v1582 - hx tobacco use (%)	16.14%	18.53%	0.55%	11.84%	<b>0.47%</b>
3051 - tobacco use disorder (n)	2478	2367	1037	3405	<b>230</b>
3051 - tobacco use disorder (%)	11.08%	12.63%	1.53%	10.73%	<b>3.20%</b>

Clinical Data would not pick this up as it is an effect of documentation

©2007 University HealthSystem Consortium 31

## The average number of diagnoses coded per case

Diagnoses Profile			
HCO	Cases	Mean #	Max #
140088	22,374	10.072	59
140119	26,969	9.253	55
<b>140150</b>	<b>15,008</b>	<b>6.380</b>	<b>25</b>
140276	22,056	9.929	77
140281	32,380	8.230	30

This hospital does not seem to be giving  
Itself 'credit' for the severity of their patients

This will also negatively effect reimbursement

©2007 University HealthSystem Consortium 32

## Summary

- For use in performance improvement, administrative data (if proper checks are in place) can be an effective portrayal of clinical practice
- In addition, the CDB can assist a hospital in improving the accuracy of administrative data quality and accuracy

# Galaxy Data Quality Program MIT IQ Industry Symposium

16-17 July 2008

Ingenix  
United Health Analytics  
Galaxy – Shared Data Warehouse  
Laura Sebastian-Coleman  
IS Manager – Data Quality & End User Support

**INGENIX.**

## Overview

- Background / context
  - Ingenix and Galaxy
  - Galaxy's DQ program
- Value Measurement Initiative – or why we need to measure our SDLC to improve Galaxy's data quality
- Some things never change – or How Galaxy's experience applies to other situations

©2006 Ingenix, Inc.

**INGENIX.**

## Ingenix Background

- A global healthcare information company
- Founded in 1996 to develop, acquire, and integrate some of the nation's best-in-class healthcare information capabilities
- Significant and rapidly evolving portfolio of tools and services now transform data into actionable, fact-based, technology-enabled decision support
- Ranked among the top 10 providers of informatics by *Healthcare Informatics* magazine in June 2006
- Today there is an Ingenix product at work in nearly every U.S. healthcare organization.
- Ingenix is a wholly owned subsidiary of UnitedHealth Group (UHG).

©2006 Ingenix, Inc

**INGENIX.**

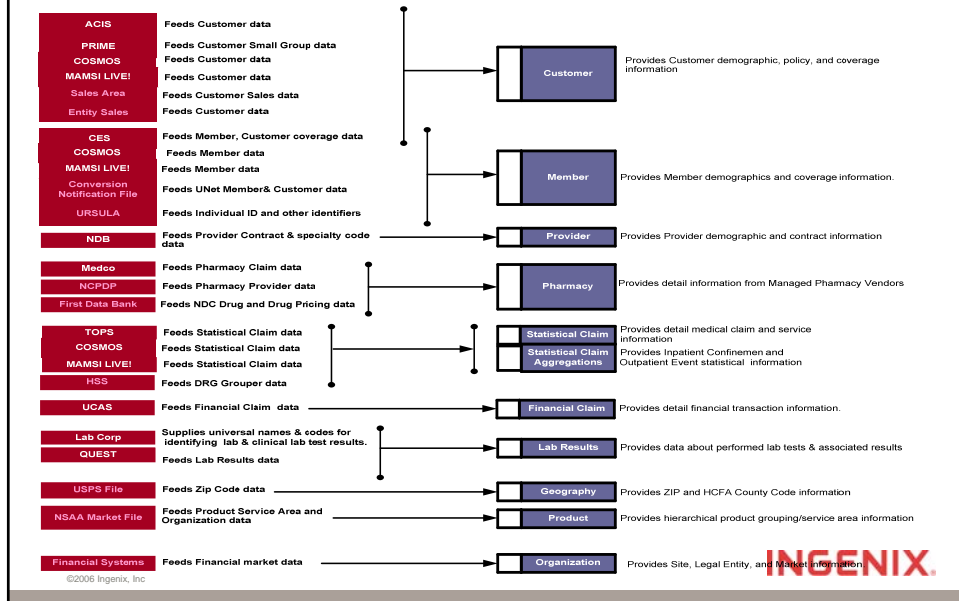
## Galaxy Overview

- Atomic Data Warehouse with transformations
- Integrates data from more than a dozen subject areas (claim, membership, customer, provider, etc.) across multiple sources
- Size
  - 350 source input files from more than 25 distinct internal and external sources (and counting)
  - 15 TB of data; 62 TB footprint
  - 3,438 attributes across 15,069 columns in more than 500 user-facing tables
  - Largest table: more than 1.5 billion rows
    - 1,782,687,382 on Claim Statistical Service as of 3/31/08
- Usage
  - Over 1,000 registered users
  - So far in 2008, averaging more than 450,000 queries / month
  - Ad hoc, scheduled queries, production extracts to applications and marts
  - Direct access to Galaxy via user-selected tools – Sagent is administratively supported

©2006 Ingenix, Inc

**INGENIX.**

## Galaxy Source Systems & Subject Areas



## Last year, I described our current situation

- Galaxy = a mature, enterprise data warehouse
- High demand for data and for organizational services
- Galaxy's DQ program also relatively mature
  - Defined metrics – baseline, semi-annual data profile
  - Automated data collection – complete with alerts, statistically established thresholds on key attributes
  - Regular reporting – post load, quarterly
  - DQ Community – user group
- UHG growing, largely through acquisitions and partnerships
- Healthcare industry changing – relation of government to health care, new products, esp. consumer driven

## And how we would meet new challenges

- Demand for more data from acquisitions
- Demand for faster integrations
- “Common Interface” approach – same structure for incoming data, regardless of source
- “Gateway” to drive consistency across sources
- DQ built into the process
- I was anticipating smooth sailing, since the pieces were all falling into place....

©2006 Ingenix, Inc

**INGENIX.**

## What we did not count on was

- New, new challenges: evolution of the user community at the same time that demand is increasing for Galaxy data.
- The down side of success
- Revenue model
- Users new to Galaxy
- New employees
- Desire for faster integrations
- New business relationships

©2006 Ingenix, Inc

**INGENIX.**

## Changes within user community

- Different users of data
- Different uses for data
- Different assumptions about the data
- Different questions about the data
- Different perceptions of the data
- These things throw open the flood gates to problems with the foundational necessity of “fitness for use” as a standard for quality.

©2006 Ingenix, Inc

**INGENIX.**

## Effort to meet demand

- More projects
- Larger projects
- More complex projects
- New expectations about projects
- New tools, each with a learning curve
- New employees
- Competition for resources, especially “knowledge workers”
- These factors put stress on the organization and especially on the software development life cycle.

©2006 Ingenix, Inc

**INGENIX.**



## Result

- A negative impact on data quality
- Actual – as defined by DQ metrics
- Perceived – as defined by end user perceptions

©2006 Ingenix, Inc

**INGENIX.**

## How to Respond? Metrics

- Launched program for new metrics in January 2008
  - Measure where the pain is
  - Project work
    - On time, on budget
  - Project quality
    - Defect tracking
  - Data delivery
    - Are new sources delivering as promised?
- Prevent new pain from emerging
- Continue standard DQ metrics – conformance to business rules, expected populations, etc.
- Continue production database metrics – availability

©2006 Ingenix, Inc

**INGENIX.**

## Sample metrics

*See handouts*

- Project delivery
- Project quality
- Data delivery

©2006 Ingenix, Inc

**INGENIX.**

## What's the take-away? Sticking to Basics

- Data in the warehouse is only as good as data in the source – needs constant vigilance
- Manufacturing model: Data as a product produced through a process – SDLC = a key part of that process
- Measure to improve (not just to measure...)
  - Baseline key SDLC processes (budget, schedule, spec)
  - Keep measures simple –
    - on time, or not?
    - How early, how late?

©2006 Ingenix, Inc

**INGENIX.**

## Can Accountants Help Reduce Medication Errors?

Scott R. Boss

Janis L. Gogan

James E. Hunton

Bentley College

### Medication Errors → Adverse Events

- In 1994 Boston Globe's Betsy Lehman dies
  - Dosage 4X higher than prescribed
- In 2006 three Indiana infants die and in 2007 Dennis Quaid's twins narrowly escape harm
  - Adult Heparin instead of Pediatric Hep-Lock
- Institute of Medicine: 400,000 preventable medication-related injuries/year

### Proposal:

#### Accountants can help clinicians

- Accountants help assure information quality in financial systems and business processes.
- **Validity**: The record describes a transaction which was authorized and actually occurred.
- **Accuracy**: The record contains a correct description of the transaction.
- **Completeness**: A record is kept of every authorized transaction.

### REA Modeling

- One way to document business processes.
  - Others: system flow charts, DFDs, BPMN, etc.
  - REA Advantage: simplicity, ease of training
- Key elements:
  - **Resources** (cash, inventory)
  - **Events** (sale, purchase, payment, etc.)
  - **Agents** (employee, computer)
  - **Relationships**/Dependencies

Figure 1. Self-service sales example

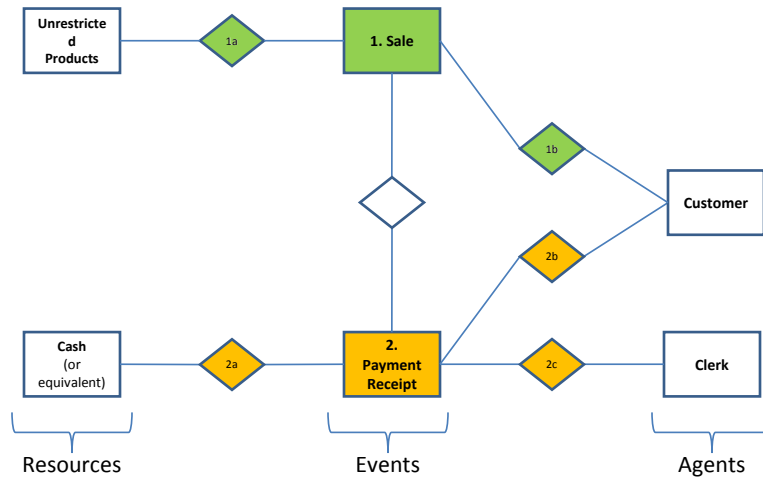


Figure 2. Restricted sale example

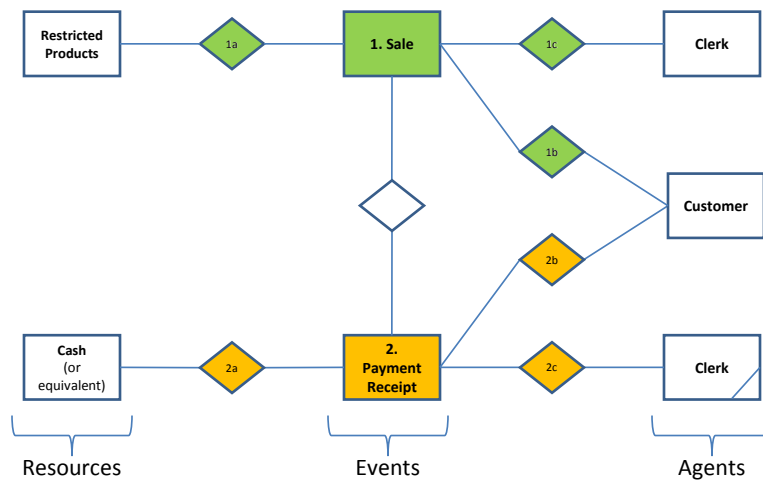
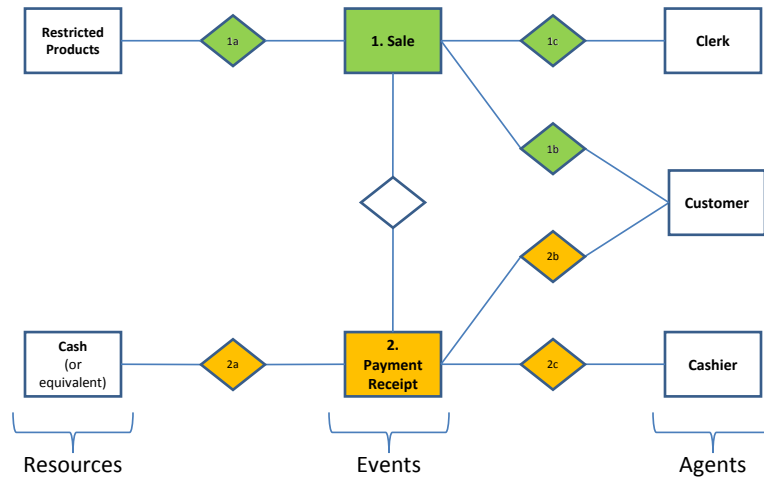


Figure 3. Manual separation of duties



### Questions auditors ask about process issues and information quality

- Have **all reasonable threats** to validity, accuracy and completeness been identified?
- Is there a **reasonable balance**?
  - **Preventive**: Ensure that processes and related data won't be invalid, inaccurate, or incomplete.
  - **Detective**: If a transaction or related data is invalid, inaccurate, or incomplete, we'll find out.
  - **Corrective**: When a problem is detected, we have a plan to fix the data and the problems it caused.

Auditor's view of  
manual and computerized controls

- Humans make more **errors** than computers, but exercise better **judgment**.
  - Where is judgment really needed/desired?
- Cost-effective control **redundancy** is desired.
  - Control is not free, but no single control is perfect.
- All controls are tested and **validated**.
  - Functional tests: unit, integration, end-to-end
  - Other tests: capacity/volume, security, usability
  - **Compliance must be verified.**

A clinical scenario:  
Pediatric intensive care

- *Infant admitted for treatment of a staph infection with IV antibiotics. MD prescribes pediatric Hep-Lock but the baby gets adult Heparin.*
- Interlinked processes and actors:
  - stock medications: pharmacist, pharmacy tech
  - prescribe medications: doctor, computer
  - administer medications: nurse or doctor
  - check patient condition: nurse, doctor, computer, lab
  - decide appropriate corrective clinical action: doctor

Figure 4 Stock Cabinets



Figure 5 Prescribe Medication

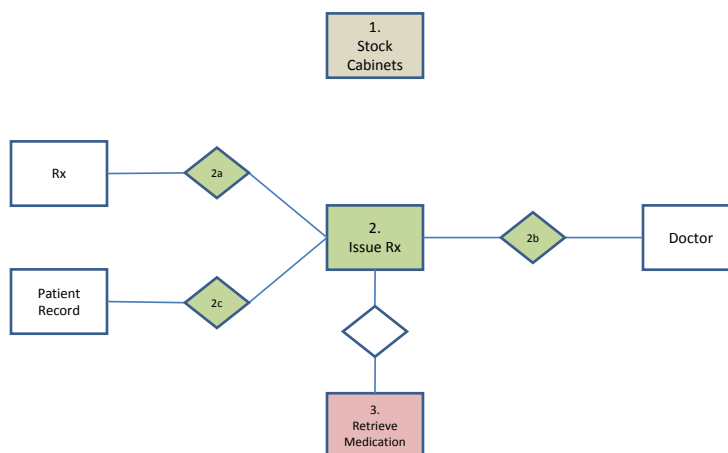




Figure 6 Retrieve and administer medication

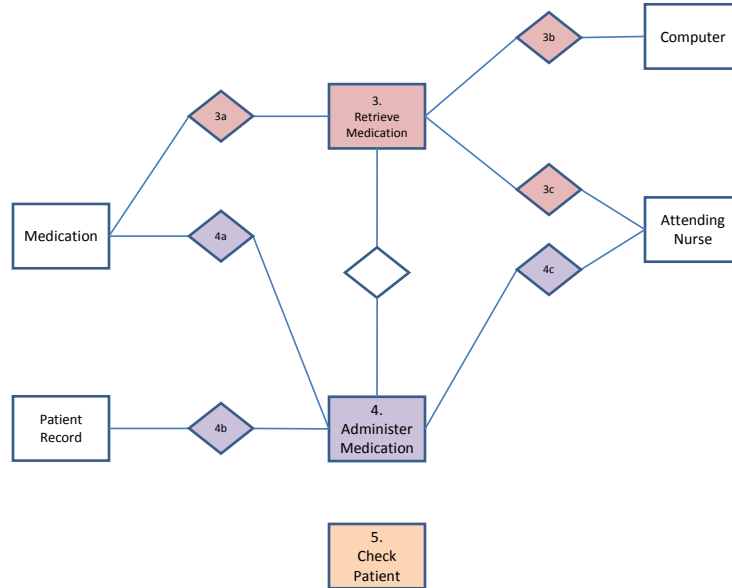


Figure 7, Further tests and treatment

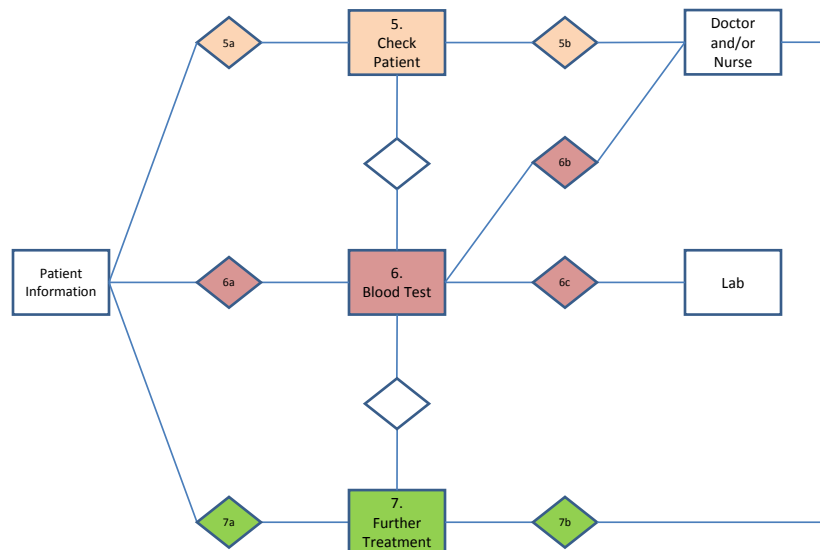


Figure 8 Stock cabinets: segregation of duties

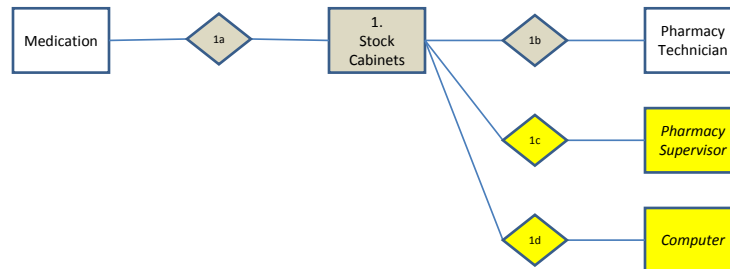


Figure 9 ePrescribing with control

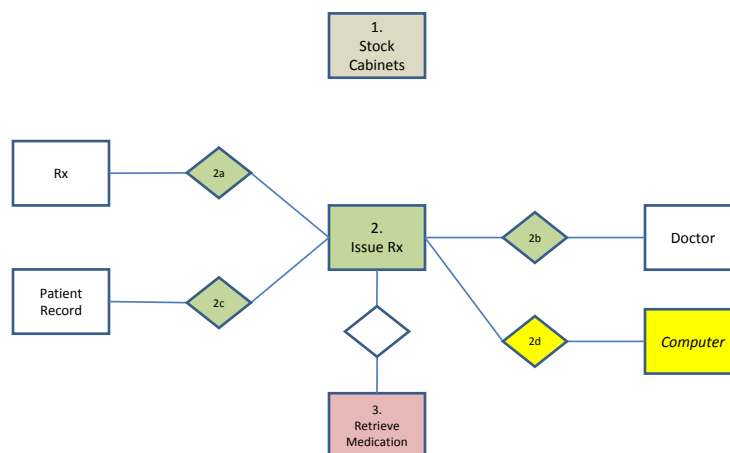
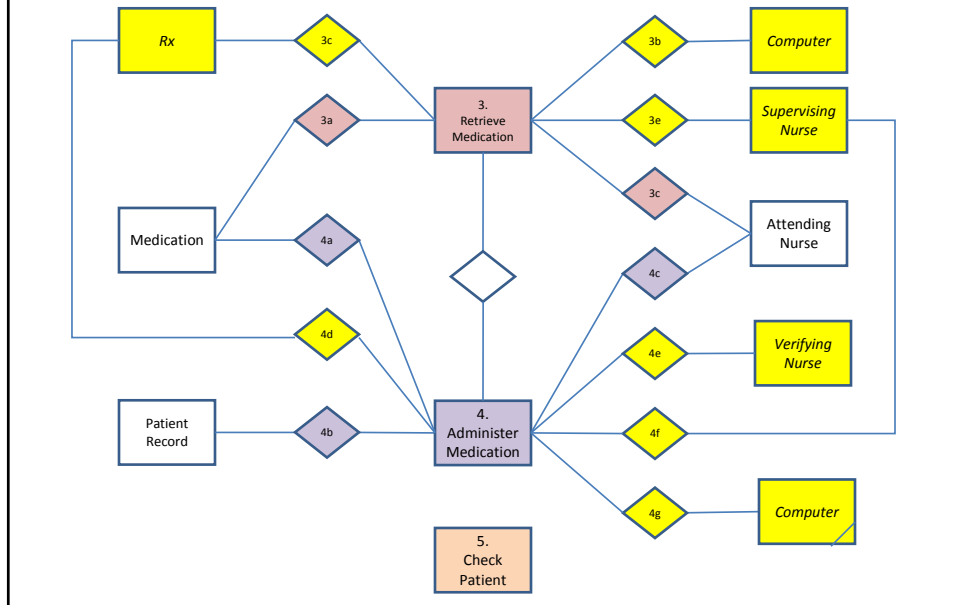


Figure 10 Retrieve and administer medication

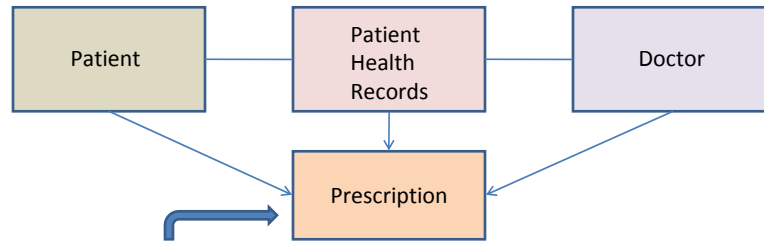


## Observations

REA models:

- do reveal opportunities to add controls that improve **validity, accuracy, and completeness** in clinical processes.
- do reveal opportunities to improve the mix of **preventive, detective, and corrective** controls.
- do *not* reveal an *optimal* mix of controls or optimal level of *control redundancy*.
  - Other techniques are used by auditors for this.

## eXtensible Markup Language (XML)



Example XML tags that could be attached to each prescription

Drug Interaction Check Performed?	By Whom?	Check Date?
Prescription Filled?	By Whom?	Filled Date?
Prescription Checked?	By Whom?	Checked Date?
Drug Transferred to Nurse?	By Whom?	Transfer Date?
Nurse Received Drug?	By Whom?	Received Date?
Drug Administered to Patient?	By Whom?	Administration Date?
Drug Administration Verified?	By Whom?	Verified Date?

# Using Data Quality Methods at the Federal Railroad Administration

Improving the Archiving and  
Retrieval of Safety Information



## Agenda

- Introduction
- Background
- The Context of the FRA Enterprise
- Data Quality
- Information As Product
- Information Product Maps
- Conclusions

5/27/2008

2

Federal Railroad Administration



# Introduction

## ■ Mission

- *Ensure the safety and efficiency of passenger and rail freight services in the United States.*

## ■ Task

- *500 Inspectors*
- *Inspect 230,000 miles track, 1.2 million freight cars, 20,000 locomotives, and 89,000 track miles of signal and train controls.*

## ■ Challenge

- *Optimize efficiency while guaranteeing quality*

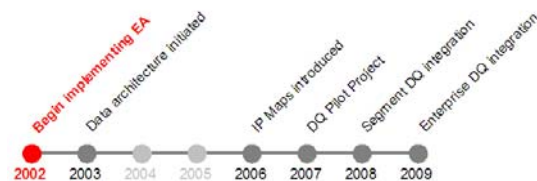
5/27/2008

3

Federal Railroad Administration



# Timeline 2002



## ■ FRA began implementing EA In 2002.

- *Eight functional areas*
- *Three view levels*
- *First cut, since refined.*

Federal Railroad Administration Enterprise Architecture							
EA Artifact Inventory		Reference Documents		EA & CPIC Meetings		Contact Us	
Strategy & Goals	Organization & Responsibility	Business Architecture	Enterprise Architecture	Technology Infrastructure	Capital Planning	Project Management	Security & Privacy
Strategic View	Mission & Vision	Organization Structure	Business Areas	EA Program	Systems	Investment Process	Project Management Plans
Operational View	Goals	IT Governance	Lines of Business	EA Framework	Networks	Current Investment Portfolio	Project Information
Tactical View	E-Gov Strategy	Training & Development	FRA Business Processes	EA Deliverables	Technical Reference Model	Estimates 53 & 300 (E-CPIC)	Project Planning
							Certification and Accreditation

5/27/2008

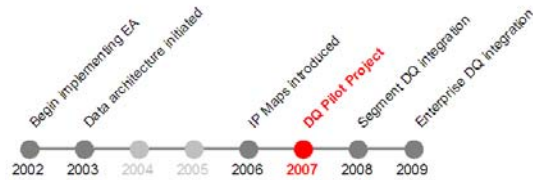
4

Federal Railroad Administration





## Timeline 2007



- Data Quality Pilot Project conducted in Fall 2007.
- Focus of project on eight essential data attributes shared between two LOBs.
  - Office of Chief Counsel (RCC)
  - Office of Safety (RRS).

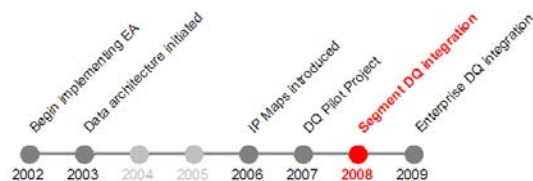
5/27/2008

7

Federal Railroad Administration



## Timeline 2008



- Segment-by-segment analysis of FRA enterprise throughout 2008.
  - New Initiatives
    - Disaggregation and cataloging of institutional reports and data calls.
    - Coordination, cooperation, and data sharing between EA and Security data collection teams.

5/27/2008

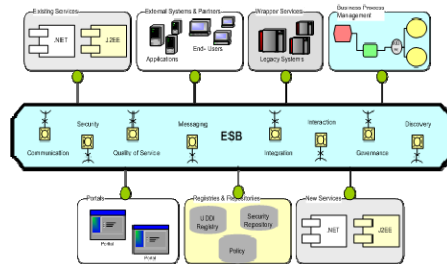
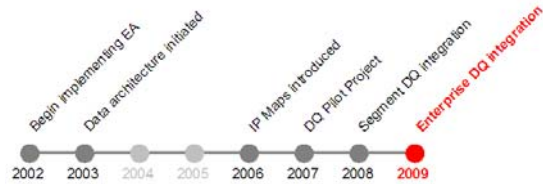
8

Federal Railroad Administration





## Timeline 2009



### Database consolidation and normalization

- Rationalized data structure as keystone of Data Quality Program and Service Oriented Architecture (SOA).

5/27/2008

9

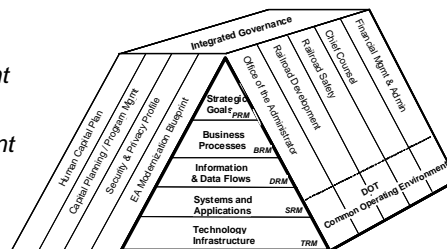
Federal Railroad Administration



## The Context of the Enterprise

### ■ Five Lines of Business (LOB)

- Office of the Administrator
- Office of Safety
- Office of Railroad Development
- Office of the Chief Counsel
- Office of Financial Management and Administration



5/27/2008

10

Federal Railroad Administration



## Data Quality

# ***“Fitness for Use”***

5/27/2008

11

Federal Railroad Administration



## Data Quality (cont.)

Information Quality Category	Information Quality Dimensions
Intrinsic Information Quality	Accuracy, Objectivity, Believability, Reputation
Accessibility Information Quality	Accessibility, Access Security
Contextual Information Quality	Relevancy, Value-Added, Timeliness, Completeness, Amount of Information
Representational Information Quality	Interpretability, ease of understanding, concise representation, ease of manipulation

5/27/2008

12

Federal Railroad Administration



## Information As Product

- Product vs. By-Product
  - *Consumer vs. Producer measures of quality.*
- Formal recognition of the criticality of data quality to mission.
  - *Organizational structures, positions, and processes to manage data quality.*

5/27/2008

13

Federal Railroad Administration



## Information As Product (cont.)

	Information as Product	Information as By-product
<b>What is managed?</b>	Information; information product life cycle	Hardware and software; systems life cycle
<b>How is it managed?</b>	Integrated, cross-functional approach that encompasses information collectors, custodians, and consumers	Integrate stove-pipe systems; control individual components; control costs
<b>Why manage it?</b>	Deliver high-quality information to consumers	Implement high-quality hardware and software system
<b>What is success?</b>	Deliver high-quality information continuously over the product life cycle; no GIGO (garbage in, garbage out)	The system works; no bugs
<b>Who manages it?</b>	Chief information officer (CIO); information product manager	CIO; information technology director and database administrators

5/27/2008

14

Federal Railroad Administration



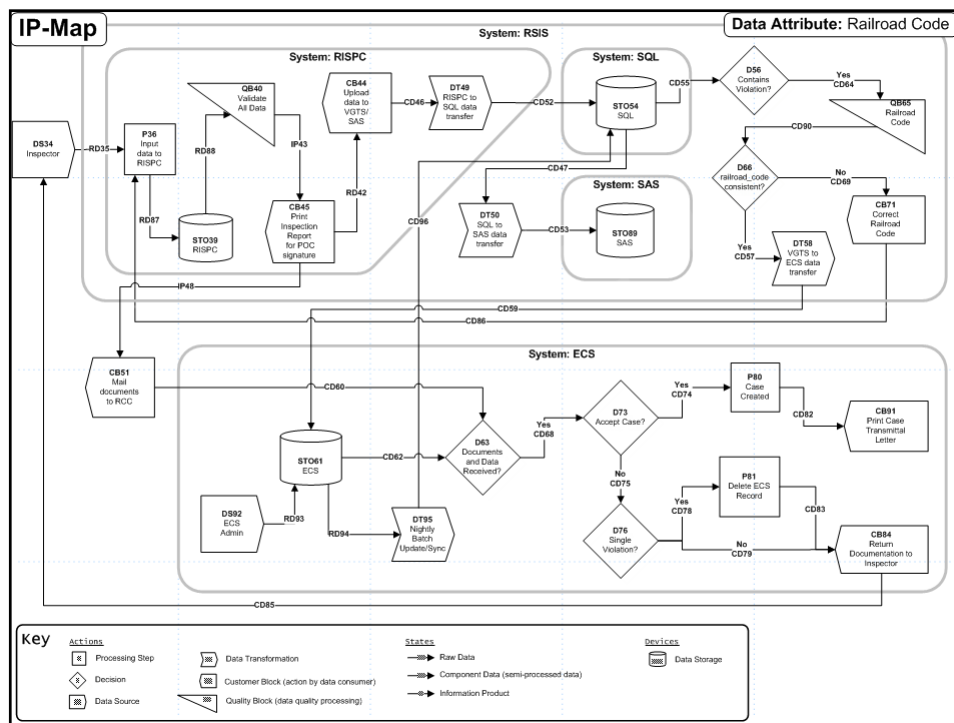
## Information Product Maps (IP Maps)

- Designed to graphically describe data workflows and transformations.
  - Can represent single data elements or logical groupings of multiple data elements.
- Identifies source and destinations of Information Products.
  - A “collection of data element instances that meet the specified requirements of a data consumer.” (Wang)

5/27/2008

15

Federal Railroad Administration



## Conclusions

- ***Be ready.***
  - *It should be anticipated that data quality efforts will recommend change and possibly even Change.*
- ***Be clear and purposeful.***
  - *Know what you're doing and why... and be able to articulate this in practical terms that managers and workers will understand and appreciate.*

5/27/2008

17

Federal Railroad Administration



## Questions

5/27/2008

18

Federal Railroad Administration



*Thank You!*

**Scott Bernard**

Chief Architect  
Federal Railroad Administration  
U.S. Department of Transportation  
202-493-6125  
[scott.bernard@dot.gov](mailto:scott.bernard@dot.gov)

**Mark Trimble**

Data Architect  
Federal Railroad Administration  
U.S. Department of Transportation  
202-493-1343  
[mark.trimble@dot.gov](mailto:mark.trimble@dot.gov)

5/27/2008

19

Federal Railroad Administration





## **Embedding Information Quality in the Lockheed Martin Enterprise Architecture Framework: An IPMAP Approach**

**Edwin F. Nassiff**

Director, Architecture - Enterprise Business Services  
Lockheed Martin Corporation

**Paul B. Pierson**

Principal Computing Systems Architect - Enterprise Business Services  
Lockheed Martin Corporation

**John P. Slone, Ph.D.**

Sr. Manager, Enterprise Architecture - Enterprise Business Services  
Lockheed Martin Corporation

### **Agenda**



- ***About Lockheed Martin***
- ***Strategic Focus for IT***
- ***Lockheed Martin Enterprise Architecture Framework***
- ***Information Quality and Enterprise Architecture***
- ***Information Product Maps in the Context of Enterprise Architecture***
- ***Early Lessons and Next Steps***

## The Men and Women of Lockheed Martin

- **140,000 Employees**
- **70,000 Scientists and Engineers**
  - **25,000 IT Professionals**
- **Operations in 1,000 Facilities, 500 Cities, 46 States and 63 Countries**

**Partners to Help Customers Meet Their Defining Moments**

Corporate Overview 3

## Our Core Markets

**Defense & Intelligence**

**Civil Government**

**IT**

**Homeland Security**

**IT: Common Denominator**

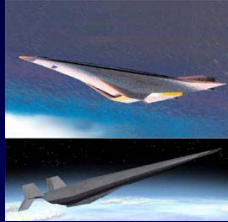
4



## Redefining What Is Possible



Hypersonics



Biometrics



Return of Crew Space Exploration



Persistent Surveillance



Information Fusion



Unmanned and Autonomous Systems



A Passion for Invention

5

## IT Strategy



Industry Leading Security



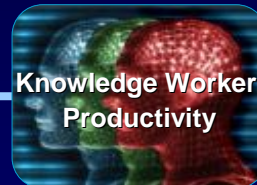
Enterprise Architecture & Business Process Optimization



Innovation



Knowledge Worker Productivity



World Class Infrastructure



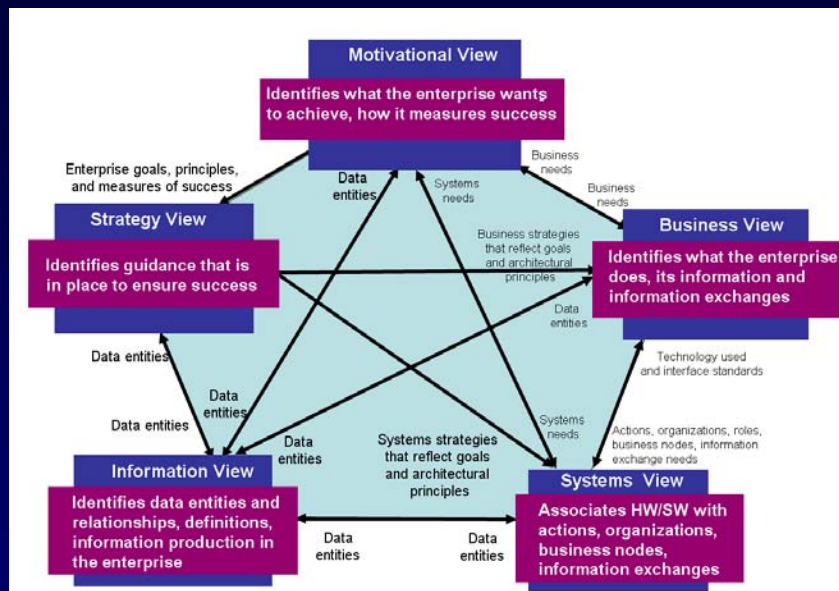
6

## Lockheed Martin EA Framework

- *The LM Enterprise Architecture Framework establishes a common set of EA practices for the corporation*
- *The common practice of EA facilitates alignment of people, processes, and technology, enables business agility, and acts as a catalyst to derive maximum value from IT investments*
- *These practices are also used to facilitate reduction of unnecessarily redundant efforts, and create and reuse architectures and their descriptions across the corporation*

7

## LM EA Framework



8





## Information View Artifacts

<b>Logical Data Model (OV-7)</b>	Represents objects about which the enterprise records information. Is a fully attributed, keyed, normalized entity relationship model.
<b>Information Production Map</b>	Illustrates how data and information are produced in the enterprise.
<b>Information Exchange Matrix (OV-3)</b>	Decomposes each needline from the Business Node Connection Model into its constituent information exchanges.
<b>Systems Data Exchange Matrix (SV-6)</b>	Specifies characteristics of data exchanged between systems (the automated information exchanges from the Business Information Exchange Matrix.)
<b>Integrated Dictionary (AV-2)</b>	Central source for all definitions and metadata of terms used in the architecture description, represents the underlying architecture database.

**Combines Several Standard DoDAF Products with IPMAP**

9





## Relationship between IPMAP Objects and LM EA Objects

<b>IPMap Object</b>	<b>Symbol</b>	<b>LM EA Equivalent</b>
<b>Data Source/Data Vendor/Point of Origin Block</b>		<b>Business node</b>
<b>Processing Block</b>		<b>System action or human action</b>
<b>Data Storage Block</b>		<b>A System</b>
<b>Quality/Evaluation/Check Block</b>		<b>System Action or human action</b>

10

## Relationship between IPMAP Objects and LM EA Objects (Cont.)

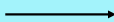


IPMap Object	Symbol	LEAF Equivalent
Data Sink/ Consumer/ Point of Destination Block		Business node
Decision Block		A special type of Processing Block
Information System Boundary		Sending and Receiving Business or System Node
Organizational Boundary		Sending and Receiving Business node

11

## Relationship between IPMAP Objects and LM EA Objects (Cont.)



IPMap Object	Symbol	LEAF Equivalent
Arrows		Information exchanges
Data Quality Attributes		Attributes in Information Exchange Matrices

12

## IPMAP in LM EA Context



- *EA Framework*
  - *Integrated the IPMAP methods and descriptions into the LM EA Framework*
  - *Developed a set of criteria for selecting data elements for IPMAP analysis*
- *Practical application*
  - *Targeting the IPMAP approach as a “bottom-up” entry point into our LEAF methodology*
    - *identify a problematic data element*
    - *identify the business processes that create, maintain and use the data element*
    - *identify the organizations that perform those processes, and the software applications used to automate the processes*
    - *illuminate a “micro-view” of the enterprise with “laser focus” on specific problematic data elements*
    - *By completing EA models for those aspects of the enterprise, and integrating with those models created from more of a “top-down” perspective, we ensure that the solutions were compatible with the rest of the enterprise*
  - *Integrated the IPMAP descriptions into our integrated architecture framework*

13

## Selection Criteria for Targeted Data Elements



- *Critical to the Organization*
- *Recognized Pain Point*
- *Dollar Impact at or Above Minimum Threshold*
- *Illustrative Power re IP Mapping*
- *Illustrative re EA Products*
- *Practical to Model*
- *Practical to Implement*
- *Owner Identified*
- *Commitment by Collector/ Creator, Custodian, Consumer*

14

## Early Lessons and Next Steps



- **Early Lessons**
  - *Widely accepted with enthusiasm as a concept*
  - *Getting commitment to act is a challenge*
- **Next Steps**
  - *Establish the adoption of our EA methodology*
  - *Work with practitioners to transition problematic data elements into EA via the IPMAP approach*

15

## Acknowledgements



- *We gratefully acknowledge the support and contributions made by the following individuals:*
  - *Kathie Sowell, for her support in integrating the IPMAP concept into our LEAF Architecture Description Language*
  - *Dr. Richard Wang, for his vision in the application of IPMAP within an EA Framework*

16



The MIT 2008 Information Quality Industry Symposium



## Patterns in Data Quality

A Method for Organizing  
Enterprise Data Quality (Web) Services  
in Service Oriented Architectures

Michael Overturf – VP of Strategy

Navin Sharma – Dir. of Product Management



The MIT 2008 Information Quality Industry Symposium



### Patterns in Data Quality

- A methodical approach to structuring rules for data user satisfaction
- Patterns provide simplification
- Patterns focus measurement
- System architects are the primary beneficiaries of simplification



The MIT 2008 Information Quality Industry Symposium



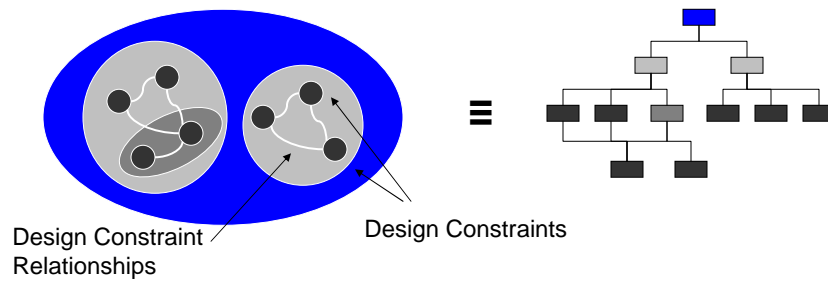
## Background

Structured constraint patterns developed by Christopher Alexander (Architect and Mathematician) in 1964 in '*Notes on the Synthesis of Form*'

*M* – Set of Design Constraints

*L* – Set of Design Constraint Relationships

$G(M,L)$  – Linear Graph of Design Constraints

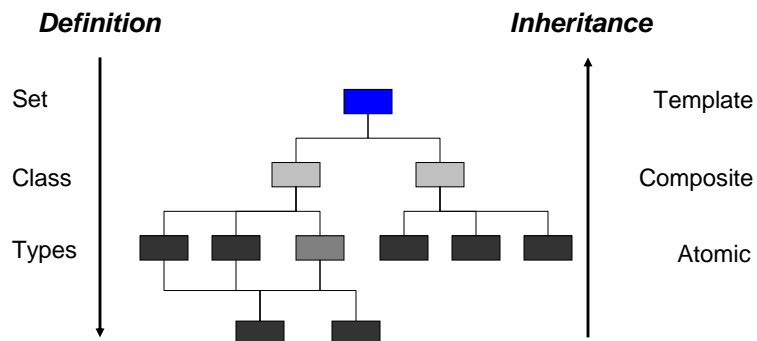


The MIT 2008 Information Quality Industry Symposium



## Constraint Tree Dynamics

Definition-Inheritance dynamic describes how we get from DQ Constraint Set to DQ Service







The MIT 2008 Information Quality Industry Symposium



## Constraint Patterns in Data Quality

*Data quality constraints  
define  
the form of data quality rules*

<u>Metadata</u>	<u>Platform</u>	<u>Access</u>	<u>Modality</u>	<u>Input</u>	<u>Output</u>
Consistency	Operating System	Web Services/SOAP	Transactional	Type	Type
Completeness	SaaS	C/S, API	Batch	Binding	Binding
Accuracy		Security	Microbatch	Source	Sink
Uniqueness/ Singularity			Multi-modal (Any)		

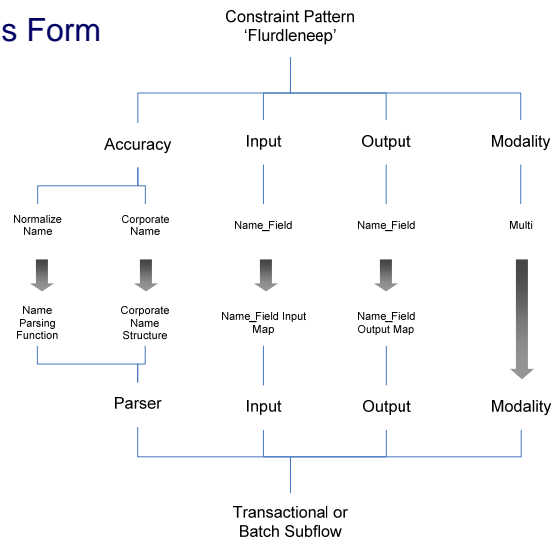


The MIT 2008 Information Quality Industry Symposium



## Pattern yields Form

*A simple example*



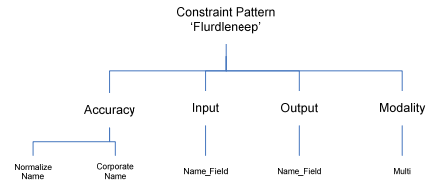


The MIT 2008 Information Quality Industry Symposium

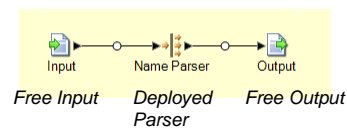


## Pattern yields Form

*A simple example*



## Subflow



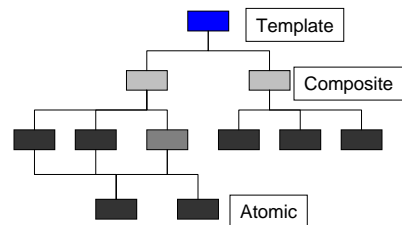
The MIT 2008 Information Quality Industry Symposium

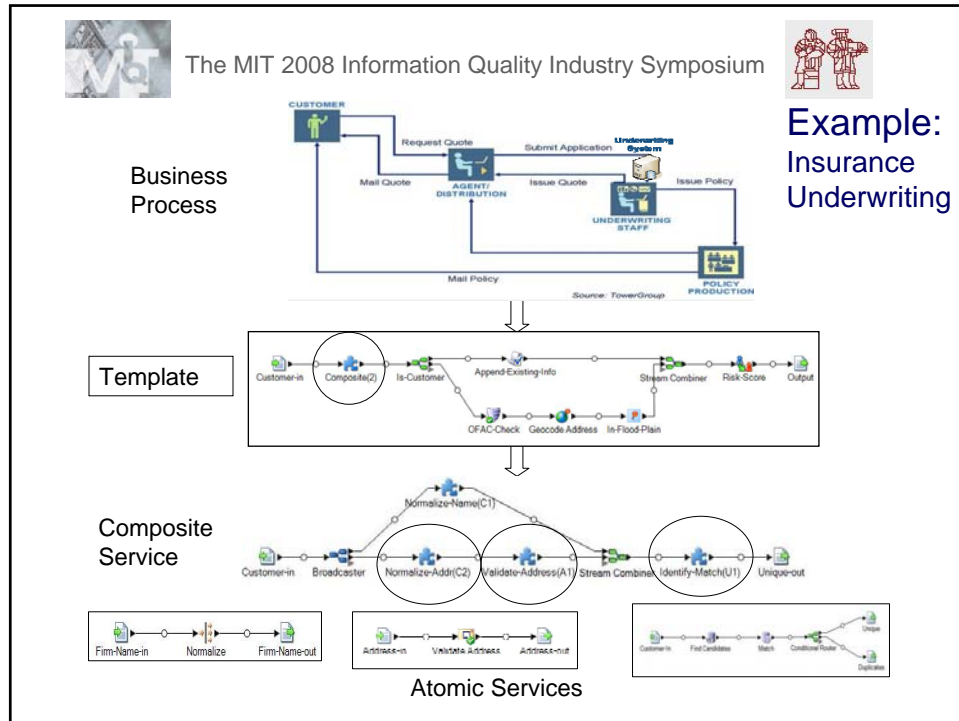


## Quality as a Service (QaaS)

Patterns naturally translate into services (in a SOA context):

- *Atomic* services that are fine grained to provide a monotonic function;
- *Composite* services that encompass two or more atomic services;
- *Templates* make up one or many composite and atomic services
  - by specific data quality metrics (consistency, uniqueness, etc.)
  - industry specific business processes





The MIT 2008 Information Quality Industry Symposium

## Web Services implement Pattern Constraints

- Widely adopted and accepted within enterprises and vendor communities
- Software community agreement on standards for distribution
  - WS-I and W(3)C standards,
  - WSDL (Web Services Description Language)
- Principles encourage and enable –
  - Ease of integration (loose coupling)
  - Re-use
  - Ease of management
  - Agility
  - Active use
- Satisfies modality constraints (batch, transactional)

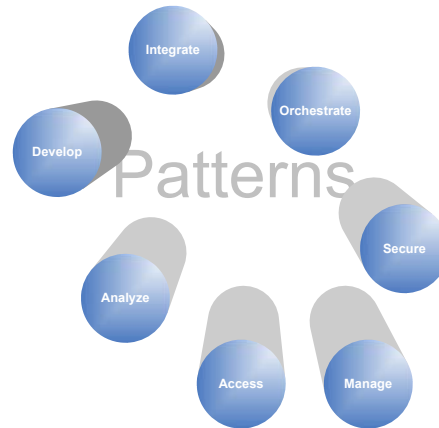


The MIT 2008 Information Quality Industry Symposium



## Laying the foundation

- **Create** – a Composition Interface that allows data stewards and enterprise architects to use and define templates, composite and atomic data quality services
- **Publish** – for distribution
- **Integrate** – via standards based interfaces (WSDL, etc.)
- **Manage** – for governance



## Consistent Measurement of Data Quality Results



The MIT 2008 Information Quality Industry Symposium



## Summary

- A methodical approach to structuring rules for data user satisfaction
- Patterns provide simplification
- Patterns provide a structure for measurement of data quality
- System architects manage data quality using standard Web Service Management Lifecycle



The MIT 2008 Information Quality Industry Symposium



## Contact Information

- Michael Overturf
  - Email: [Michael\\_Overturf@g1.com](mailto:Michael_Overturf@g1.com)
  - Phone: 413-695-5500
- Navin Sharma
  - Email: [Navin\\_Sharma@g1.com](mailto:Navin_Sharma@g1.com)
  - Phone: 240-447-6801



The MIT 2008 Information Quality Industry Symposium



## Rapid Corporate Growth and Information

Steve Sarsfield, Trillium Software



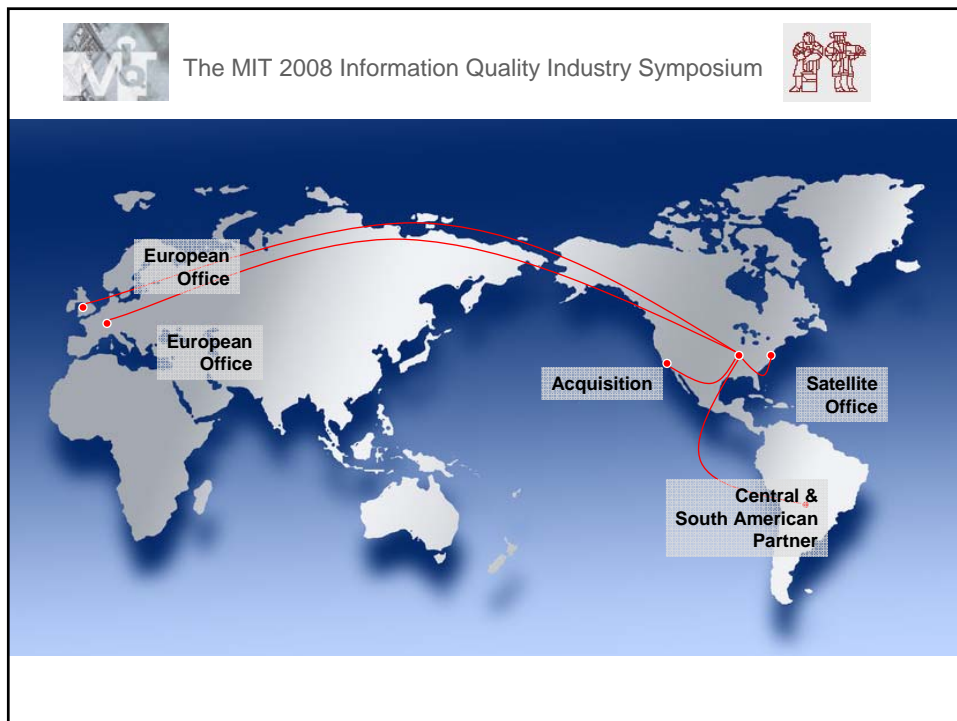
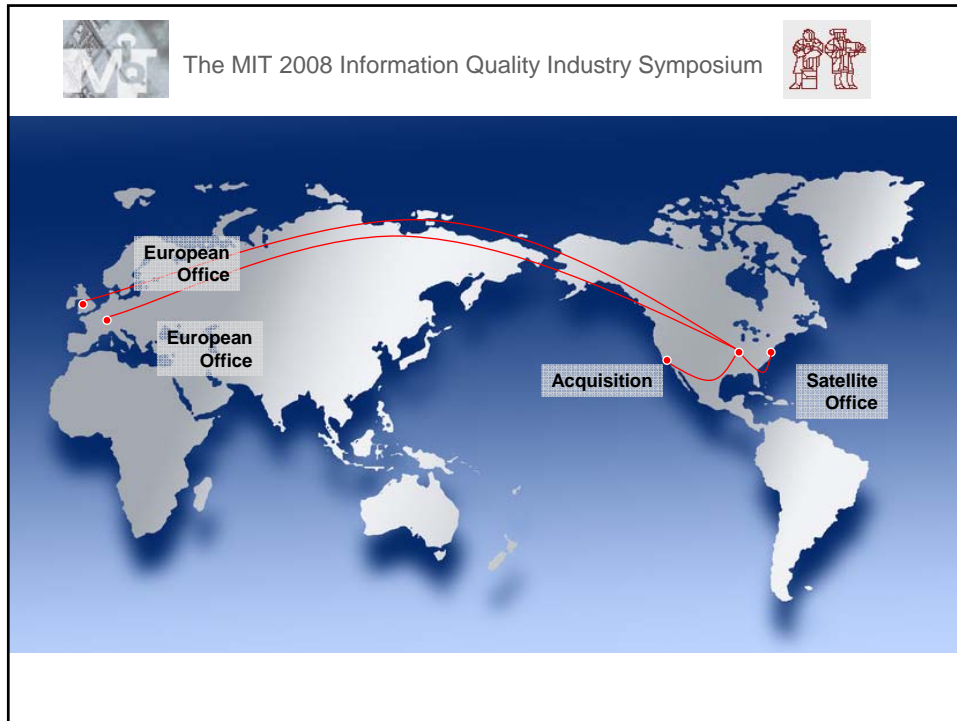
The MIT 2008 Information Quality Industry Symposium



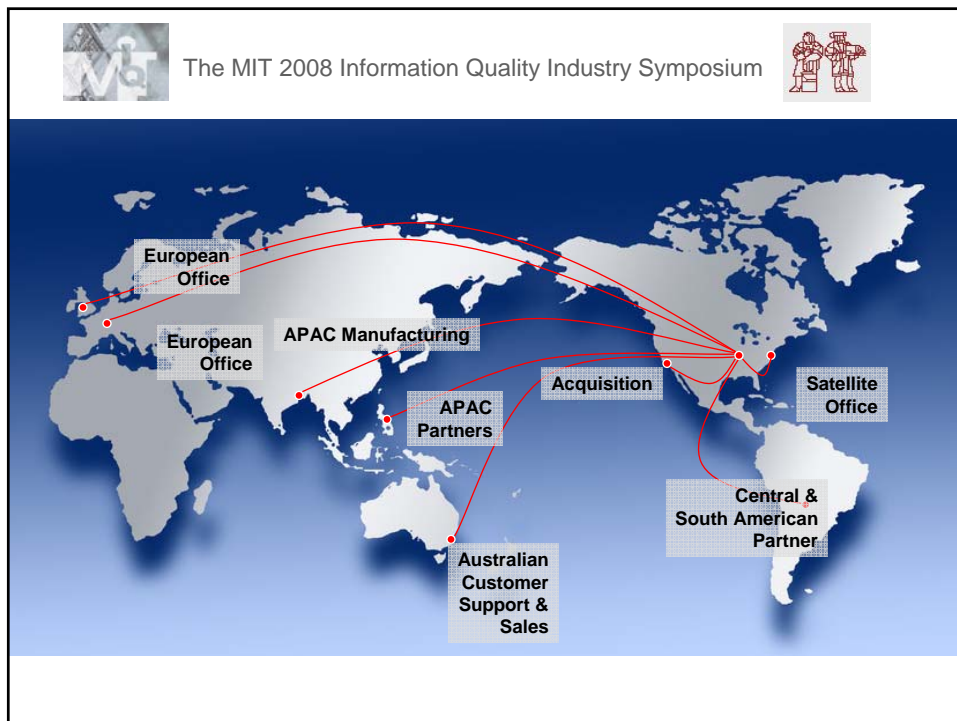
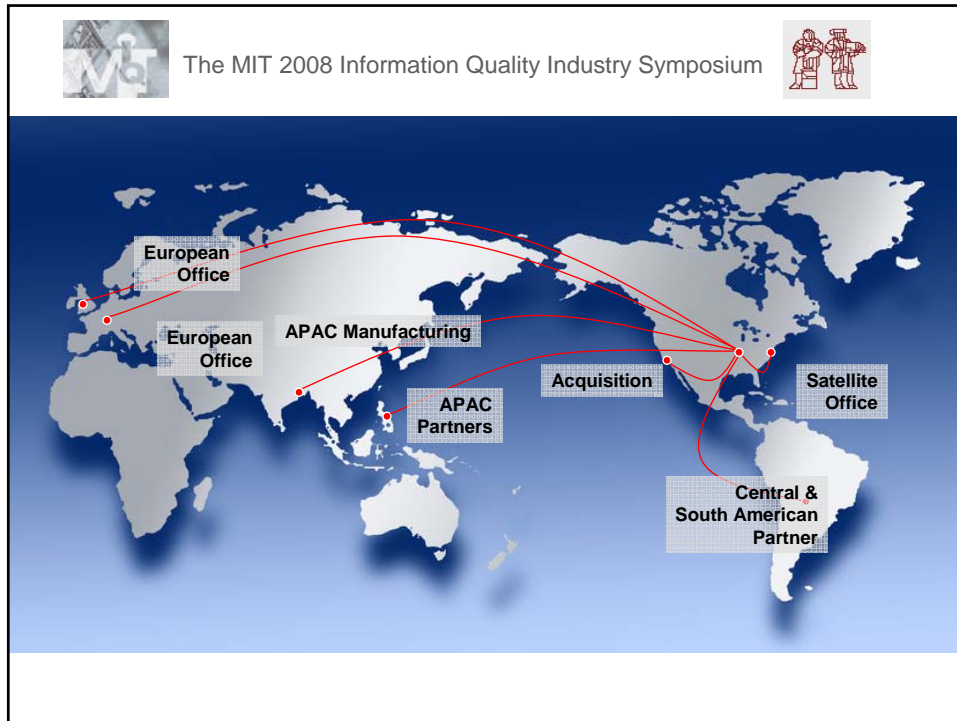
### Agenda

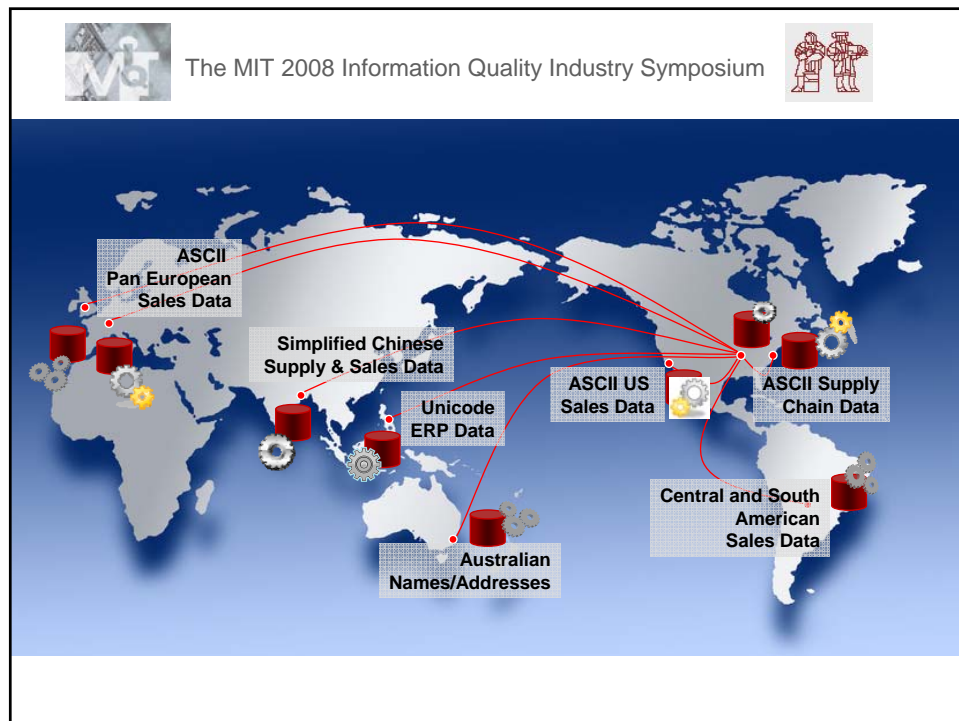
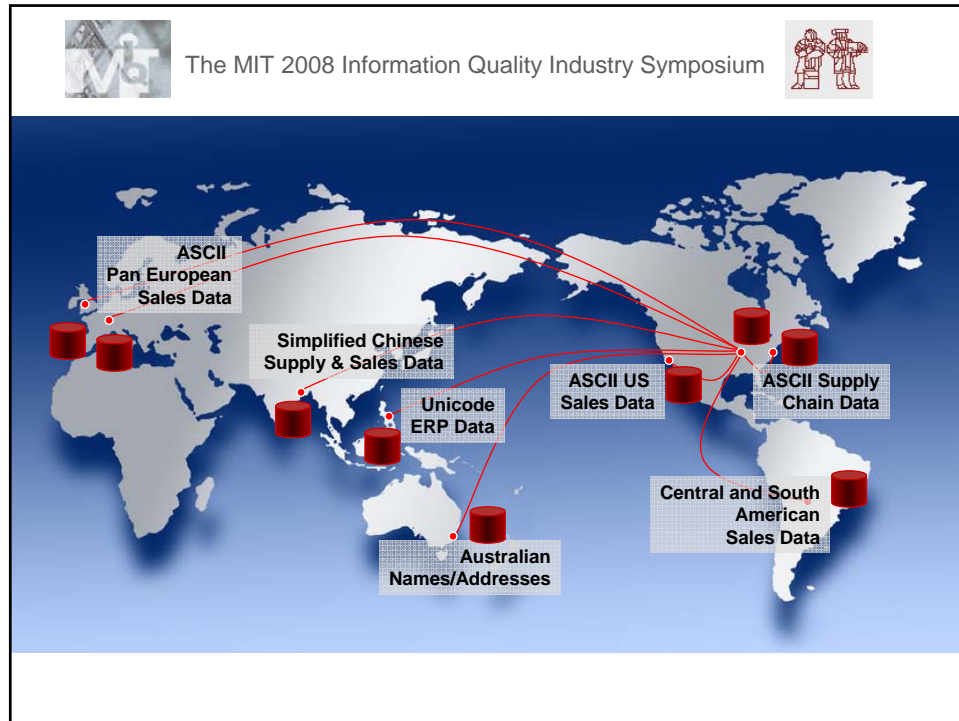
- How Companies Grow
- The Data Components of Company Value
- Effects of Rogue Data Quality Processes
- Sorting Out a Large Company's Problems
  - Cleansing and Matching
  - Domain Specific Knowledge
  - Platform Unification

























The MIT 2008 Information Quality Industry Symposium



<ul style="list-style-type: none"> <li> <b>ASCII Pan European Sales Data</b></li> <li> <b>Simplified Chinese Supply &amp; Sales Data</b></li> <li> <b>Unicode ERP Data</b></li> <li> <b>Australian Names/Addresses</b></li> <li> <b>ASCII US Sales Data</b></li> <li> <b>HQ</b></li> <li> <b>ASCII Supply Chain Data</b></li> <li> <b>Central and South American Sales Data</b></li> </ul>	<p><b>Vision Blocking Factors</b></p> <ul style="list-style-type: none"> <li>Typos and Duplicates</li> <li>Lack of standards</li> <li>Competing Information Quality Processes</li> <li>Code pages                         <ul style="list-style-type: none"> <li>ASCII</li> <li>Unicode</li> <li>EBCDIC</li> </ul> </li> <li>Platforms                         <ul style="list-style-type: none"> <li>SAP</li> <li>Oracle</li> <li>Siebel</li> <li>Tibco</li> <li>SalesForce</li> <li>Etc</li> </ul> </li> <li>Operating Sys.</li> <li>Language</li> <li>Local Nuances</li> <li>Data Age &amp; Reliability</li> <li>Unknown Data                         <ul style="list-style-type: none"> <li>M&amp;A</li> <li>Suppliers</li> <li>Partner</li> </ul> </li> </ul>	<p><b>THE BASICS</b></p> <p>How much did we sell yesterday?</p> <p>What's the sales pipeline?</p> <p>What do we have in inventory worldwide?</p> <p>Can I trust these results?</p> <p><b>OPPORTUNITY LOST</b></p> <p>Supply Chain/Inventory Problems?</p> <p>Can we reach our customers effectively?</p> <p>Are we paying too much to suppliers?</p> <p>Are we making the right business decisions?</p> <p>Are users avoiding new systems because of data?</p>
--	--	---



The MIT 2008 Information Quality Industry Symposium



## How Company Value is Measured

• Number of Customers	= Customer Data
• Hard Assets	= ERP and Supply Chain Data
• Number of Valued Employees	= Knowledge Management Data
• Sales Channels	= Supplier and Partner Data



The MIT 2008 Information Quality Industry Symposium



## Differences in Data Quality Processes

Ivan Madar  
75 Calle del Norte  
Sedona, AZ 86336



Certain Solutions can't understand  
this street address.  
No BLVD, ST, RD, AVE



The MIT 2008 Information Quality Industry Symposium



## Differences in Data Quality Processes

Debra Shaw  
203 Old Meadow Drive  
Leave at front door  
Greensburg, PA 15601



Certain solutions can't handle  
delivery info  
intermingled with name and address.

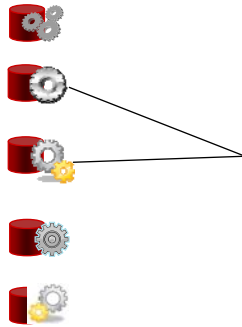


The MIT 2008 Information Quality Industry Symposium



## Differences in Data Quality Processes

Gary Wright  
C/O Allstate Insurance  
1466 S Potomac St  
Hagerstown, MD 21740



Certain Solutions  
Can't handle the C/O

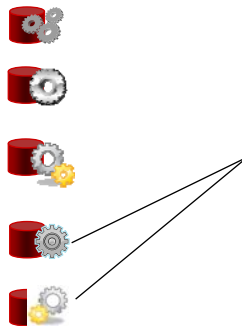


The MIT 2008 Information Quality Industry Symposium



## Differences in Data Quality Processes

Marilyn E Vogt  
N105W21040 Parkland  
Colgate, WI 53017



Certain Solutions  
Can't handle a number and letter-rich  
address such as this.



## The MIT 2008 Information Quality Industry Symposium



### Differences in Data Quality Processes

Ivan Madar  
75 Calle del Norte  
Sedona, AZ 86336

Debra Shaw  
203 Old Meadow Drive  
Leave at front door  
Greensburg, PA 15601

Gary Wright  
C/O Allstate Insurance  
1466 S Potomac St  
Hagerstown, MD 21740

Marilyn E Vogt  
N105W21040 Parkland  
Colgate, WI 53017



Ivan Madar  
75 Calle del Norte  
Sedona, AZ 86336

Debra Shaw  
203 Old Meadow Drive  
Leave at front door  
Greensburg, PA 15601

Gary Wright  
C/O Allstate Insurance  
1466 S Potomac St  
Hagerstown, MD 21740

Marilyn E Vogt  
N105W21040 Parkland  
Colgate, WI 53017

Debra Shaw  
203 Old Meadow Drive  
Greensburg, PA 15601

Gary Wright  
Allstate Insurance  
1466 S Potomac St  
Hagerstown, MD 21740

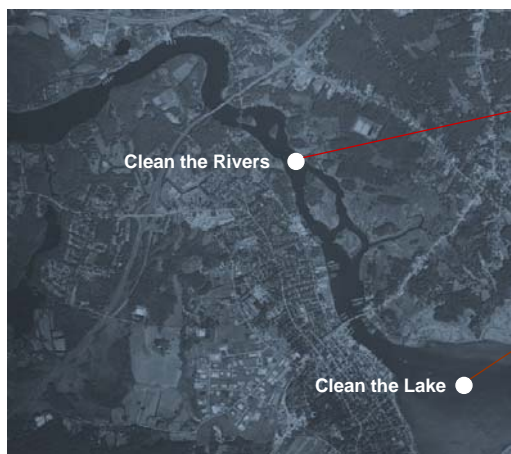
Marilyn E Vogt  
N105 W21040 Parkland  
Colgate, WI 53017



## The MIT 2008 Information Quality Industry Symposium



### Dirty Data Strategy

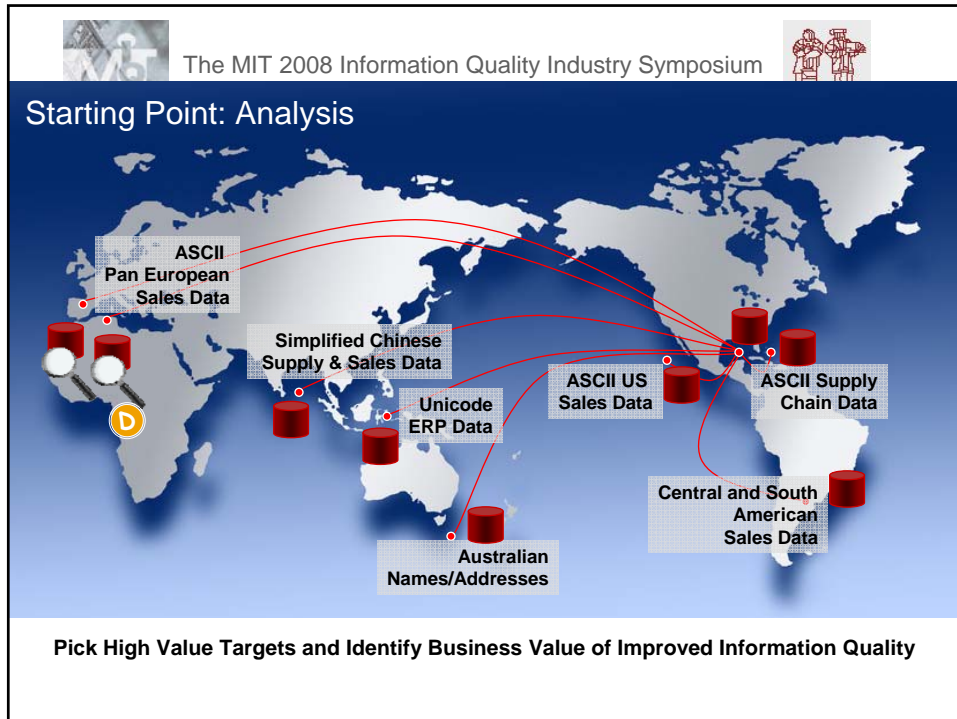


#### Clean the Rivers

- Real-time cleansing of incoming data

#### Clean the Lake

- Batch cleansing of existing data



The MIT 2008 Information Quality Industry Symposium

### Cleansing Is Key to Matching

<p><b>Original Record 1</b></p> <p>Name: Peggy Smith  Address: 345 6<sup>th</sup> Ave  City: NY  State: NY  Zip: 01012  Country:</p>	<p><b>Original Record 2</b></p> <p>Name: Margaret Smith  Address: 345 Avenue of the Americas  City: Manhattan  State: NY  Zip: 1012  Country: USA</p>
<p><b>Standardized Record 1</b></p> <p>Root First Name: Margaret  Last Name: Smith  Address: 345 Ave of the Americas  City: New York  State: NY  Post Code: 10012  Country: USA</p>	<p><b>Standardized Record 2</b></p> <p>Root First Name: Margaret  Last Name: Smith  Address: 345 Ave of the Americas  City: New York  State: NY  Post Code: 10012  Country: USA</p>





The MIT 2008 Information Quality Industry Symposium



## Cleanse: Domain-Specific Standardization

### How to Repair and Make Sense of Legacy Data

Name1: Flugtaggen GMBH  
Name 2: rhamer strasse 20  
Address: dus  
City/Town: 40489  
Post Code:  
Country:



Business Name: Flugtaggen GMBH  
Personal Name:  
Street Name: Rhamer  
Street Type: Str.  
Street Number: 20  
City/Town: Düsseldorf  
Post Code: 40489  
Country: DE

### Value Added for CRM

- Fully automate data cleansing
- Apply country intelligence (names geographic, etc.)
- Standardize critical data elements
- Context-sensitive data interpretation
- Enrich data (geocoding, etc.)



**Increased accuracy = better business processes & better matching**



The MIT 2008 Information Quality Industry Symposium



## Understanding Global Data (Korean)

광주시 오포읍양벌리94-5대주파크빌2차207동 101호

Level 1

Level 2

Level 3

Level 4

Block Number

Sub-Block Num

Apt. Number

House Number

Postal Code





The MIT 2008 Information Quality Industry Symposium



## Understanding Global Data (Korean)

**Level 1** 경기도

**Block Number** 94

**Level 2** 광주시

**Sub-Block Num** 5

**Level 3** 오폭읍

**Apt. Number** 207

**Level 4** 양벌리  
대주파크빌2차아파트

**House Number** 101

**Postal Code** 464-764



The MIT 2008 Information Quality Industry Symposium



## Understanding Product Data

Product Description	
12oz D. Pepsi 12pack	Multiple Meanings 12 oz vs. 12 Pack
12pk C Orange Slice	
Mtn Dew 2ltr	Unstandardized
Code Red 24pk Bottles	
2L Mountain Dew Cs	
D.P. Cans 12p	

Free-Form Text: No Common Format

Duplicates



The MIT 2008 Information Quality Industry Symposium



## Identify Attributes/Categories

Product Description	Product	Container Size	Container Type	Packaging
12oz D. Pepsi 12pack	DIET PEPSI	12 OZ	<b>CANS</b>	12 PACK
12pk C Orange Slice	ORANGE SLICE	<b>12 OZ</b>	CANS	12 PACK
Mtn Dew 2ltr	MOUNTAIN DEW	2 L	<b>BOTTLES</b>	<b>8 CASE</b>
Code Red 24pk Bottles	CODE RED	<b>20 OZ</b>	BOTTLES	24 PACK
2L Mountain Dew Cs	MOUNTAIN DEW	2 L	BOTTLES	<b>8 CASE</b>
D.P. Cans 12p	DIET PEPSI	<b>12 OZ</b>	CANS	12 PACK



The MIT 2008 Information Quality Industry Symposium



## Key Take-aways

- Big Company = Big DQ Problems
  - Faster Growth = Bigger DQ Problems
- Unified Process for Data Quality is Key
- Domain Coverage is Important
  - Think enterprise solution, not point solution
- First Steps: Data and Metadata Comprehension

**Steve Sarsfield, Trillium Software**

[Steve\\_Sarsfield@trilliumsoftware.com](mailto:Steve_Sarsfield@trilliumsoftware.com) (978) 436-8768



The MIT 2008 Information Quality Industry Symposium



## What's In a Name? Multi-cultural Name Recognition and Data Quality

**Presenter:** Mala Narasimharajan  
IBM Corporation  
Product Marketing Manager

© 2008 IBM Corporation



The MIT 2008 Information Quality Industry Symposium



## Objectives

- What's In a Name
- Why are names complex
- Why even need name recognition
- Role of name-based information in data quality

© 2008 IBM Corporation



The MIT 2008 Information Quality Industry Symposium



## What's In A Name?

- Name are everywhere
- There are multiple variations of Andrew, Manual, John, Jeurgen – different ways of spelling the same name
- And every time; someone applies for a checking account, savings account, transfers money, cashes a check, boards a flight, or applies for credit these names must be looked up!



Nicknames, *Drew, Andraz, Drue*  
Shortened names, *Andy*,  
Prefixes, *Abdul, Fitz, O', De La*,



Andrewes,  
Andrews,  
Andrey,  
Andrezj,  
Andrian,  
Andriel,



Andros,  
Andru,  
Andruw,  
Andrzej,  
Andy,  
Drew,



Drue,  
Ohndrae,  
Ohndre,  
Ondre,  
Ondrei,  
Ondrej,



*Name.*, *Rev, Order, Hussein, Mohammed*  
*Abu Ali*  
Titles, *Dr Haj, Sri., Col*  
Phonetics, *Worchester, Wooster, "Worcester"*

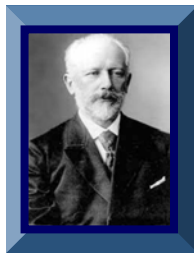
© 2008 IBM Corporation



The MIT 2008 Information Quality Industry Symposium



## Chaikovsky with a "T"



Tchaikovsky

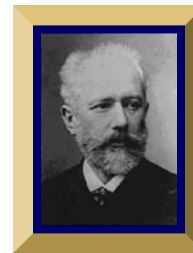
Names of people, places, and businesses

- There are no dictionaries for them,
- There's no way to look up a name and say this is wrong or right
- We run into very, very intractable problems [in transliteration from] other writing systems

So if somebody's looking for a name coming from the Cyrillic—for example, Tchaikovsky with a 'T' in front of it, matching will be very difficult as this is a French, not Russian transcription of his name


And thats just the surname, never mind the multiple variations of first:

- ✓ **Piotr Illich,**
- ✓ **Pyotr Ilich,**
- ✓ **Peter Illich**




Chaikovsky

© 2008 IBM Corporation

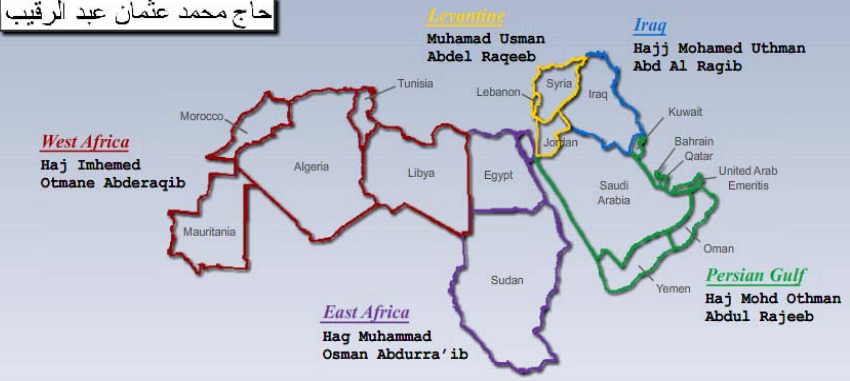


The MIT 2008 Information Quality Industry Symposium




## The Same Name across the Arabic World

حاج محمد عثمان عبد الرقيب




The Same Name Across the Arabic World


© 2008 IBM Corporation








The MIT 2008 Information Quality Industry Symposium




## Example: The Same Name across SE Asia




張丘蘇

	Zhang Qiusu
	Chang Ch'iu-Su
	Chiusu Sae Chang
	Cheung Yau So
	Cheung Yau So

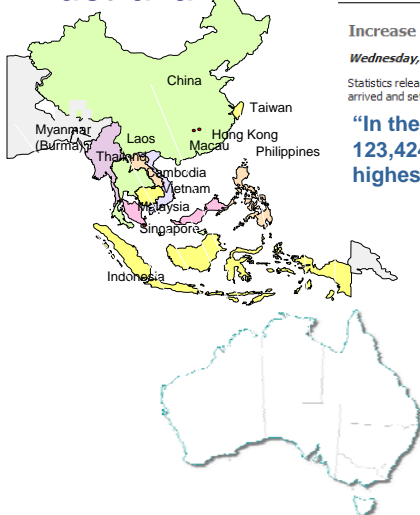
© 2008 IBM Corporation



The MIT 2008 Information Quality Industry Symposium



## Simple Name Recognition - A Critical Challenge For Australia



### Australian Immigration News

**Increase in number of new migrants**

*Wednesday, 18 January 2006*

Statistics released by the Immigration Minister on 31 December 2005 show that more than 123,000 people arrived and settled in Australia during 2004-2005, reflecting a 40% increase over the past 10 year period.

**“In the Australians’ case they have welcomed 123,424 new immigrants in 2004-2005, the highest number in more than 15 years”**

**The Problem:**

- Many have no proof of identity
- Many have disposed of all personal papers en route to Australia
- It is not uncommon for them to **‘change identity’** either during the journey or processing, in the hope that it may be easier to stay if they claim a different nationality

**The Question:**

- This raises legitimate questions about their intentions

© 2008 IBM Corporation



The MIT 2008 Information Quality Industry Symposium



## Sure You Can Standardize An Address Or Phone Number But Names Have Always Been A Challenge

### Address & ## Standardization

4737 Simeron Drive  
Easton, MA 02334  
(978)36 5-5312

↓

4737 **Cimarron** Drive  
Easton, MA 02334  
(978) 365-5312

### Name Standardization?

Teddd Kennedy	Edl Kennedy
xEd Kennedy	Ed Kennedy
Kim June Joe	Kimie Spacek



- No single “dictionary” of “right” spellings
- No one-to-one correspondence among nicknames to names
- Poor data quality is common
- Cultural syntax variations are problematic

© 2008 IBM Corporation



The MIT 2008 Information Quality Industry Symposium



## Data Collection Forms Create Name Parsing Errors

### Name Formats

- Last Name \_\_\_\_\_  
First Name \_\_\_\_\_  
Title \_\_\_\_\_
- Name \_\_\_\_\_
- Middle Name \_\_\_\_\_  
Last Name \_\_\_\_\_  
First Name \_\_\_\_\_
- Family Name \_\_\_\_\_  
Given Name \_\_\_\_\_  
Middle Initial \_\_\_\_\_

© 2008 IBM Corporation

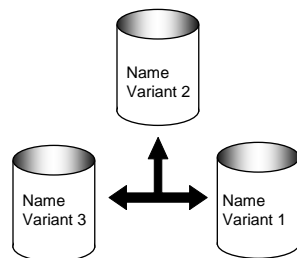


The MIT 2008 Information Quality Industry Symposium



## Common Problems with Names

### Database Problems



Same name is represented differently across corporate databases – standard approaches to match them are often inadequate

**Exact Match**

**Soundex (1918)**

**NYSIIS (1963)**

**“Home - Grown”**

© 2008 IBM Corporation



The MIT 2008 Information Quality Industry Symposium



## Why do you need Name Recognition ?

- Ability to recognize multi-cultural names from around the world – and provide insight, analysis and matching capabilities
- For organizations where names – personal or business constitute a vital data element
- Dealing with name is essential part of data quality
- Applications and solutions that rely on identity – rely on names
- Names are important to data quality as they enhance identification, merging, standardization and enrichment steps
- Many organizations must be able to recognize and match names (e.g., banks, insurance companies, law enforcement, airlines)
- Global business interactions demand ability to process multi-cultural names with greater accuracy

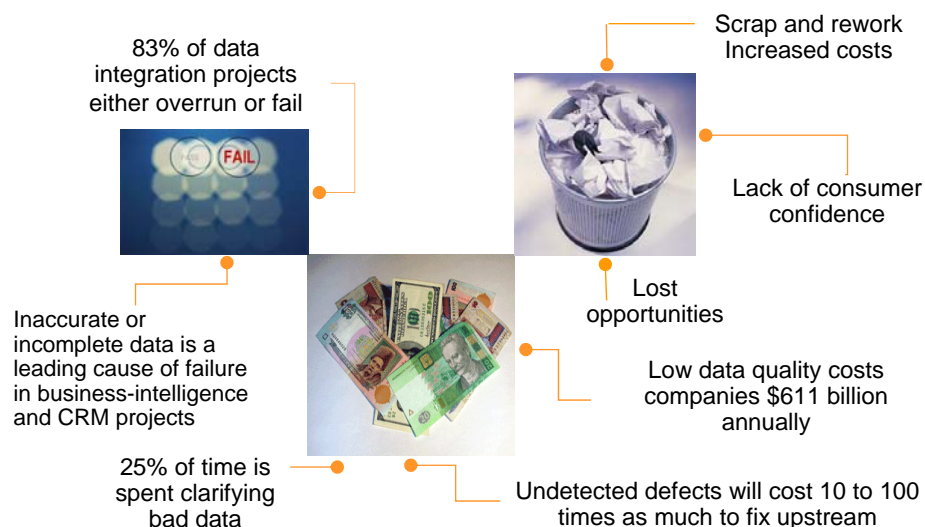
© 2008 IBM Corporation



The MIT 2008 Information Quality Industry Symposium

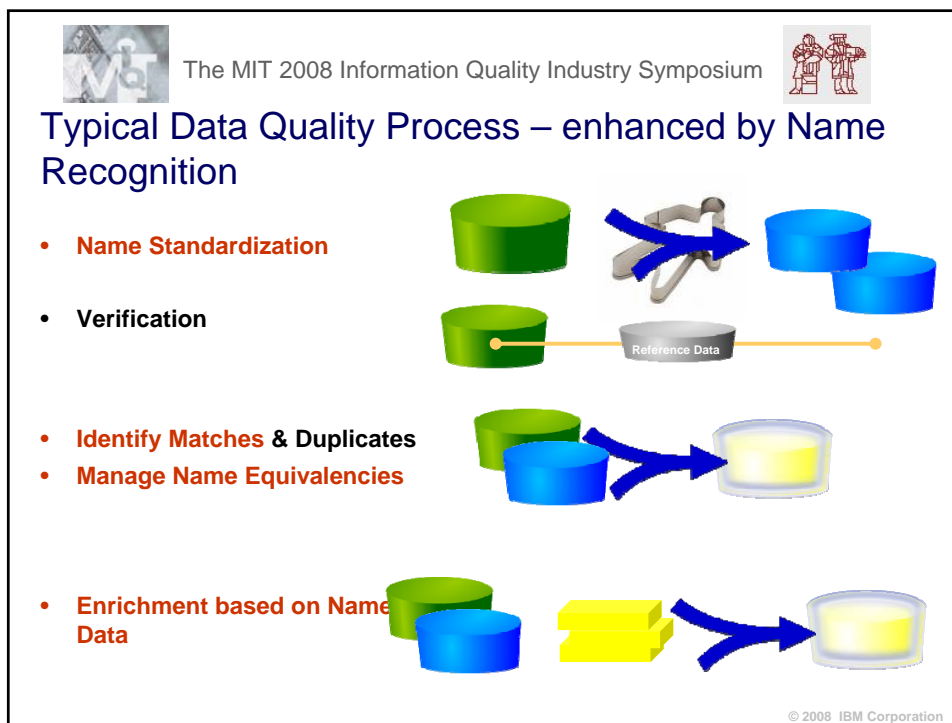
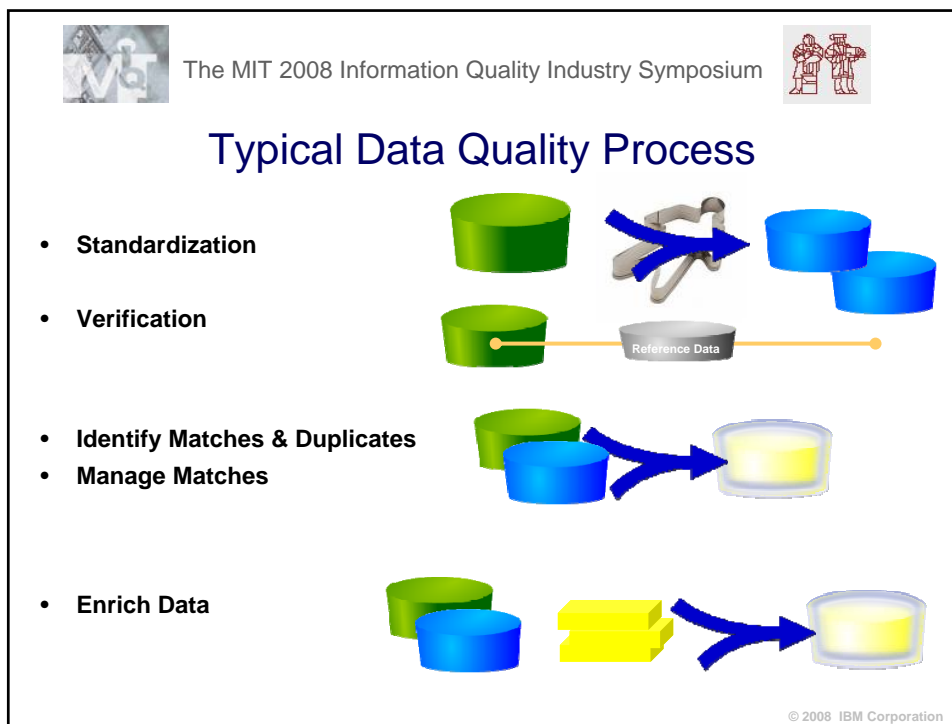


## Costs of Dirty Data



© 2008 IBM Corporation







The MIT 2008 Information Quality Industry Symposium



## Capabilities of Global Name Recognition

*How does this apply to Data Quality ?*

**Transliterate**  
Incoming name  
in non-Roman  
native script



**Parse**  
Analysis and  
remediation of  
name data



**Classify**  
Identifies culture  
and the proper  
search techniques



**Genderize**  
Identifies the  
most likely  
gender of a  
name



**Name Variations**  
Ranked name  
variations used  
as query terms  
for search



**Search**  
Ranked list of  
potential matches  
from one or more  
data sources



Standardization

Enrichment

© 2008 IBM Corporation



The MIT 2008 Information Quality Industry Symposium



## Conclusions

- Data quality and name recognition are complementary to each other
- Today- there are strong integration synergies between Global Name Recognition and QualityStage
- Global Name Recognition abilities significantly enhance data quality efforts by:
  - Providing enhanced insight, matching and standardization around names from around the world
  - Providing better match rates and higher precision and recall

© 2008 IBM Corporation



# Homeland Security

## “Meaningful Engagement for Information Quality”

Presenter: Glenn Norton, U.S. Citizenship and Immigration Services

MIT 2008 Information Quality Industry Symposium



Homeland Security

## Topics

- Operational Context
- System vs. Information Focus
- Engagement Model
- Lessons Learned
- Next Steps



Homeland Security

## Why was DHS Created?

*“It was not created merely to bring together different agencies under a single tent. It was created to enable these agencies to secure the homeland through joint coordinated action. Our challenge is to realize that goal to the greatest extent possible”\**

Three focus areas for achieving the goal:

- Operate under a common picture of threats
- Provide active and appropriate policy response to those threats
- Ensure unified execution of Component operations to carry out the mission

Success in each of these areas depends on knowing what information is available and appropriately sharing the right information with the right people at the right time

\*April 20, 2005 Statement for the Record by Secretary Chertoff before the Senate Subcommittee on Homeland Security



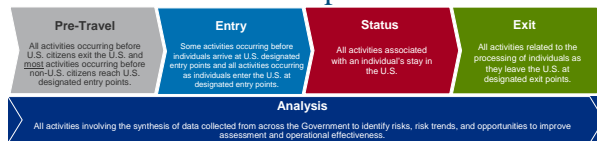
## Complex Environment

- Organizational
- Operational
- Technical
- Budgetary

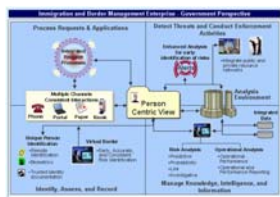


## Evolving Vision

### Changing Vision of the Immigration & Border Management Enterprise



- Dept. of Homeland Security
  - Customs & Border Protection
  - Immigration & Customs Enforcement
  - Citizenship & Immigration Services
  - US-VISIT
- Dept. of State
- Dept. of Justice
- Dept. of Transportation
- Dept. of Commerce
- State and local governments
- Foreign governments

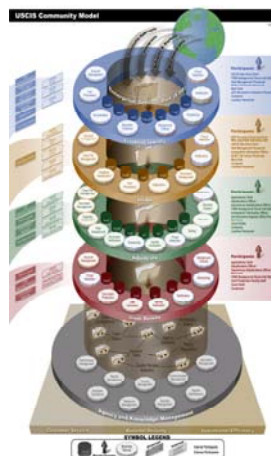


**Homeland Security**

## Transforming Operations

U.S. Citizenship & Immigration Services is embarking on an enterprise-wide transformation effort that will transition the Agency from a fragmented, paper-based filing system to a centralized and consolidated electronic environment. This effort will require re-engineering agency-wide business processes and updating information technology systems to provide new capabilities to our employees and our customers.

The USCIS Transformation Program Office is charged with facilitating the development of a flexible and efficient organizational business model supported by an integrated technical environment for both its customers and employees.



**Homeland Security**

## Topics

- Operational Context
- System vs. Information Focus
- Engagement Model
- Lessons Learned
- Next Steps



Homeland  
Security

## Time Travelers and the Walking Dead

Decisions are being made about the information in our systems.  
Who is making them?

Developers make an assumption and time travel is possible  
System migration identifies deceased on vacation

How can we ensure the information in our systems fit for use?



Homeland  
Security

## Five Business Questions that Must be Asked

1. How do I find the best source of information available to meet my specific business needs?
2. Is the information I create in my business also needed by other organizations or lines of business (within DHS, and ultimately outside of DHS as well)?
3. How can I have confidence that the information I receive from others is actionable for my business purposes?
4. How can I be sure that the information I provide to others will be construed correctly and handled appropriately?
5. How can I best leverage available resources to obtain optimal business value from data and information assets?



Homeland  
Security

## Information Management Environment

- The ability of an organization to answer the five business questions establishes the context and maturity of the organization's data and information management environment
- Lack of answers indicate a high probability that mission goals are not being effectively attained

*Who do we ask?*

*Who can correctly answer?*



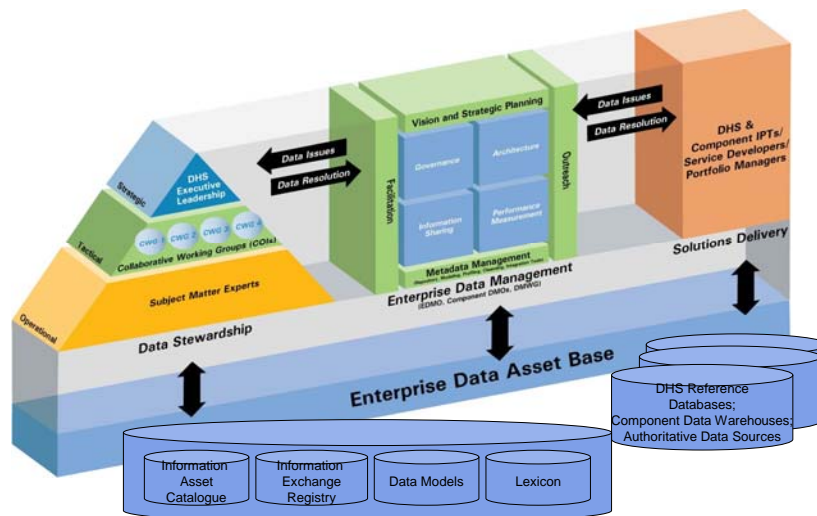
Homeland  
Security

## Topics

- Operational Context
- System vs. Information Focus
- Engagement Model
- Lessons Learned
- Next Steps



## Notional Operational & IT Engagement Model



DHS Enterprise Architecture Information Repository



## Notional IT & Business Data Management Roles

- **Data Governance And Management Strategies**
  - Data Management Vision & Strategic Planning
  - Data Management Policies, Processes & Procedures
  - Data Stewardship Program Facilitation (Define Data Standards, Data-Related Business Rules, Authoritative Data Sources, Information Security Categorization and Privacy Rules)
  - Metadata Management Policies & Practices
  - Data Management Outreach Program
  - Data Management Performance Measurement Strategy Planning
  - Data Change Management Policies & Practices
- **Data Management Services & Solutions**
  - Enterprise Data Architecture And Integration
    - Baseline Current Architecture, Gap Analysis And Recommendations
    - Define Data Management Standards And Monitor Compliance
    - Define Enterprise Data Architecture And Monitor Compliance
    - Oversee Increment Data Models (Harmonization & Integration)
  - Program-Wide Data Administration & Technical Metadata Management
    - Manage Metadata (Discover, Define & Contribute Classifications, Taxonomy & Vocabulary)
    - Manage Conceptual, Logical & Physical Data Models
    - Define And Implement Data Exchange/Sharing
    - Execute Data Standardization Procedures (E.G., Naming Standards)
    - Implement Security And Privacy Protections, Information Security Categorization
  - Database Administration
  - Data Management Tools Support
- **Data Assurance; Performance Measurement; Outreach**
  - Data Quality Management (Audit, Recommend And Monitor)
  - Performance Scorecard (Measurement And Reporting)
  - Data Management Services and Solutions Outreach Program

**Homeland Security**

## Data Governance

*Data governance is the exercise of authority and accountability for decision making and conflict resolution to ensure data assets are managed in the best interest of the enterprise*

- Data Governance establishes policies, processes, and authority and accountability for decision making to ensure data is managed as valued asset, like money, equipment, facilities, etc.
- Data Stewardship provides a structured framework to ensure business people are accountable for managing the business aspects of data (definitions, business rules, quality, security and privacy parameters)

**Homeland Security**

## Topics

- Operational Context
- System vs. Information Focus
- Engagement Model
- Lessons Learned
- Next Steps



Homeland  
Security

## Lessons Learned

- Management of information quality directly affects mission performance and thus must be treated as a mission critical function, not simply an administrative IT function
- Information quality is really all about the data and information being “fit for use” within each specific (relative) business context
- When data is shared across systems or organizations, the business context in which that data will be used is often altered, sometimes significantly
- Information sharing carries with it a serious responsibility for ensuring information quality, fitness for use of the data within the new business context
- Acknowledgement of this responsibility across organizations and development of mechanisms to ensure information quality are some of the most significant challenges that must be addressed in information sharing



Homeland  
Security

## Lessons Learned

- Systems engineering efforts are required when the information needed to address business needs is contained in separate systems and must be shared
- However, without business context or responsible parties to explain that data and information as it travels through its full lifecycle to all stakeholders, assumptions and misunderstandings may occur
- The goal of systems engineering is to produce a technical capability; the goal of data and information management is to provide information that is “fit for use” in specific business contexts

---

*Do not confuse information management with systems engineering*

*Do not assume systems engineering efforts will adequately address information management*

*Explicitly define the information management function in program planning and execution*



Homeland  
Security

## Topics

- Operational Context
- System vs. Information Focus
- Engagement Model
- Lessons Learned
- Next Steps



Homeland  
Security

## Next Steps

- Incorporate lessons learned into future initiatives
- Devote concerted effort into converting tacit understanding (or misunderstanding) of organizational, operational and IT roles and responsibilities into explicit engagement models
- Leverage developing DHS data management capabilities and Communities of Interest
- Continue outreach efforts



Homeland  
Security



MIT 2008 Information Quality  
Industry Symposium  
July 16-17, 2008  
Boston, Massachusetts

## Ten Steps to Quality Data and Trusted Information™ - An Overview

Danette McGilvray  
Granite Falls Consulting, Inc.  
President and Principal  
Phone: 510-501-8234  
Email: danette@gfalls.com  
Web: www.gfalls.com  
Fremont, California USA

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

1

## Copyright Information

These materials, and any part thereof, are protected under copyright law. Course participants are granted the right to "fair use" of these materials. The contents of this document may not be reproduced or transmitted in any form, in whole or in part, or by any means, mechanical or electronic, for any other use, without the express written consent of Danette McGilvray or Elsevier Inc.

Portions of this work are from the forthcoming book available Summer 2008, *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information™*, by Danette McGilvray, published by Morgan Kaufmann Publishers, Copyright 2008 Elsevier Inc. All rights reserved.

All uses of Ten Steps to Quality Data and Trusted Information™, The Ten Steps™, or Ten Steps™ throughout these materials are protected by trademark law.

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

2

## Presentation Description

Data quality situations familiar to many organizations:

- Business has invested heavily in data purchased from external sources, yet cannot depend on the quality to meet the company needs.
- The data warehouse has been in production for over a year. Users from the business intelligence group don't trust the reports, complain about the quality, and are reverting to their own spreadsheets for verification.
- An ERP (Enterprise Resource Planning) application has been implemented. Data previously used by one business function is being used in end-to-end processes – with poor results.
- The organization is starting a data integration project. The project team has a tight schedule, yet already knows there are quality issues with the source data to be moved.

There is help available! This presentation provides an overview of a methodology, Ten Steps to Quality Data and Trusted Information™, which is a systematic approach to improving and creating data and information quality. The methodology combines a conceptual framework for understanding information quality and The Ten Steps™ process which provides instructions, techniques, and best practices for implementing the key concepts.

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

3

## Background

Danette McGilvray is President and Principal of Granite Falls Consulting, Inc., a firm specializing in information quality management. Projects include enterprise data quality services, data warehousing strategies, data governance, and best practices for large-scale ERP data migrations for Fortune 500 organizations. Her book on data quality, "Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information™" (Morgan Kaufmann) will be available Summer 2008.

Danette is an invited speaker at conferences throughout the US and Europe. She is a member of DMReview.com's Ask the Expert panel. Her previous experience as a leader of enterprise data quality within a company and now working with clients in various industries, gives her understanding of the information quality challenges faced daily by organizations. She has been profiled in PC Week and HP Measure Magazine and was an invited delegate to the People's Republic of China to discuss roles and opportunities for women in the computer field.

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

4

## Dealing with Poor Health

**“Doctor, my left arm hurts!”**

The doctor puts your arm in a sling, gives you an aspirin and tells you to go home.



**But what if you were really having a heart attack?**

You would expect the doctor to **diagnose** your condition and take **emergency measures** to save your life.

After you were **stabilized** you would expect the doctor to:

- Run **tests**
- Get to the **root cause** of the heart attack
- Recommend measures to **correct damage** done (if possible) and **prevent** another heart attack.

The doctor would have you come in for **periodic additional tests** and **follow-up to assess** your condition.

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

5

## Dealing with Unhealthy Data and Information

When it comes to data and information quality, how often do we:

- Address the immediate problem, then
  - Go for the “easy fix” (the aspirin and sling) and
  - Expect that to take care of our problems?
- 
- No tests or assessments are run to determine the location or magnitude of our problems
  - No root cause analysis is performed
  - No preventive measures are put into place

And then we are surprised when the problems appear and reappear!

Just like your own health, you can:

- **Prevent** data quality “health” problems
- **Assess** and take **action** if they appear

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

6

## Dealing with Unhealthy Data and Information

This presentation will introduce you to a methodology that will help with your organization's data and information quality health.

Common data quality situations:

- Company has implemented an **ERP** (Enterprise Resource Planning) application. Data previously used by one business function is being used in end-to-end processes – with poor results.
- The company is starting a **data integration** project. The project team has a tight schedule, yet already knows there are quality issues with the source data to be moved.
- The company invests in a major **data clean-up project**, and a few years later starts **another data clean-up project** because data quality declined and is causing issues for the business.
- The **data warehouse** has been in production for over a year. Users from the business intelligence group don't trust the reports, complain about quality, and are reverting to their own spreadsheets for verification.
- Business has invested heavily in **data purchased from external sources**, yet cannot depend on the quality to meet the company needs.
- Data quality is an important part of your **daily responsibilities**.

## Data vs. Information

- **Data** – Known facts or other items of interest to the business
- **Information** – facts within context
- **Are there differences between the two?**
- This approach does not generally differentiate between data and information. Some organizations respond to “data quality” and others respond to “information quality.” Use the term most meaningful to your organization and those with whom you are speaking.



## Ten Steps to Quality Data and Trusted Information

A systematic approach to **improving and creating data and information quality** within your business.

The methodology combines:

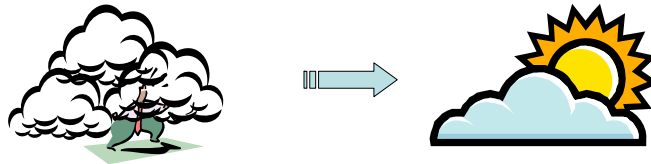
- A **conceptual framework** for understanding information quality with
- The **tools, instructions, and best practices** for improving information quality.

**Your company's\* "wellness" program for data and information.**

\* Company includes any organization such as for-profit businesses, non-profit and charitable organizations, government agencies, and educational institutions.



## Outline



- **Key Concepts**
- **The Ten Steps Process and Projects**
- **Summary and Best Practices**

## Framework – Like the Food Pyramid

Just as the Food Pyramid provides **guidelines and a visual of the components** for healthy eating and physical activity, the Framework for Information Quality provides the components necessary for “healthy” data.



Source: mini poster.pdf available at <http://www.mypyramid.gov/>, accessed July 16, 2007. The Food Pyramid was developed by The Center for Nutrition Policy and Promotion, an organization of the U.S. Department of Agriculture. It was established to improve nutrition and promote dietary guidance for all Americans.

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

11

## Framework for Information Quality (FIQ)



© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

08-01

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

12

## FIQ Sections Explained (1)

- ① **Business Goals/Strategy/Issues/Opportunities:** The “why”– Anything done with information should support the business in meeting its goals.
- ② **Information Life Cycle\*** Use POSMAD to help remember the information life cycle:
  - **P**lan - Identify objectives, plan information architecture, develop standards and definitions; model and design applications, databases, processes, organizations, etc.
  - **O**btain – Data or information is acquired in some way, e.g. create records, purchase data, load external files, etc.
  - **S**tore and Share – data is stored and made available for use.
  - **M**aintain – Update, change, manipulate data; cleanse and transform data, match and merge records, etc.
  - **A**pply - “Retrieve” data, use information. Includes all information usage such as completing a transaction, writing a report, making a management decision, completing automated processes, etc.
  - **D**ispose –Archiving information; delete the data or records.

\* Also known as an Information Chain, Information Value Chain, Information Resource Life Cycle

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

13

## FIQ Sections Explained (2)

- ③ **Key Components** affecting information quality
  - **Data (What)** - Known facts or other items of interest to the business.
  - **Process (How)** - Activities, actions, tasks, or procedures that touch the data or information (business processes, data management processes, processes external to the company, etc.).
  - **People/Organizations (Who)** - Organizations, teams, roles/responsibilities or individuals.
  - **Technology (How)** – Forms, applications, databases, files, programs, code, or media that store, share, or manipulate the data, are involved with the processes, or are used by the people and organizations.
- ④ **Interaction Matrix** between information life cycle phases and key components
- ⑤ **Location (Where) and Time (When and How Long)**

Note: Top half of the framework along with the first orange bar answers the interrogatives of who, what, how, why, where, when, and how long

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

14

## FIQ Sections Explained (3)

- ⑥ **Broad-Impact Components** – additional factors affecting information quality. Lower your risk by ensuring these factors have been considered and appropriately addressed. If they are *not* addressed, you are still at risk (RRISCC) as far as information quality is concerned.
- **R**equirements and Constraints
  - **R**esponsibility
  - **I**mprovement and Prevention
  - **S**tructure and Meaning
  - **C**ommunication
  - **C**hange
- ⑦ **Culture and Environment**- Your company's attitudes, values, customs, practices and social behavior; conditions that surround people in your company and affect the way they work and act. Take into account to better accomplish your goals.

## Using the Framework

Use as a tool for:

- **Diagnosis** – Realize where breakdowns are occurring; assess your practices, determine if all components necessary for information quality are present.
- **Planning** – Design new processes, determine where to invest time, money, and resources.
- **Communication** – Explain the components required for and impacting information quality.

The framework allows us to organize our thinking in a way so we can plan and **take effective action**.

## Data Quality Dimensions



Aspects or features of information and a way to classify information and data quality needs. Dimensions are used to define, measure, and manage the quality of the data and information.

- In order to improve information quality, there must be a way to measure it.
- Measure the dimensions that best address your business need.
- There is no industry standard for the types of data quality dimensions.
- The dimensions defined here are derived from experience and are those most feasible and useful within the usual constraints of most businesses.

## Data Quality Dimensions



<b>Data Specifications</b>	A measure of the existence, completeness, quality, and documentation of data standards, data models, business rules, metadata, and reference data.	<b>Ease-of-Use and Maintainability</b>	A measure of the degree to which data can be accessed and used and the degree to which data can be updated, maintained, and managed.
<b>Data Integrity Fundamentals</b>	A measure of the existence, validity, structure, content and other basic characteristics of data.	<b>Data Coverage</b>	A measure of the availability and comprehensiveness of data compared to the total data universe or population of interest.
<b>Duplication</b>	A measure of unwanted duplication existing within or across systems for a particular field, record, or data set.	<b>Presentation Quality</b>	A measure of how information is presented to and collected from those who utilize the information. Format and appearance support the appropriate use of the information.
<b>Accuracy</b>	A measure of the correctness of the content of the data (which requires an authoritative source of reference to be identified and accessible).	<b>Perception, Relevance, and Trust</b>	A measure of the perception of and confidence in the data quality; the importance, value, and relevance of the data to the business needs.
<b>Consistency and Synchronization</b>	A measure of the equivalence of information stored or used in various data stores, applications, and systems, and the processes for making data equivalent.	<b>Data Decay</b>	A measure of the rate of negative change to the data.
<b>Timeliness and Availability</b>	A measure of the degree to which data are current and available for use as specified and in the timeframe in which they are expected.	<b>Transactability</b>	A measure of the degree to which data will produce the desired business transaction or outcome.

## Assessments and Dimensions of Quality

Different tools, techniques, and processes are used to assess, measure, and manage the various dimensions of quality (with varying levels of time, money, and resource required).

Why differentiate the dimensions of quality?

- Match dimensions against a business need and prioritize which assessments to complete and in what order.
- Understand what you will (and will not) get from assessing each dimension
- Better define and manage the sequence of activities in your project plan within time, money, and resource constraints

## Determining What to Assess for Quality

First, understand the business issues driving the data quality assessment and improvement activities. Then ask yourself:

- **Should** I assess the data?
  - Only spend time testing when you expect the results to give you actionable information related to your business needs
- **Can** I assess the data?
  - Is it possible or practical to look at this quality dimension?
  - Sometimes you cannot practically assess the dimension of quality or the cost to do so is prohibitive

Only assess and manage quality for those dimensions where the answer to both of the questions above is “yes.”

## Business Impact Techniques

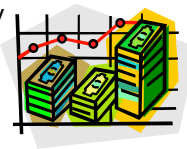
Awareness of data quality issues leads to questions:

- “Why does this matter?”
- “Why should I care?”
- “What impact does this have on the business?”

Answer those questions by using quantitative and qualitative techniques to assess the impact of data quality on the business.

Use results from assessing business impact to:

- Establish the business case for information quality
- Gain support for investing in information quality
- Determine the optimal level of investment



MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

21

## Business Impact Techniques

1	Anecdotes	Collect examples or stories of the impact of poor data quality.
2	Usage	Inventory the current and/or future uses of the data.
3	Five “Whys”	Ask “Why” five times to get to real business impact.
4	Benefit vs. Cost Matrix	Analyze and rate the relationship between benefits and costs of issues, recommendations, or improvements.
5	Ranking and Prioritization	Rank impact of missing and incorrect data to specific business processes.
6	Process Impact	Illustrate the effects of poor quality data to business processes.
7	Cost of Low Quality Data	Quantify the costs and revenue impact of poor quality data.
8	Cost-Benefit Analysis	Compare potential benefits of investing in data quality with anticipated costs through an in-depth evaluation. Includes Return on Investment (ROI) – profit from an investment as a percentage of the amount invested.

Less Complex/ Less Time      **Relative Time and Effort**      More Complex/ More Time

←      1   2   3   4   5   6   7   8      →

08-01

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

22

## Choosing Which Techniques to Use

- Use the techniques that best fit your situation, time, and resources available.
  - Many of the techniques work together or can be used alone
- The continuum shows relative effort – not relative results:
  - You can understand business impact even without completing a full cost/benefit analysis
  - Less complicated does not necessarily mean less useful results
  - More complex does not necessarily mean more useful results
  - The best results come from using the techniques most appropriate to your situation

## Root Cause Analysis

Root cause analysis is the process of analyzing all possible causes of a problem, issue, or condition to determine the actual cause.

1	<b>Five “Whys” for Root Cause</b>	Leverage a basic quality approach by asking “Why” five times to get to root cause.
2	<b>Track and Trace</b>	Identify the location of the problem by tracking the data through the information life cycle and determining the root cause where the problem first appears.
3	<b>Cause-and-Effect / Fishbone Diagram</b>	Use a standard quality technique to identify, explore, and graphically display all possible causes of an issue.

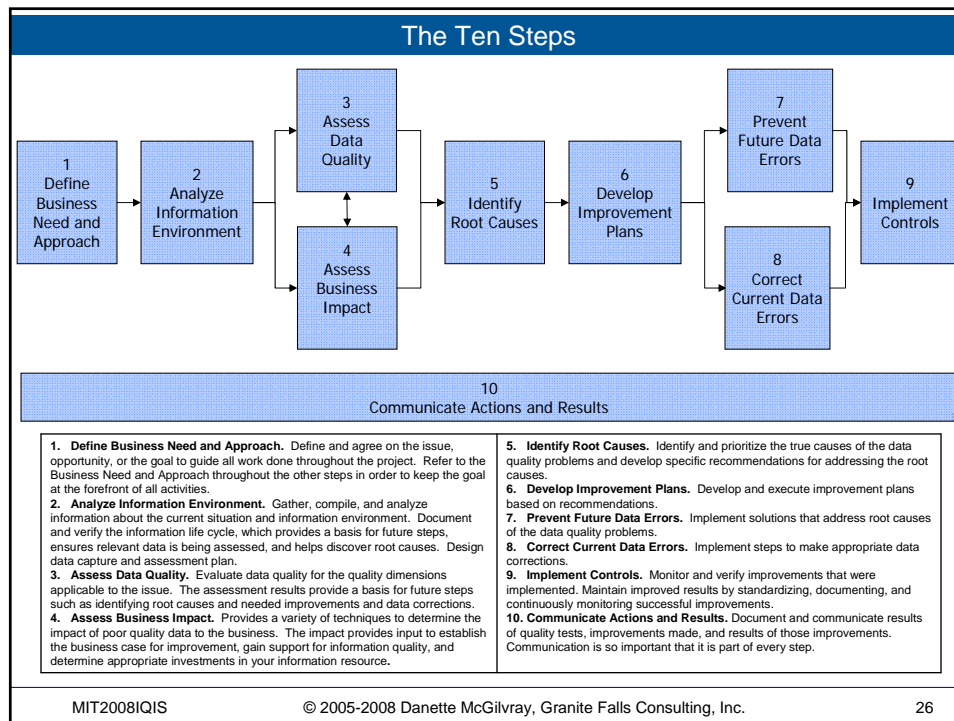


## The Ten Steps Process and Projects

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

25



## Using the Ten Steps Methodology

**Pick and choose** appropriate steps, activities, and techniques from the methodology:

- For **information-quality focused** projects
- In the course of **daily work** where you have responsibility for managing data quality or the work you do impacts data quality
- To **integrate** specific DQ/IQ activities into **other projects and methodologies** (e.g. ERP migration or building a data warehouse)

## Approaches to Data Quality in Projects (1)

### Establish Business Case

- Exploratory assessment or quick proof of concept assessing data quality on a very limited set of data. As an individual, you can implement a brief project that will help you make a business case for further data quality improvements. If you already have a specific data quality problem, you may just want to assess the business impact of that problem without further quality assessment.

### Establish Data Quality Baseline

- When the business has committed to improving data quality and there is support for a project team and resources.

### Determine Root Causes

- Use this approach when you already know the data quality issues and have determined the impact of those issues warrants further investigation into the real cause.

## Approaches to Data Quality in Projects (2)

### Implement Improvements

- Execute the recommendations developed when the data quality assessment and business impact analysis generate a plan for data quality improvement.

### Implement Ongoing Monitoring and Metrics

- Focus on instituting operational processes for monitoring, evaluating, reporting, and acting on results.

### Address Data Quality as an Individual

- Use data quality techniques in the course of daily work where you have responsibility for managing data quality or to address a specific data quality issue as an individual.

### Integrate Data Quality Activities into Other Projects and Methodologies

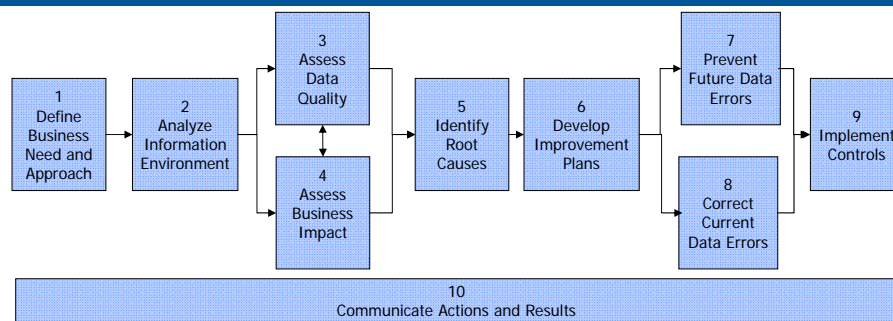
- Combine The Ten Steps activities with your company's favored project management and project life cycle and include in your specific project plan.

MIT2008IQIS

© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

29

## Approaches to Data Quality and The Ten Steps Process



**Establish Business Case:** Steps 1, 2, 3, 4, 10

**Establish Data Quality Baseline:** Steps 1, 2, 3, 4, 5, 6, 10

**Determine Root Causes:** Steps 1, 2, 3, 4, 5, 6, 10

**Implement Improvements:** Leverage baseline results plus Steps 7, 8, 10

**Implement On-going Monitoring and Metrics:**

Leverage baseline results and improvements implemented plus Steps 9 and 10

**Address Data Quality as an Individual:** Steps vary

**Integrate Data Quality Activities Into Other Projects and Methodologies:** Steps vary

MIT2008IQIS

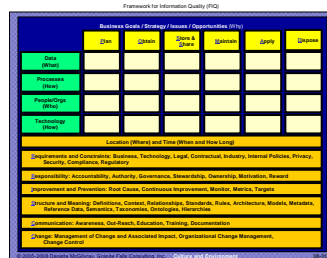
© 2005-2008 Danette McGilvray, Granite Falls Consulting, Inc.

30

## Summary and Best Practices

## The Methodology Has Two Main Components

### Framework for Information Quality (FIQ) and Other Key Concepts



### Ten Steps Process



- Provides the foundation for understanding information and data quality
- Shows the components necessary for information quality
- Concrete instructions for implementing, improving, and creating data quality
- Process for implementing framework and key concepts
- Contains examples, templates, techniques, and advice

## Best Practice – Apply to Any Data and Information

Use what you have learned to improve any data:

- Customer
- Order Management
- Sales and Marketing
- Finance
- Procurement
- Manufacturing
- Etc.



Apply to any category of data:

- Master data
- Transactional data
- Reference data
- Metadata

## Guidelines for Applying the Methodology

- **Relevant.** Ensure your work is associated with the business issue to be resolved.
- **Pick-and-choose.** Use only those steps applicable to your project.
- **Level of detail.** Start at a high level and go to more detail only if needed.
- **Scale.** Use for one-person few week project to a several-month project with project team. Use in your individual work.
- **Reuse (80/20 rule).** Bring together existing knowledge in such a way that you can understand it better. Supplement existing material with original research only as needed.
- **Tool independent.** Make better use of the tools you have.

## Do's and Don'ts

- You don't have to have the CEO's support to get started
  - You DO have to have the appropriate level of management support to start while continuing to obtain management support as high up in the organization as possible
- You don't have to have all the answers
  - You DO need to do your homework, know your company, and be open to many options
- You don't need to do everything all at once
  - You DO need to have a plan of action and get started

## Questions??

Thank you for attending and your participation!  
Feel free to contact me if you have comments or questions:  
[danette@gfalls.com](mailto:danette@gfalls.com)

# THANK YOU!

Danette McGilvray  
Granite Falls Consulting, Inc.  
President and Principal      Email: [danette@gfalls.com](mailto:danette@gfalls.com)  
Phone: 510-501-8234      Web: [www.gfalls.com](http://www.gfalls.com)  
Fax: 510-505-9898      Fremont, California USA

## Selected Resources (1)

Brassard, Michael and Diane Ritter. *The Memory Jogger II – A Pocket Guide of Tools for Continuous Improvement and Effective Planning*. GOAL/QPC. Methuen, MA. 1994.

English, Larry. *Improving Data Warehouse and Business Information Quality*. Wiley. New York. 1999. ISBN: 0471253839.

Eppler, Martin. *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes*, Springer 2003. ISBN: 3540003983.

Hay, David C. *Requirements Analysis: From Business Views to Architecture*. Prentice Hall PTR, New Jersey. 2003.

Huang, Kuan-Tsae, Yang W. Lee and Richard Y. Wang. *Quality Information and Knowledge*. Prentice Hall. New Jersey. 1999.

Loshin, David. *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann. San Francisco. 2001. ISBN: 0124558402.

McGilvray, Danette. *Data Quality and the Project Life Cycle*. The Data Administration Newsletter – tdan.com. July 10, 2007. <http://www.tdan.com/view-articles/5092>

## Selected Resources (2)

McGilvray, Danette. *Data Governance: A Necessity in an Integrated Information World, Part 1*. DM Review. December 2006.

[http://www.dmreview.com/article\\_sub.cfm?articleId=1069951](http://www.dmreview.com/article_sub.cfm?articleId=1069951).

*Part 2*. DM Review. January 2007.

[http://www.dmreview.com/article\\_sub.cfm?articleId=1072431](http://www.dmreview.com/article_sub.cfm?articleId=1072431)

Olson, Jack E. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann. San Francisco. 2003.

Redman, Thomas C.: *Data Quality: Management and Technology*, Bantam, New York, NY, 1992.

Redman, Thomas E. *Data Quality for the Information Age*. Artech House. Boston. 1996.

Redman, Thomas C. *Data Quality The Field Guide*. Digital Press. Boston. 2001.

Rummler, Geary A. and Alan P. Brache. *Improving Performance: How to Manage the White Space on the Organization Chart*. Jossey-Bass. San Francisco. 1990.



The MIT 2008 Information Quality Industry Symposium



# Using Conceptual Data Modeling to ensure high Information and Data Quality

Pete Stiglich

Senior Consultant

[PStiglich@ewsolutions.com](mailto:PStiglich@ewsolutions.com)

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 1

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



## EWSolutions' Background

EWSolutions is a Chicago-headquartered strategic partner and full life-cycle systems integrator providing both award winning strategic consulting and full-service implementation services. This combination affords our clients a full range of services for any size enterprise information management, managed meta data environment, and/or data warehouse/business intelligence initiative. Our notable client projects have been featured in the Chicago Tribune, Federal Computer Weekly, Crain's Chicago Business, and won the 2004 Intelligent Enterprise's RealWare award, 2007 Excellence in Information Integrity Award nomination and DM Review's 2005 World Class Solutions award.

**Information  
Integrity Coalition**  
2007 Excellence in  
Information Integrity Award  
Nomination

**intelligent  
Enterprise REAL  
Awards TRANSFORM**  
**2004 WINNER**  
Best Business Intelligence  
Application  
Information Integration  
Client: Department of Defense

**Chicago Tribune**



World Class  
Solutions Award  
Data Management

For more information on our Strategic Consulting Services, Implementation Services, or World-Class Training, call toll free at 866.EWS.1100, 866.397.1100, mail number 630.920.0005 or email us at [Info@EWSolutions.com](mailto:Info@EWSolutions.com)

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 2

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™





## Professional Profile / Contact Information

**Pete Stiglich** is a Senior Consultant with EWSolutions with nearly 25 years of IT experience in the fields of Data Modeling, Data Warehousing, Business Intelligence, meta data Management, Data Integration, Customer Relationship Management (CRM), Customer Data Integration (CDI), Database Design and Administration, Data Quality, and Transaction Processing. Pete has architected Enterprise Information Management solutions for diverse industries such as Insurance, Credit Card, Medical, Retail, Banking, Manufacturing, Telecom, and Government.

Pete has developed and taught courses on Dimensional Data Modeling, Conceptual Data Modeling, ER/Studio, and SQL. Pete has presented for DAMA at the international and local level, as well as at the 2007 IADQ Conference. Pete's articles on Data Architecture have been published in *Real World Decision Support*, *DMForum*, *InfoAdvisors*, and the *Information and Data Quality Newsletter*. Pete is a listed expert in SearchDataManagement on the topics of data modeling and data warehousing.

For the current issue of Real World Decision Support  
**See:** [http://www.ewsolutions.com/resource-center/rwds\\_folder/rwds-curr-issue/](http://www.ewsolutions.com/resource-center/rwds_folder/rwds-curr-issue/)

Email: [PStiglich@EWSolutions.com](mailto:PStiglich@EWSolutions.com) Phone: [602-284-0992](tel:602-284-0992)



[www.EWSolutions.com](http://www.EWSolutions.com)  
 © 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 3

**Strategic Partner & Systems Integrator**  
 Intelligent Business Intelligence<sup>sm</sup>



## EWSolutions' Partial Client List

Arizona Supreme Court	Ford Motor Company	Neighborhood Health Plan
Bank of Montreal	GlaxoSmithKline	NORC
BankUnited	Harris Bank	Physicians Mutual Insurance
Basic American Foods	The Hartford	Pillsbury
Becton, Dickinson and Company	Harvard Pilgrim HealthCare	Quintiles
Blue Cross Blue Shield companies	Health Care Services Corporation	Sallie Mae
Branch Banking & Trust (BB&T)	Hewitt Associates	Schneider National
British Petroleum (BP)	HP (Hewlett-Packard)	Secretary of Defense/Logistics
California DMV	Information Resources Inc.	South Orange County Community College
College Board	International Paper	SunTrust Bank
Corning Cable Systems	Janus Mutual Funds	Target Corporation
Countrywide Financial	Johnson Controls	The Regence Group
Defense Logistics Agency (DLA)	Key Bank	Thomson Multimedia (RCA)
Delta Dental	LiquidNet	United Health Group
Department of Defense (DoD)	Loyola Medical Center	United States Air Force
Driehaus Capital Management	Manulife Financial	United States Navy
Eli Lilly and Company	Mayo Clinic	United States Transportation Command
Federal Aviation Administration	Microsoft	USAA
Federal Bureau of Investigation (FBI)	National City Bank	Wells Fargo
Fidelity Information Services	Nationwide	Wisconsin Department of Transportation
		Zurich Cantonal Bank




**For more information on our Strategic Consulting Services, Implementation Services, or World-Class Training, call toll free at 866.EWS.1100, 866.397.1100, main number 630.920.0005 or email us at [Info@EWSolutions.com](mailto:Info@EWSolutions.com)**

[www.EWSolutions.com](http://www.EWSolutions.com)  
 © 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 4

**Strategic Partner & Systems Integrator**  
 Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## What will we talk about?

- Data Models and Data/Information Quality
- What is a Conceptual Data Model?
- Benefits of Conceptual Data Models for Information Quality
- Developing the Conceptual Data Model
- Phased modeling approach (conceptual, logical, physical)
- Conceptual Data Model expressiveness

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 5

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



# Data Models and Quality

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 6

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Information and Data Quality

- Information and Data Quality is a huge issue for every business, government, or institution.
- Poor Information and Data Quality affects every type of information system – OLTP or decision support
- Often leads to a lack of confidence and credibility of IT and IT systems.

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 7

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



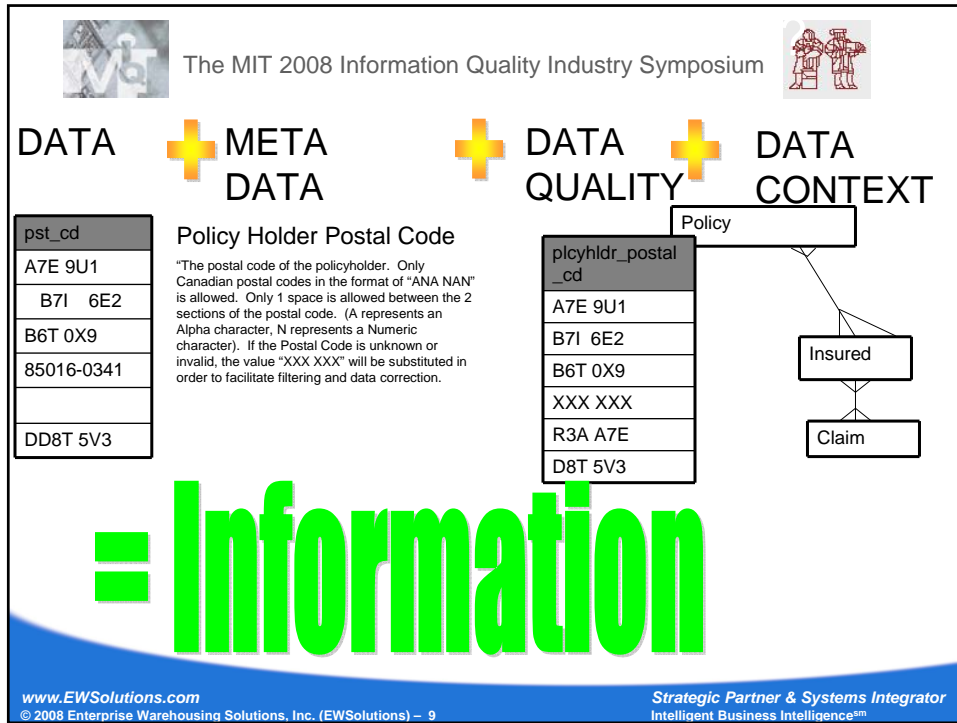
## Information and Data Quality

- What is Data Quality?
  - ➡ Accurate, complete, and valid data that is captured, stored and maintained according to business requirements.
- What is Information Quality?
  - ➡ First, what is the difference between data and information?

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 8

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium

■ Information Quality allows us to ask (and answer with confidence) questions such as?

- ➡ How many unique customers do we have across all lines of business?
- ➡ What geography would be the best to focus on for a new marketing campaign?
- ➡ What are patterns to look for in order to identify a potential disease outbreak?
- ➡ etc, etc, ...

**www.EWSolutions.com**  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 10

**Strategic Partner & Systems Integrator**  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Information and Data Quality

### ■ There are many causes of poor Data Quality

- Lack of system constraints when data is originally captured
- Focus on quantity not quality (let's get these projects done as quickly as possible, and move on to the next thing...)
- Poor data management practices, e.g. authorization, archival
- Programmatic bugs
- Lack of management support for Data Governance and Stewardship
- Data Profiling tool not acquired/used!
- Lack of automated audits and alerts when actual/potential data quality events occur
- Etc....

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 11

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Information and Data Quality

### ■ There are many causes of poor Information Quality

- Stovepiped, independent data marts – different people get different numbers for the same data
- Lack of an integrated Enterprise Data Warehouse, with dependant data marts
- Data not structured in an easy to use format (e.g. Dimensional) that can help prevent misunderstandings
- Users directly querying (e.g. via SQL tools) databases
- Lack of a Managed Meta Data Environment (MME)
  - What does this data mean?
  - Where did it originate from?
  - What were the conditions of the data at the time of the query – e.g. were any loads delayed
- Lack of Data Governance and Stewardship
- Etc...

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 12

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Information and Data Quality

- However, an often overlooked cause of poor information and data quality is:

**Poor or non-existing  
data models**

**Especially Conceptual Data Models!!**



[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 13

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



The MIT 2008 Information Quality Industry Symposium



## Data models and quality

- Data models are often an afterthought or developed only to meet immediate requirements.
- Data Models are often developed by application developers or DBA's – not by Data Architects.
- It is very common (and very bad practice) to see physical data models being the only data model developed for a system. Better practice is to develop a logical model before a physical – but this is still **not BEST practice!!**
- **Physical data models are optimized for performance – NOT for understandability.** Often, foreign key relationships are not utilized in Physical Data Models – making the physical model difficult to understand.

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 14

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



The MIT 2008 Information Quality Industry Symposium



## Data models and quality

- ✚ The physical data model, forward engineered to become the database schema may be in place for years or decades!!!
- ✚ Often much easier to change a program than to change a data model once a system is operational (or even while still in development)
- ✚ Ergo, data models should be developed with due rigor following industry best practices

**Best Practice** is to use a phased modeling approach –  
conceptual, logical, and finally physical models

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 15

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



The MIT 2008 Information Quality Industry Symposium



## Data models and quality

- ✚ What are some of the data and information quality issues that can arise from poor data models?

- ✚ The application does not meet business expectations. Rework often required.
- ✚ The model may meet the immediate needs of the application but may miss the larger needs of the enterprise.
- ✚ M:M relationships may be missed which can lead to significant data duplication/missing data and increased development and maintenance costs
- ✚ Business rules not identified, or not identified well. Business exceptions not identified possibly causing system outages.
- ✚ If cardinality, optionality not properly identified, database constraints may be configured inaccurately leading to data quality problems.
- ✚ If relationship identification not properly captured, granularity may be affected - data not being captured at the detail necessary, other problems.
- ✚ Lack of good business meta data (attributes in business terms, business descriptions, identified data steward, etc)
- ✚ More...



[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 16

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



The MIT 2008 Information Quality Industry Symposium



## Data models and quality

- A bad data architecture practice is developing Physical Data Models without developing Conceptual and Logical Data Models first
- IT needs to “Resist the Urge” to design physical (and logical) data models first.



[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 17

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Resisting the Urge

What does this mean?

- There is a tendency to build physical data models first and ask questions later!!
- Not uncommon to see database schemas being developed in tandem with the application development process
- These models may meet initial requirements but break down when additional requirements and functionality are identified
- **These models often allow or even force Data Quality problems to creep in**
- Need to develop a conceptual data model as the first step of a phased modeling approach and use the conceptual data model as a tool to validate and communicate understanding of business requirements with the business



[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 18

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>





The MIT 2008 Information Quality Industry Symposium



## Causal factors

- Lack of data modeling experience and training
- IT professionals often don't feel productive unless they're "doing something" – e.g. developing a database or writing code.
- Temptation to cut corners when management wants things done yesterday
- Designing and creating databases is fun!! Why did we get into IT but to design and build systems?
- In IT, there are many ways that something can be accomplished – not each way is equal in value



[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 19

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Result

- Systems which may not fully meet business requirements
- Physical structures that may initially be easy to load and query but over time become more difficult to use
- **Poor data quality!**
- Maintenance headaches
- Inflexible for future change
- Longer load cycles
- Etc...



**END RESULT: Unsatisfied customers, increased expense, lack of confidence in IT, etc**



[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 20

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



# Group Exercise

[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 21

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Example

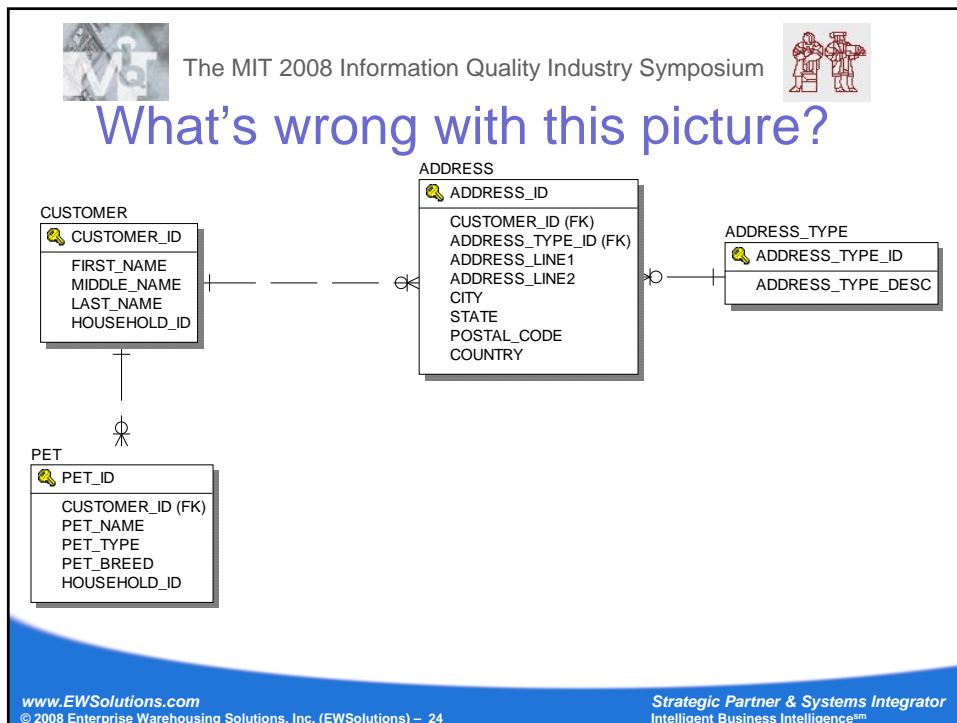
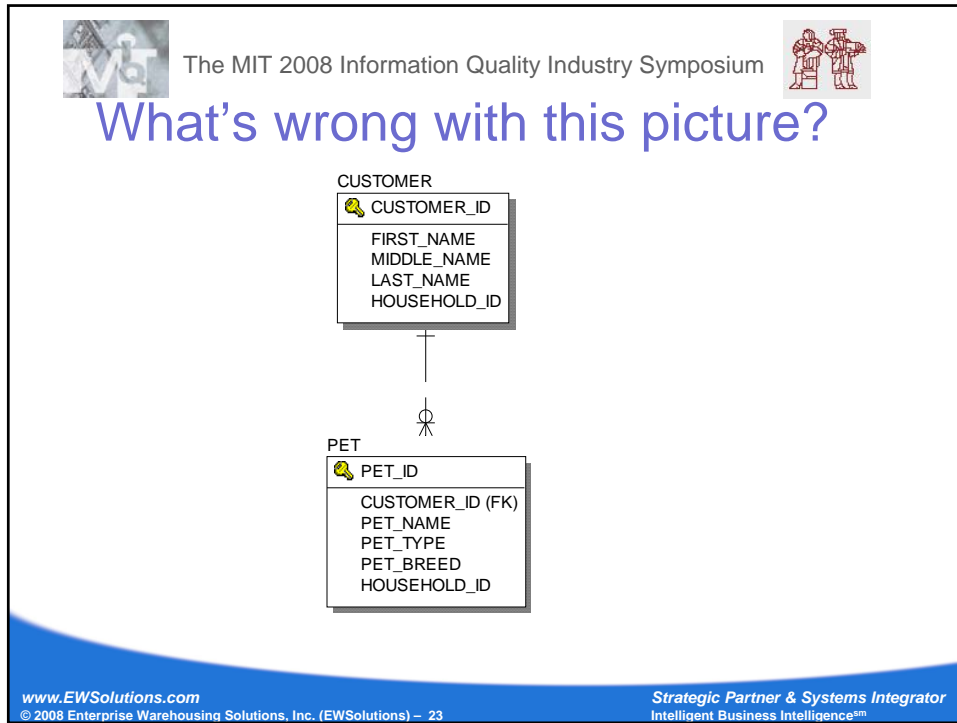
**SCENARIO:**

- A pet hospital chain that performs services and sells products started a CRM (Customer Relationship Management) undertaking and began capturing information about customers and their pets in a CDI (Customer Data Integration) Hub
- Also wanted to track household activity. Last name and address used for determining a household. A household is comprised of 1 or many customers.
- Data to be used for targeted marketing campaigns
- Wanted to be able to track multiple addresses per customer.
- Per business requirements, a Customer had only 1 household id



[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 22

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>





The MIT 2008 Information Quality Industry Symposium



## FACTORS: Example

- Developers assumed they understood the business – they interviewed the customer
- A CDM was not created due many factors such as lack of data modeling expertise and tight deadlines.
- Was incredibly difficult to make changes to the model
- This “proof of concept” required very extensive modification in order for the business to have some confidence in it . It was eventually outsourced!

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 25

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## END RESULT: Example

- Duplication all over the place, requiring unnecessarily complex processing and longer ETL processing windows
- Took heroic effort and a long amount time to adjust the system for changing business requirements – CMM Level 0! ↓
- Excessive maintenance programming
- The business rules had to be enforced primarily in the ETL and SQL and not in the database!
- The poor data model **forced** data quality problems into the system
- The data model didn't fulfill its “enforcement” role – enforcing good data quality through the data model!!

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 26

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Headache

- As the old saying goes “An ounce of prevention is worth a pound of cure”
- Taking additional time up front to understand the business and develop **conceptual data models** helps:
  - Prevent assumptions which lead to data, information quality problems
  - Uncovers “gotchas” that can surface later – fewer “OH SHOOT” moments
  - Reduce development and maintenance costs



[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 27

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



The MIT 2008 Information Quality Industry Symposium



## 7 Habits

- One of the habits in Steven Covey's “**7 Habits of Highly Effective People**” that is commonly quoted is “*Begin with the end in mind*”
- This makes great sense for many things but for good data modeling, start with the beginning in mind with an eye to the end (e.g. to limit scope for the CDM effort)
- *Understand the business first* and finally build physical structures (with many steps and iterations of steps in between)
- Understand the business first by developing a CDM, and review the CDM with the business



[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 28

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



The MIT 2008 Information Quality Industry Symposium



# What is the Conceptual Data Model?



[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 29

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## What is a Conceptual Data Model?

A diagram identifying real world concepts/objects/things (entities) and the relationships between these in order to gain, reflect, and document understanding of the business (as-is & to-be), in order to:

- ➡ foster semantic reconciliation
- ➡ improve business/IT collaboration
- ➡ serve as a framework for the development of information systems



[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 30

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## What is a Conceptual Data Model?

*“A conceptual entity-relationship model shows how the business world sees information. It suppresses non-critical details in order to emphasize business rules and user objects. It typically includes only significant **entities** which have business meaning, along with their **relationships**. “*

*Applied Information Science website*

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 31

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## What is a Conceptual Data Model

**‘A data model that represents an abstract view of the real world. A conceptual model represents the human understanding of a system.... A conceptual data model describes how relevant information is structured in the natural world. In other words, it is how the human mind is accustomed to thinking of the information.’**

**OECD Glossary of Statistical Terms**

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 32

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



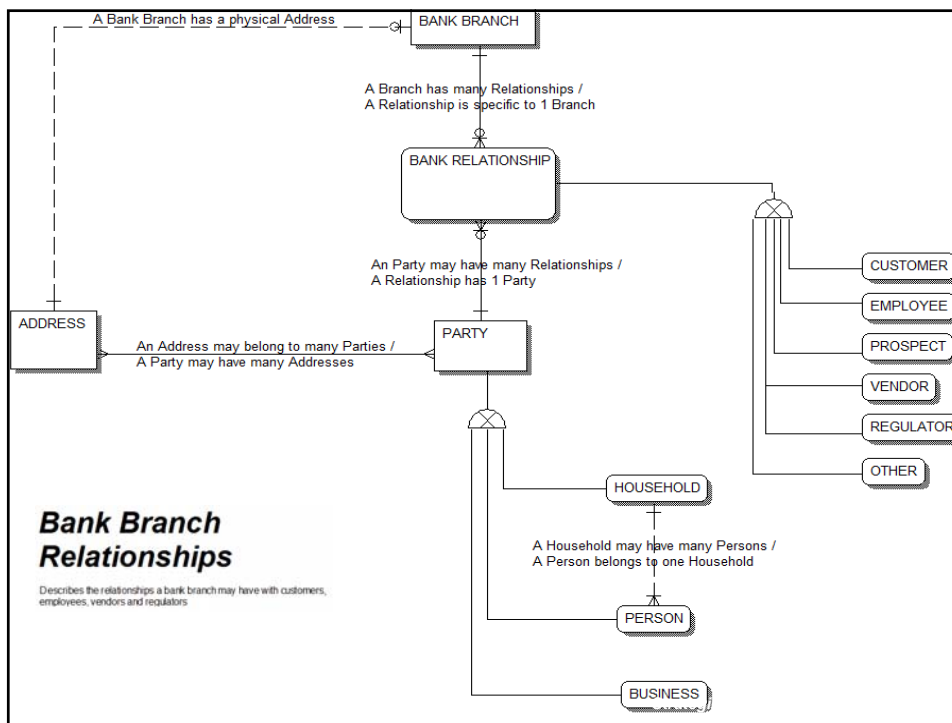
## What is a Conceptual Data Model?

- It is “stateless” - NOT a state model
- The entire possible lifecycle of a relationship should be represented, per current business practice
  - This includes business exceptions!!
  - *Not exceptions due to poor data quality or due to system limitations)*
  - The CDM should reflect the business – not IT systems
  - Review optionality and cardinality to ensure longitudinal perspective

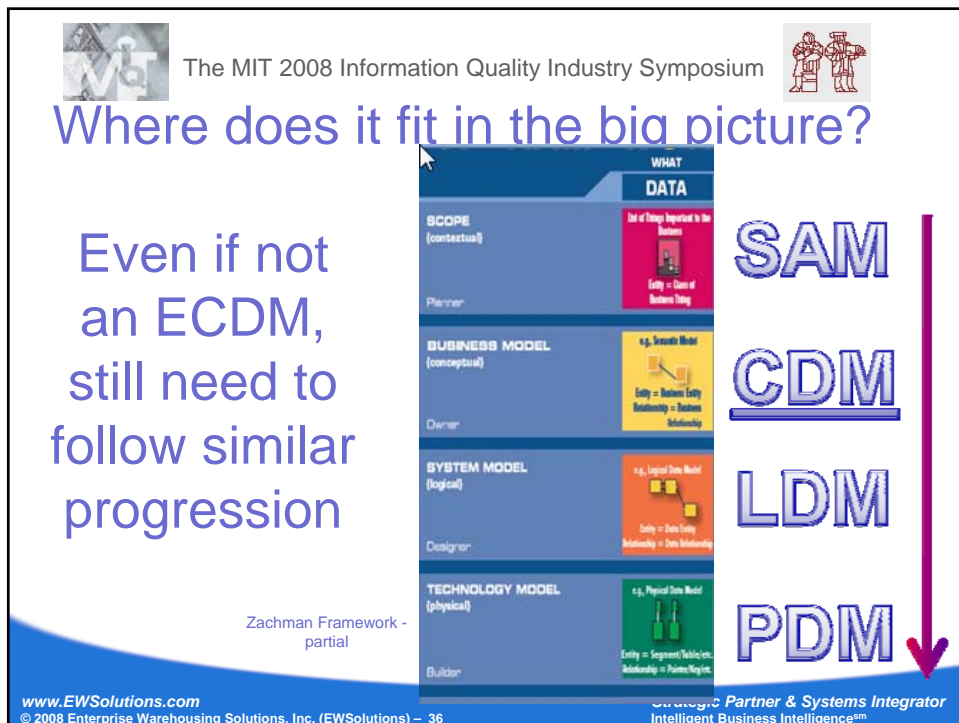
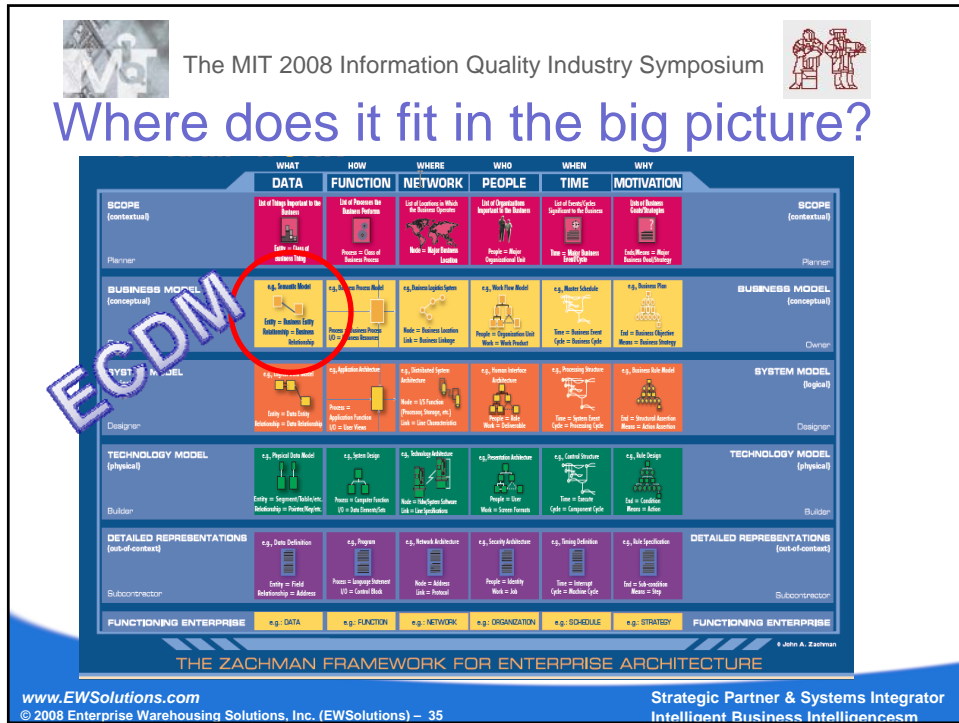
www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 33

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™









The MIT 2008 Information Quality Industry Symposium



## Semantic Resolution

- A CDM is a key tool for semantic resolution
- For enterprise applications, have to reach consensus across divisions, departments, external agencies, etc, for naming and defining data entities, and identifying correct relationships.
- Semantic resolution is a key activity of Data Governance and Stewardship, and an ECDM is a key enabler of Data Governance and Stewardship – these activities often take place in tandem, iteratively
- **Difficult to have Information Quality if synonyms, homonyms haven't been resolved.** *E.g. Is a customer a party that has placed an order, or can customer be a party who placed an order or a party that might become a paying customer?*

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 37

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



The MIT 2008 Information Quality Industry Symposium




## Semantic Resolution

- Due to fundamental differences with the LDM, the CDM often has to be contained in a separate model file and **so there is a risk that lineage from a logical entity to a conceptual entity can be lost**
- Be sure to save the association between conceptual and logical entities, logical and physical entities, etc using:
  - ➡ A meta data repository and related tool which can be used to establish these relationships
  - ➡ User defined meta data properties within the model
  - ➡ Spreadsheet, etc. Last resort
- CDM's can help drive creation of a common, corporate lexicon – fostering improved communication, standardization --- **BENEFICIAL TO THE ENTIRE ENTERPRISE – NOT JUST IT!**


[www.EWSolutions.com](http://www.EWSolutions.com)

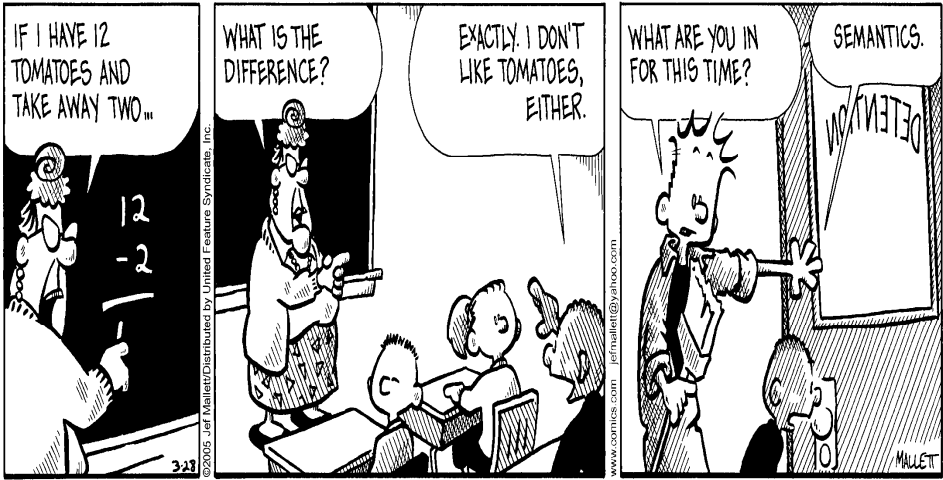
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 38

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



The MIT 2008 Information Quality Industry Symposium





IF I HAVE 12 TOMATOES AND TAKE AWAY TWO...

WHAT IS THE DIFFERENCE?

EXACTLY. I DON'T LIKE TOMATOES, EITHER.

WHAT ARE YOU IN FOR THIS TIME?

SEMANTICS.

DETERM...

323


©2005 Jeff Mallett/Distributed by United Feature Syndicate, Inc.

www.cartists.com jeffmallett@yahoo.com


MALLET

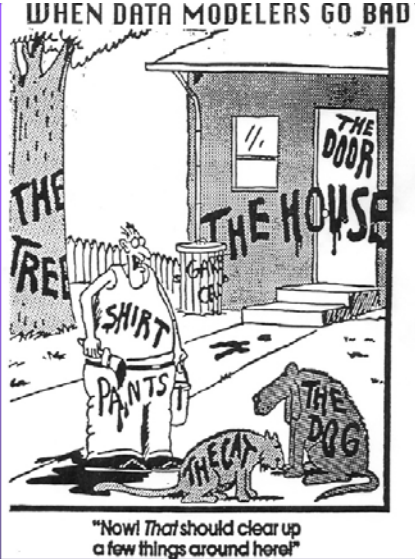
[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 39

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>SM</sup>



The MIT 2008 Information Quality Industry Symposium





WHEN DATA MODELERS GO BAD

THE TREE

THE HOUSE

THE CAT

THE DOG

"Now! That should clear up a few things around here!"

Gary Larson – The Far Side

[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 40

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>SM</sup>



The MIT 2008 Information Quality Industry Symposium



# Developing the Conceptual Data Model



www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 41

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™

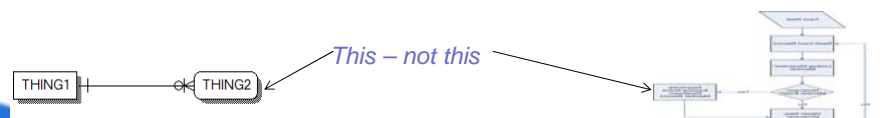


The MIT 2008 Information Quality Industry Symposium



## Getting started developing a CDM

- A major hurdle is separating “data thinking” vs “process thinking”
- For conceptual data modeling, we’re thinking about “**what**” (data) not the “**how**” (process).
- For a CDM – data is a relative term
- Data may not exist currently for a conceptual entity – but entities must be included in the CDM if it is an object of importance to the business



www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 42

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



The MIT 2008 Information Quality Industry Symposium



## Getting started developing a CDM

- When interviewing the business helpful to use a “recipe” analogy (see Steve Hoberman design challenge \*) . A recipe identifies the **ingredients, utensils, equipment (whats)** and has **directions (hows)** in order to meet the desired goal.
- If the interviewee focuses on process ask “What things are needed for the XYZ process?” “What are the components of the XYZ process?”
- Helpful starting place is to identify “nouns”,  
e.g. Customer, Product, Inventory



\* DMReview January 2008, quoting Geof Clark

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 43

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## When is a CDM finished?

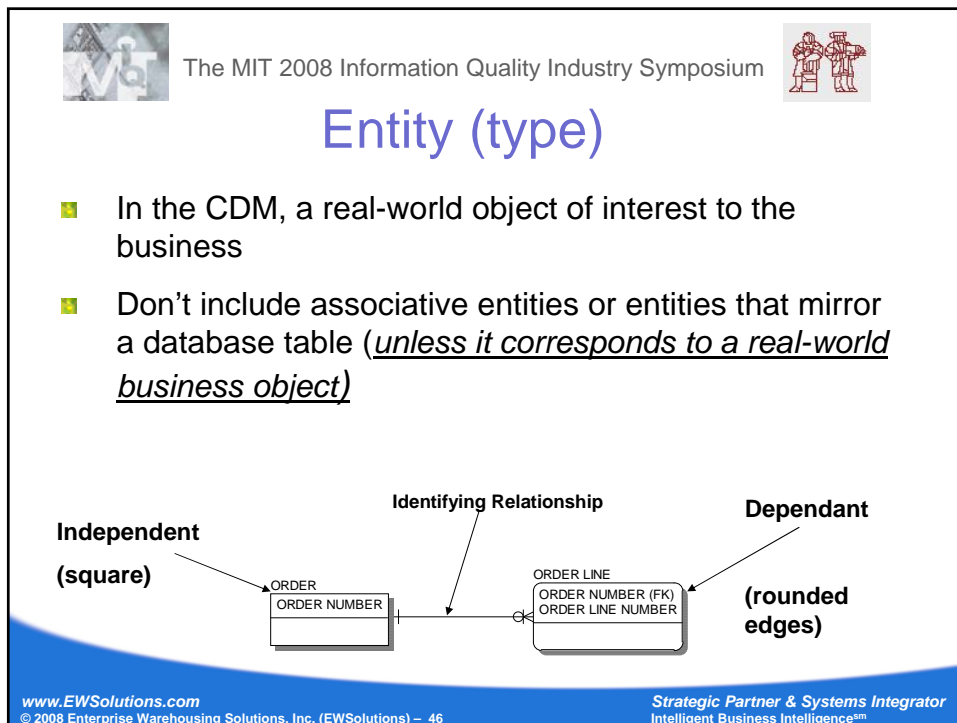
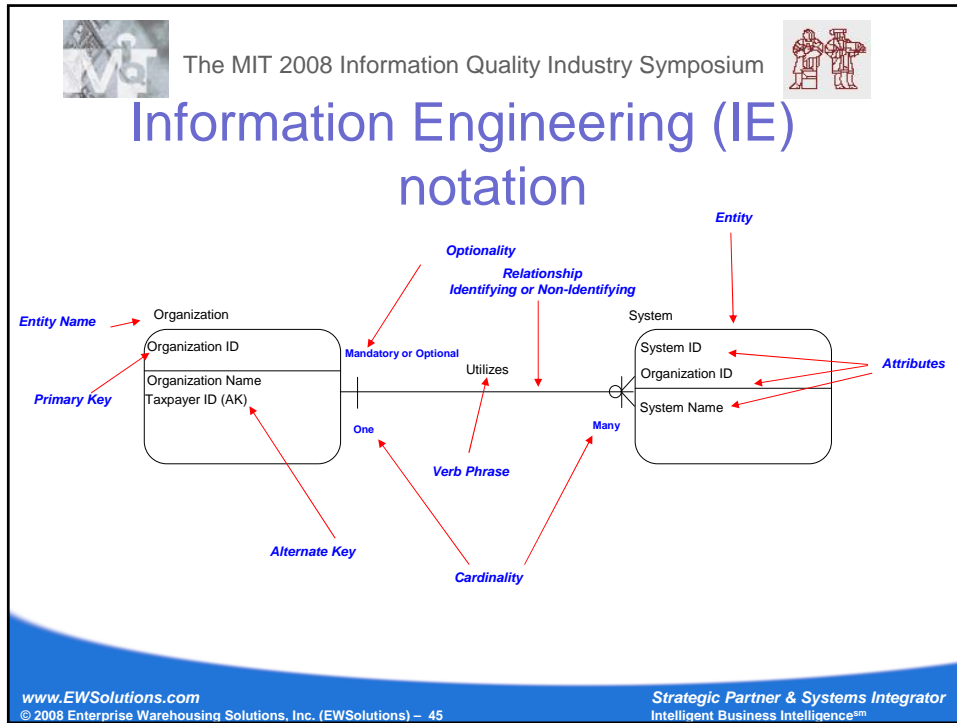
**“Perfection does not come into being,  
when nothing more can be added, but  
when nothing can be taken away”**

Antoine de Saint-Exupéry

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 44

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>





The MIT 2008 Information Quality Industry Symposium



## Identification in Relationships

—————

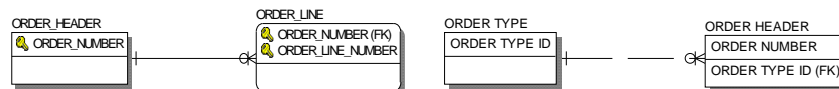
An identifying relationship is stronger -- helps determine the meaning and granularity of a child entity. Is always mandatory. **(NOTE: a solid line in a M:M relationship does not denote an identifying relationship!!)**

-----

A non identifying relationship may be mandatory or optional, but does not define meaning/granularity

Identifying Relationship

Non-identifying Relationship



www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 47

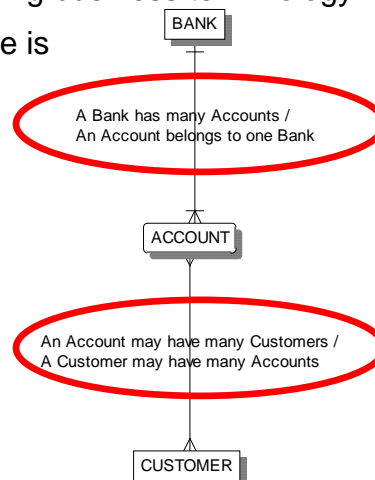
Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>

The MIT 2008 Information Quality Industry Symposium



## Relationship Verb Phrase

- Describes the relationship using business terminology
- Can be terse but IMO verbose is better
- You never know who will end up looking at the model!!!
- Business people probably won't take the time to understand the notation!



www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 48

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>

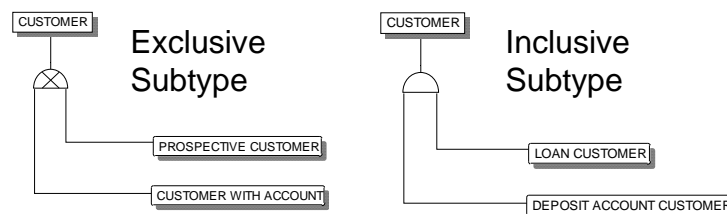


The MIT 2008 Information Quality Industry Symposium



## Subtyping

- Subtypes/Supertypes make a model more expressive and understandable
- Subtypes describe a Supertype
- A Subtype can be inclusive or exclusive, exhaustive or non-exhaustive



**More on subtyping later**

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 49

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Relationships and quality

### Many to Many Relationships

- A conceptual data model will very often have numerous M:M relationships in order to accurately reflect all possible states of a relationship
- A CDM is not a state model – it should reflect the relationship from a longitudinal (entire lifecycle of the relationship) perspective
- For example, a store clerk works for one store in almost all cases, but it is possible for a clerk to move and begin work with another store.

Clerk:Store s/b a M:M

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 50

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>





The MIT 2008 Information Quality Industry Symposium



## Relationships and quality

### Many to Many Relationships

- When a M:M relationship is not identified during requirements definition in a CDM....
  - Project scope is not measured correctly
  - Logical model design, application development, testing are all impacted – heavily!!!
  - Some M:M instances occur only occasionally – can cause bugs, outages, missed or duplicate data weeks/months later when exceptions are encountered

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 51

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Relationships and quality

### Many to Many Relationships

- Impacts
  - Logical and physical models have to be revisited, reviewed, and possibly reapproved
  - Can have a tremendous impact on applications – screen forms, program functions, load processes, reports, SQL, cubes, etc..
  - Existing data may need to be restructured
  - Impacts to downstream systems (DW/BI, ODS, MDM, etc)

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 52

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Relationships and quality

- Resolving M:M relationships involves non-trivial decisions, with benefits and impacts to weigh. Not something to decide during a 3 am support call.....
- Resolution decision can have a dramatic impact on quality
- If you choose not to allow M:M relationship in a particular instance – how are you going to impact the business near or long term?



[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 53

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Relationships and quality

- A major bank allowed multiple customers to apply for a single loan
  - However only the information for the 1<sup>st</sup> customer was retained (e.g. identifying information, credit score) in the system.
  - Bank had **no accurate idea how many customers it had**, and could not easily and accurately gauge the total customer experience.....
  - Might not know if it was marketing a new loan to a customer who had defaulted on a prior loan...



[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 54

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Relationships and quality

- Model relationships are a key source for of business rules and data quality metrics.
- For non-kernel entities, **identifying relationships** will be critical to understanding entity meaning and granularity – be sure to distinguish identifying / non-identifying
- **CAN HAVE A DRAMATIC IMPACT ON UNDERSTANDING (or misunderstanding) THE MODEL!!**

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 55

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Relationships and quality

- **If you want your application to be successful...**

**Data relationships must be  
correctly identified!!!!**

- **The only question is: when are you going to pay  
to discover the correct relationship???**

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 56

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Relationships and quality

- You can identify the correct relationships



In the CDM phase (during requirements definition)



During Logical Data Model development



During Physical Data Model development



During application development



During implementation



During production

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 57

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## CDM and requirements

- Some statistics...
- If it costs \$1 to fix a defect found in the requirements phase, it costs \$2 in design phase, and continues to rapidly increase until it costs \$68 if not found until product is released into operation - *Boehm, Barry W. Software Engineering Economics. Englewood Cliffs, NJ: Prentice-Hall, 1981*
- Requirements errors account for 70 to 85 percent of the rework cost - *Leffingwell 1997, quoting Barry W. Boehm*
- The cost to fix the defect in QA stage is eight times more than during the Requirements Development stage - *Grady 1999*

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 58

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## CDM and requirements

- Conceptual data modeling should take place in the Requirements Definition phase
- Conceptual data modeling (in general) is NOT design – it is description
- Modeling the BUSINESS – not a SOLUTION

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 59

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Data Modeling Progression

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 60

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Model progression

### ■ Conceptual Data Model (CDM)

#### ➡ Technology and application neutral

- ➡ Entities may or may not eventually translate into a physical database table
- ➡ **A data source for a conceptual entity does not need to exist!!** Only interested in understanding the business at this point
- ➡ **Physical implementation is NOT important at this point** – conceptual data modeling is all about documenting business objects. Set expectations appropriately when presenting to technologies personnel.

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 61

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Model progression

- If the CDM is wrong, your downstream models may be built upon incorrect premises!!!
- Don't shortchange the amount of time spent in this step!!
- **NEVER a waste of time!!** *At the very least you can justify it as a tool for yourself for developing LDM's – who can remember all the identification, cardinality, optionality of even a moderately complex subject area?*



[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 62

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



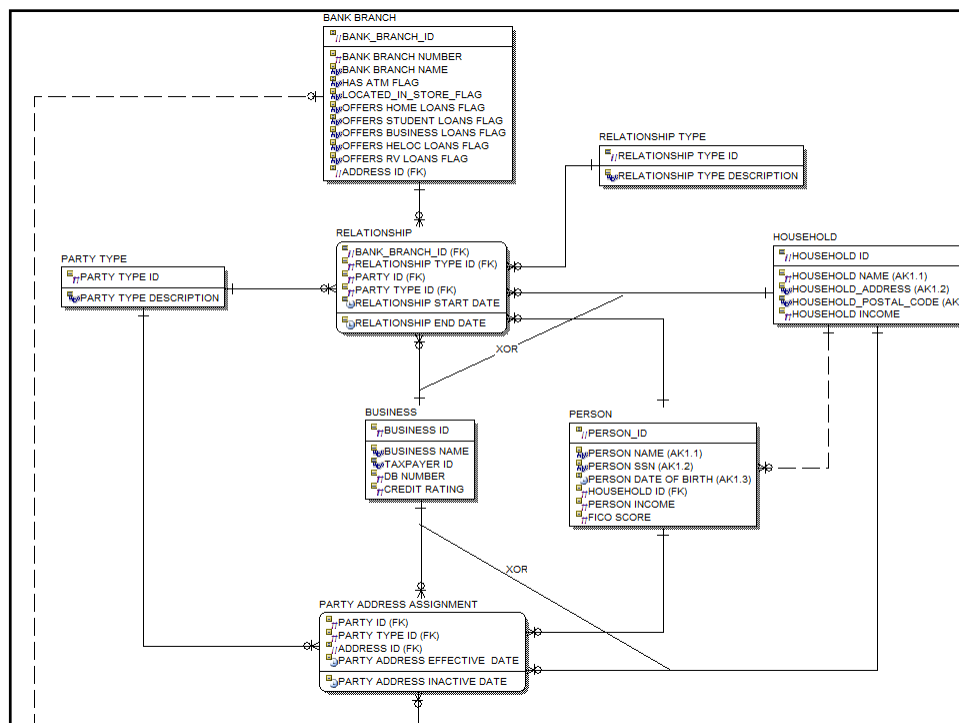
## Model progression

### ■ Logical Data Model (LDM)

- First step of SOLUTION DATA DESIGN (generally)
- Fully/mostly attributizes a conceptual data model
- Resolves many to many relationships (usually)
- Resolves subtypes/supertypes (usually)
- May introduce abstraction (generalize entities, attributes, relationships) – more later
- Formalize primary keys

www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 63

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Abstraction in the CDM

- *“Abstraction is the removal of details in such a way as to broaden applicability to a wide class of situations while preserving the important properties and essential nature from concepts or subjects” \**
- In the CDM, generally avoid abstraction in order to more closely mirror the business.
- Use supertypes when you need to abstract – for establishing broad applicability relationships (in order to avoid establishing relationships to all the subtypes)

\* Steve Hoberman – Data Modeling Made Simple

www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 65

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™

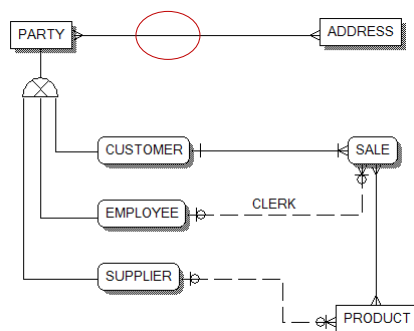


The MIT 2008 Information Quality Industry Symposium

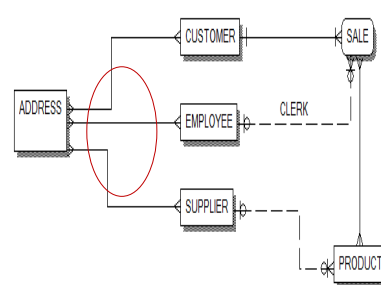


## Abstraction in the CDM

### With abstraction..



### Without....



Now add 20+ more types of parties (e.g. insurance)  
more relationships, ....

www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 66

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™





The MIT 2008 Information Quality Industry Symposium



## Abstraction in the LDM

- In the LDM, abstraction is necessary for normalization – data stored only once
- Entities, attributes, relationships can be abstracted
- Allows for flexibility in case other types need to be added in the future

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 67

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Model Progression

- Physical Data Model (PDM)
  - ◆ Represents how a logical model is applied to a particular DBMS platform
  - ◆ Assign datatypes, indexing, storage, partitioning, etc
  - ◆ Can be forward engineered to create the actual database structures
  - ◆ Complies with DBMS nomenclature restrictions
  - ◆ PDM may look different than the logical – e.g. column ordering to take advantage of partition elimination

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 68

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Model Progression

### ■ Why develop all these models?

- Follows the progression in which a Data Modeling project should be undertaken
- As more information becomes known, the more depth the models will be able to convey
- Data Models convey knowledge – and knowledge is retained in data models

www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 69

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™

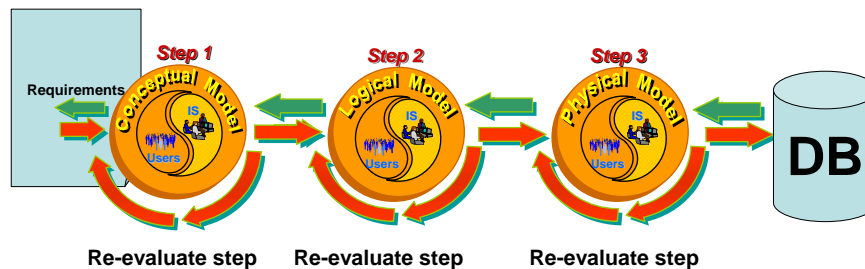


The MIT 2008 Information Quality Industry Symposium



## Phased Modeling Approach

- **Defined within the scope of the business problem**
- Data Modeling is an iterative endeavor – and it will probably be necessary to make revision to upstream models as more information becomes available



www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 70

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



The MIT 2008 Information Quality Industry Symposium



# Conceptual Data Model Expressiveness

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 71

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>

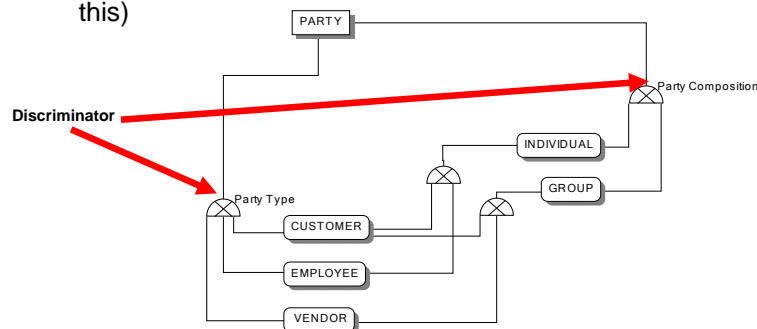


The MIT 2008 Information Quality Industry Symposium



## Subtyping

- An entity may have many subtype relationships - use a discriminator to distinguish the subtype relationship
- A Subtype may have more than one Supertype (not every tool allows this)



[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 72

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Subtyping

- **A single Subtype entity may have relationships with entities which do not apply to the other Subtypes**
- This helps the CDM to better mirror business reality.
- **Additional business rules can be expressed!!**
- When entity abstraction occurs during the LDM phase – entities and relationships are “lost”. The identification, cardinality, and optionality of these relationships is subsumed into the remaining relationships

**NOTE: abstraction is a very useful and valid tool for the logical design phase - critical for normalization to eliminate redundant data. However, business semantics and rules are harder to identify – especially by a business person.**

**www.EWSolutions.com**

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 73

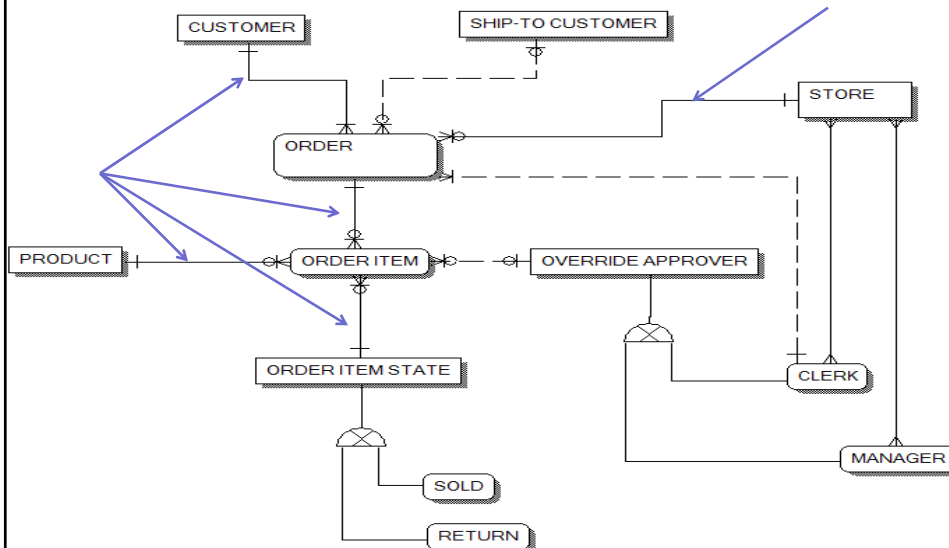
**Strategic Partner & Systems Integrator**  
Intelligent Business Intelligence<sup>sm</sup>

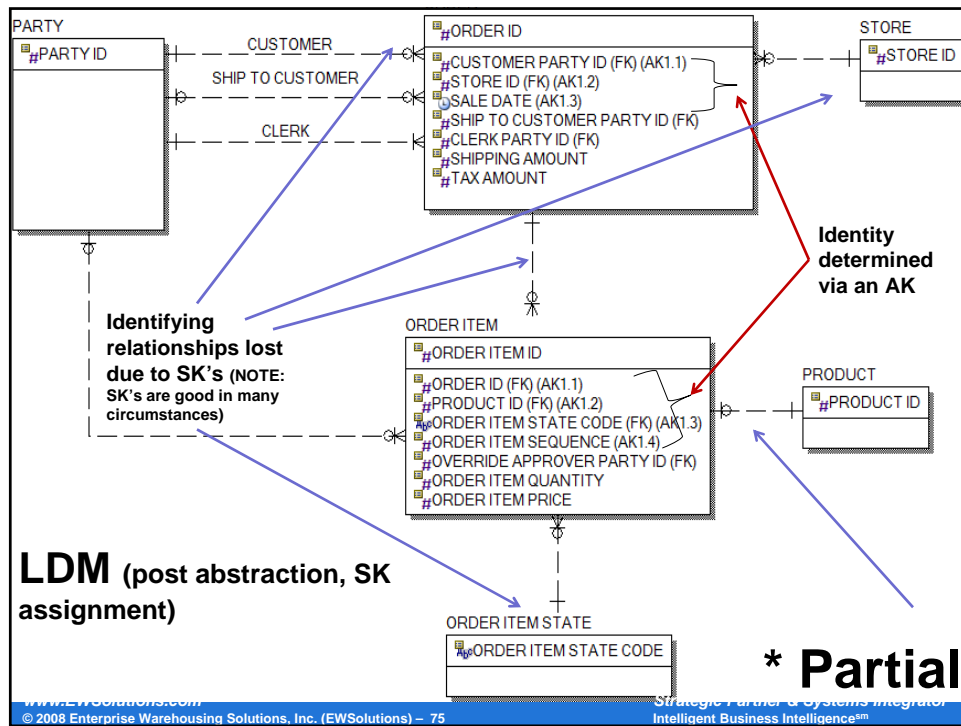


The MIT 2008 Information Quality Industry Symposium



## CDM (pre abstraction – note identifying relationships)






The MIT 2008 Information Quality Industry Symposium




## CDM Expressiveness

- The previous model, which was a normalized logical data model, partially based off of the earlier CDM, is a solution model – not a business model.
- Identifying relationships are lost due to the surrogate key assigned to the “sale” entity
- Business rules (relationships) are still there, but aren't as obvious – it isn't modeled how “the human mind is accustomed of thinking about information”

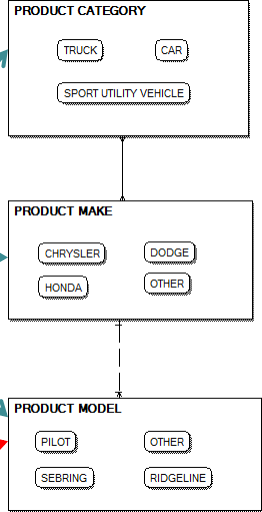


The MIT 2008 Information Quality Industry Symposium



## Subtyping and Taxonomies


- Establishing subtypes in a CDM is often the first step in developing taxonomies for classifying data
- Provides value domain for a taxon (full or partial)
- Makes abstract names more understandable




Framed subtyping  
(Euler diagram)

www.EWSolutions.com  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 77

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Wrapping it up

- There are many causes of poor Data and Information Quality
- Poor or non-existing CDM's are not the least of these causes
- Need to "Resist the urge" to develop physical (and logical) models before developing conceptual models
- Many business requirements and rules are captured and documented in the CDM
- The CDM is a key means to validate IT's understanding of business requirements, and can be used to measure data quality in implemented systems
- CDM's need to be validated by the business
- CDM's need to be presentable, understandable and tailored to the audience

www.EWSolutions.com  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 78

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## Wrapping it up

**Developing CDM's first is beneficial to your organization!!**

- Models downstream from CDM's more accurately reflect business requirements
- Fosters semantic resolution, in turn improving Information Quality
- Relationship identification, cardinality, and optionality are critical to good Data Quality
- Development and maintenance work is simplified and costs are reduced. **Once a system goes into production, it is very hard to change data structures!**



[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 79

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



The MIT 2008 Information Quality Industry Symposium



## References

- Applied Information Science website
- OECD Glossary of Statistical Terms
- Zachman Framework for Enterprise Architecture
- Steve Hoberman Design Challenge, DMReview January 2008, quoting Geof Clark
- Boehm, Barry W. Software Engineering Economics. Englewood Cliffs, NJ: Prentice-Hall, 1981
- Steve Hoberman, "Data Modeling Made Simple" Technics Publications, 2005

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 80

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>




The MIT 2008 Information Quality Industry Symposium




# Questions?

www.EWSolutions.com  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 81

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



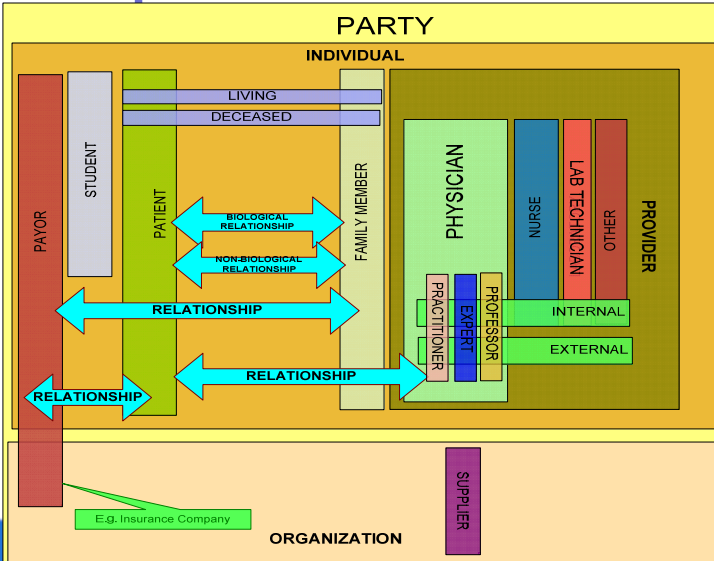
The MIT 2008 Information Quality Industry Symposium



## Graphic Model

Sample graphical representation of a CDM regarding the Party subject area in a medical educational institution

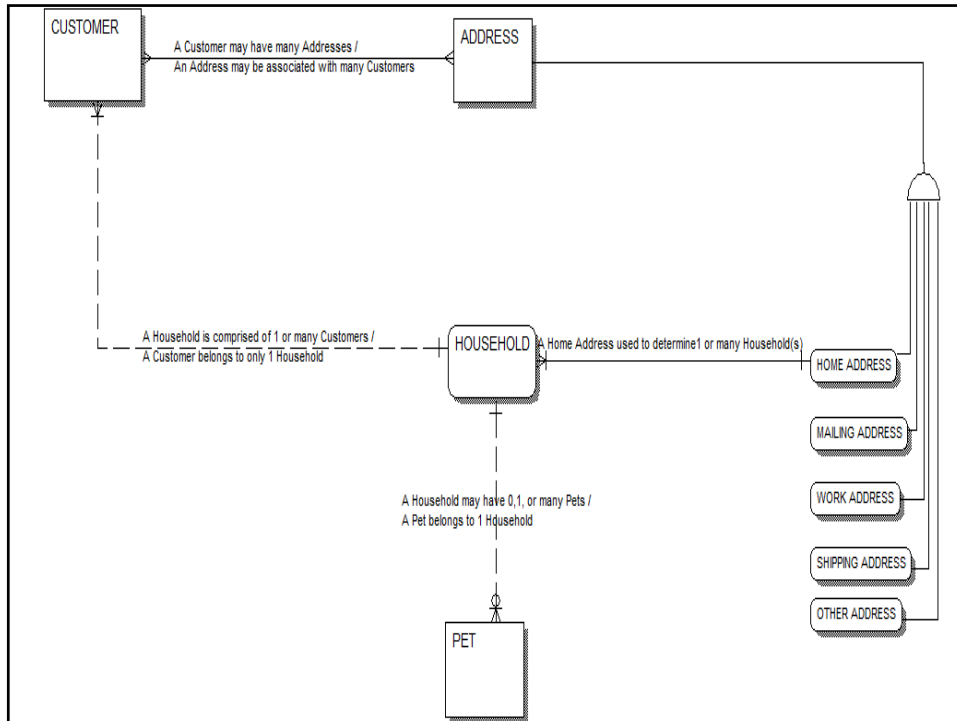
**Which are inclusive subtypes and which are exclusive subtypes?**



The diagram illustrates a 'PARTY' hierarchy. At the top is 'PARTY', which branches into 'INDIVIDUAL' and 'ORGANIZATION'. 'INDIVIDUAL' further branches into 'PAYER', 'STUDENT', 'PATIENT', 'FAMILY MEMBER', and 'PROVIDER'. 'PATIENT' is further divided into 'LIVING' and 'DECEASED'. 'PROVIDER' is divided into 'PHYSICIAN', 'NURSE', 'LAB TECHNICIAN', and 'OTHER'. 'PHYSICIAN' is further divided into 'PRACTITIONER', 'EXPERT', and 'PROFESSOR'. 'PROVIDER' is also divided into 'INTERNAL' and 'EXTERNAL'. 'ORGANIZATION' includes 'E.g. Insurance Company' and 'SUPPLIER'. Relationships are indicated by blue double-headed arrows: 'PATIENT' to 'FAMILY MEMBER' (labeled 'BIOLOGICAL RELATIONSHIP' and 'NON-BIOLOGICAL RELATIONSHIP'), 'PATIENT' to 'PROVIDER' (labeled 'RELATIONSHIP'), 'PAYER' to 'PATIENT' (labeled 'RELATIONSHIP'), and 'PAYER' to 'ORGANIZATION' (labeled 'RELATIONSHIP').

www.EWSolutions.com  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 81





The MIT 2008 Information Quality Industry Symposium



**EWSolutions, Inc.**  
**15 Spinning Wheel Road,**  
**Suite 330**  
**Hinsdale, IL 60521**  
**Office 630.920.0005**  
**Fax 630.920.0008**

<http://www.EWSolutions.com>



The MIT 2008 Information Quality Industry Symposium



# Improving your Data Warehouse's IQ



Derek Strauss  
Gavroshe USA, Inc.



The MIT 2008 Information Quality Industry Symposium



## Outline

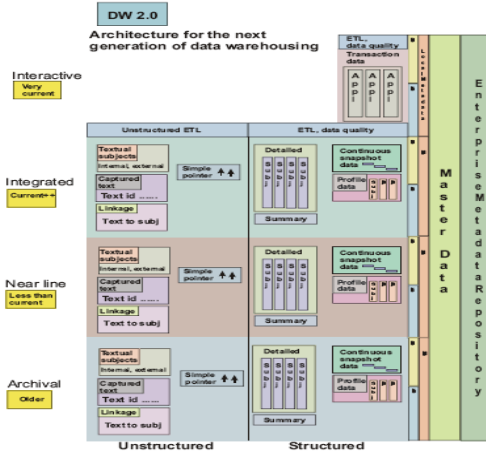
- **Data quality for second generation data warehouses**
- **DQ tool functionality categories and the data quality process**
- **Data model types across the DW2.0™ database landscape**
- **Challenging top-down from the bottom**
- **Deriving an interlocking set of models**



The MIT 2008 Information Quality Industry Symposium



## DW2.0™ – Architecture for the next generation of data warehousing



DW 2.0 is a trademark of Bill Inmon. All rights reserved

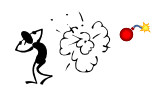

© "Architecture for the next generation of data warehousing" is copyrighted by Bill Inmon, 2006



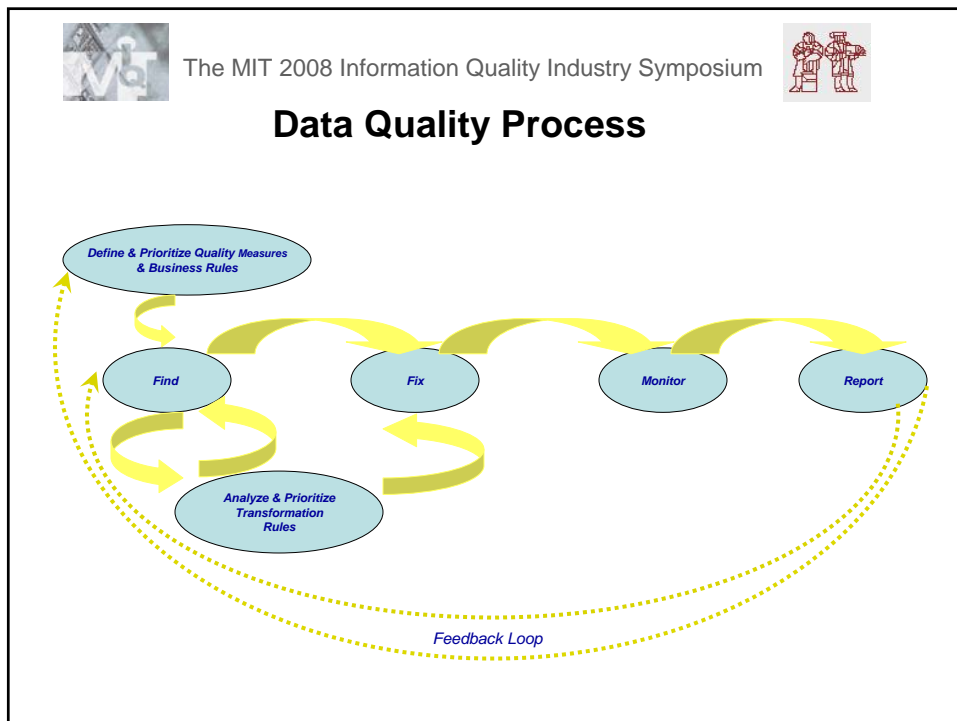
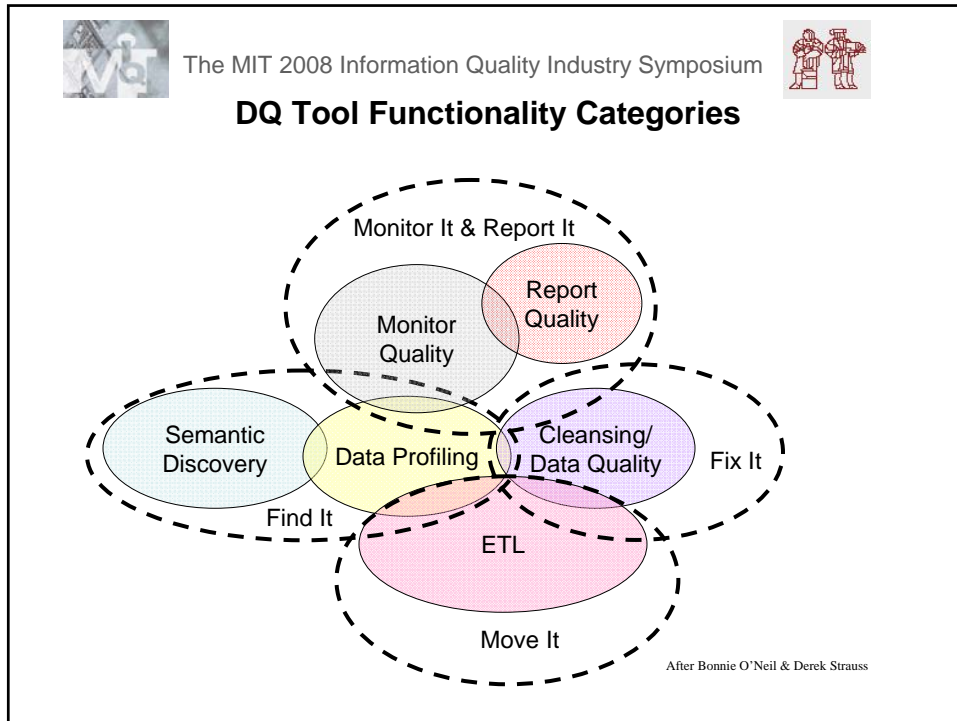
The MIT 2008 Information Quality Industry Symposium



## Data Quality for second generation data warehouses

- Getting away from "code, load and explode" 
- The "data scouts" (team with business and IT representation) 
- On finding significant DQ problem, choose from strategies such as:
  - fix the data at the source (actually go into the data store and physically zap the data)
  - fix the program the source (apply the correct edits in order to validate the data)
  - fix the business process (a broken business process is very often the main cause of poor quality data)
  - recognize and resolve situations where data attributes are being used for a purpose other than their original intent (e.g. a gender code, which has more than two distinct values)
  - transform the data on the way into the data warehouse (this is the most common of strategies, but should not be the only strategy employed)\*

\* In the case of the latter strategy, it is important to note that there are two alternative implementations for transforming data on the way into the integrated sector. The first implementation is to simply change the data and load it into the warehouse. The second implementation scenario does that and more: it will actually load the unchanged data alongside of the changed data. There are many times when this may be a preferable route to go.





The MIT 2008 Information Quality Industry Symposium



## DATA PROFILING TOOLS AND THE REVERSE ENGINEERED DATA MODEL.

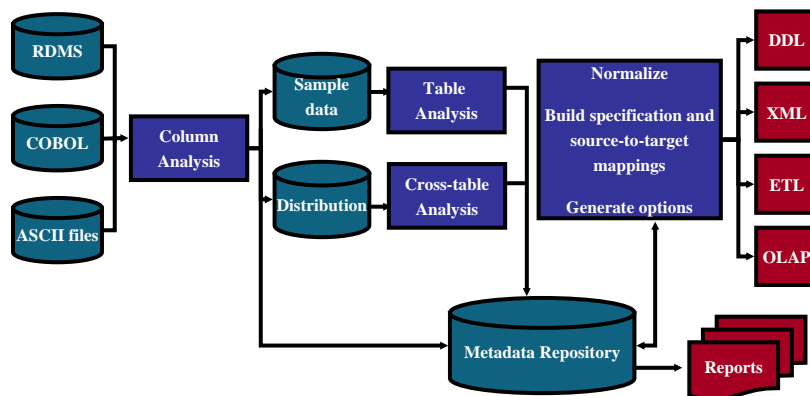
- Today there are many data profiling tools to assist the data quality team.
- The tools facilitate analysis of the data values held in a column; sometimes looking simultaneously at multiple columns in a table; sometimes even looking across tables, and even across systems to see if there are any patterns in the values held in the selected columns.
- These patterns can uncover hidden business rules, e.g. every time the value in column 1 is "a" we see that the value in column 5 can be "x" or "y".
- The best of these data profiling tools will go one step further: having analyzed the actual data values in the columns of a system, the tools can suggest a normalized schema.
- What in effect happens is the tool develops a third normal form data model, based on bottom up analysis, abstracted from the actual data values in the physical database.
- This abstracted data model is very useful input into the top down modeling process, which should be happening in parallel with the development of the warehouse.
- In fact, in the case of a DW 2.0 warehouse, we want to ensure that a high quality data model architecture is being developed, as this will greatly assist in improving the data quality of the enterprise data warehouse.



The MIT 2008 Information Quality Industry Symposium



## Automated Methods

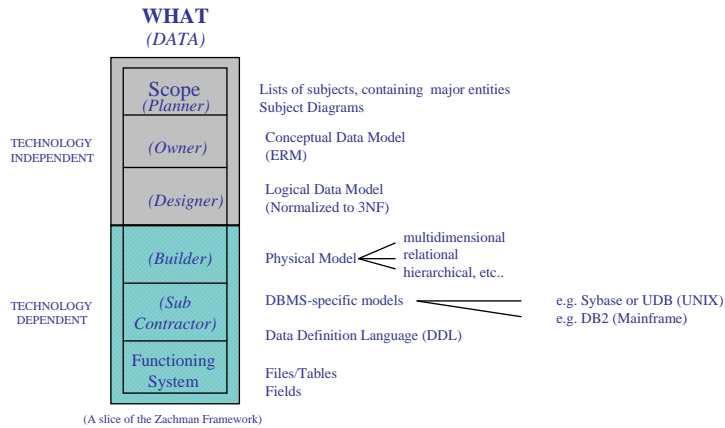




The MIT 2008 Information Quality Industry Symposium



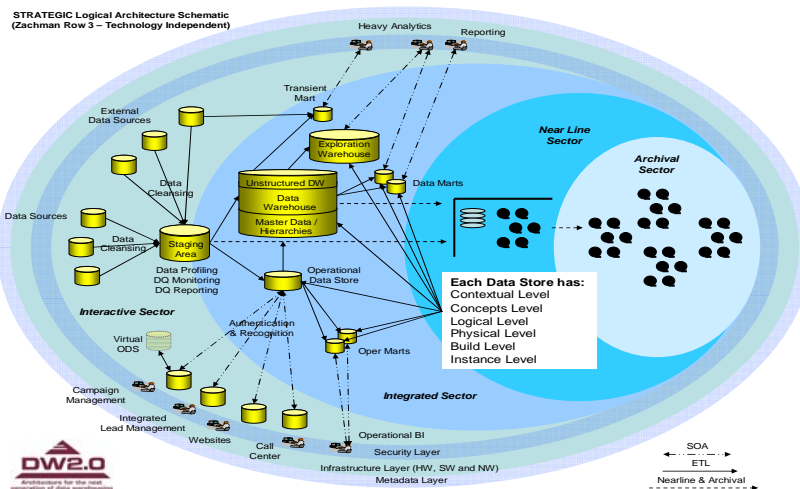
## Data Model Types



The MIT 2008 Information Quality Industry Symposium



## The DW2.0 Database Landscape





The MIT 2008 Information Quality Industry Symposium



## DATA MODELS ACROSS THE FOUR DW2.0 SECTORS

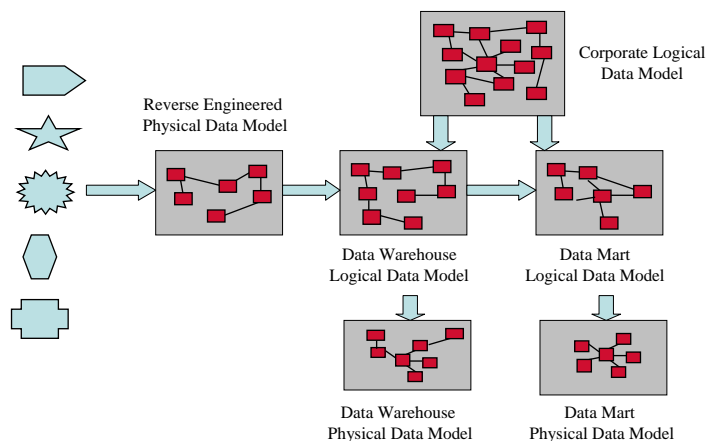
- A good concepts data model helps you to understand the major concepts in the business and how they interrelate.
- A good third normal form logical data model helps you understand all the attributes pertaining to the business entities and also the cardinality and optionality of the relationships between those entities. This model gives a great logical view of the business and its data and should be the starting point for the third model type – i.e. the physical data model.
- Physical data models for DW 2.0's integrated sector can differ widely in their structure. They will range from normalized and near normalized models for the data warehouse hub through to star schema and snowflake schema models for the data marts. Still other structures would be best suited for exploration warehouses, data mining warehouses, operational data stores, and opermarts.
- Data moving to the nearline sector should be kept as close to third normal form structure as possible; it is normal for data to be restructured as it enters the archival sector. It is important to the DW 2.0 world that there should be multi-directional traceability between these models: it should be possible to navigate from a physical model back up through the logical model and up still further to the concepts model; in like manner, we should be able to move from the top down from a concepts model to the logical data model and on to the physical data models.
- A rigorous set of interlocking models will go a long way towards improving the quality of the data in the enterprise, linking business meaning and structural business rules to physical instantiations of data entities and attributes.

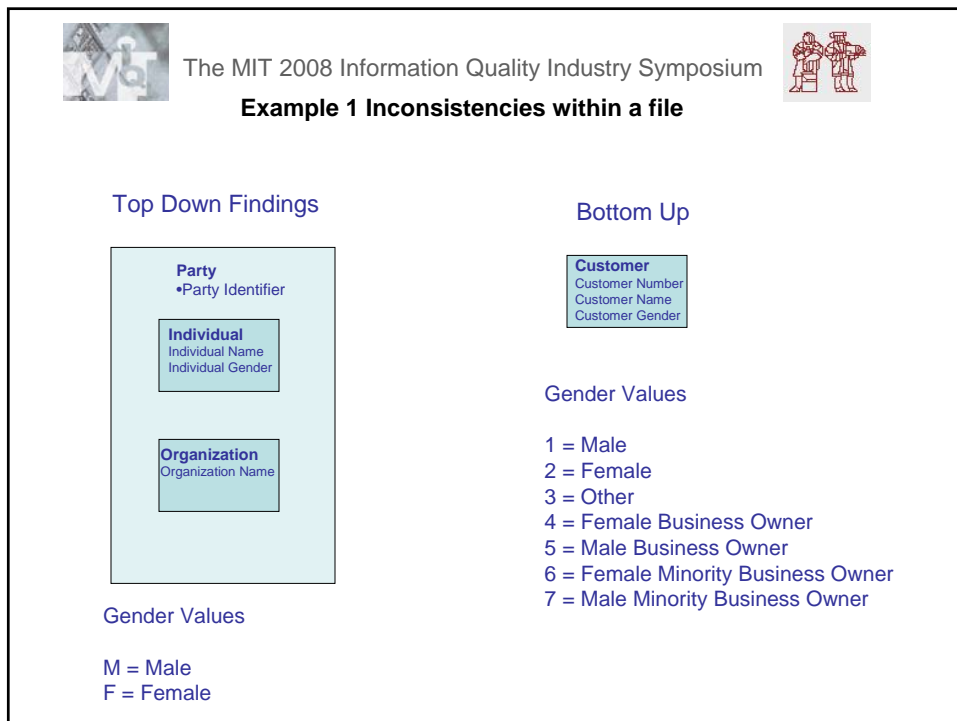
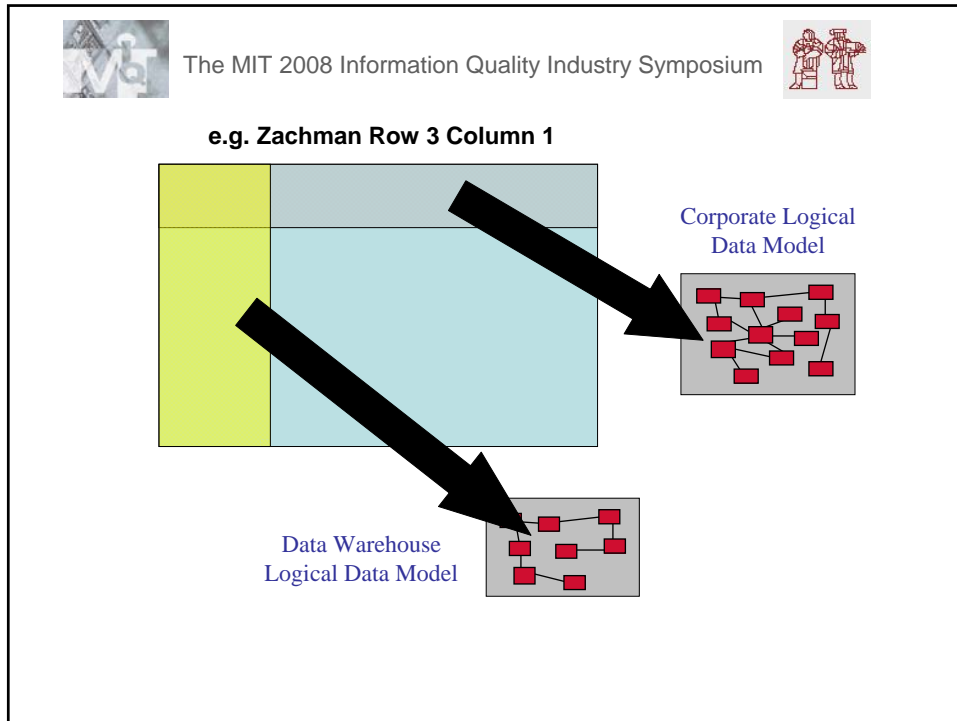


The MIT 2008 Information Quality Industry Symposium



## Data Models Needed







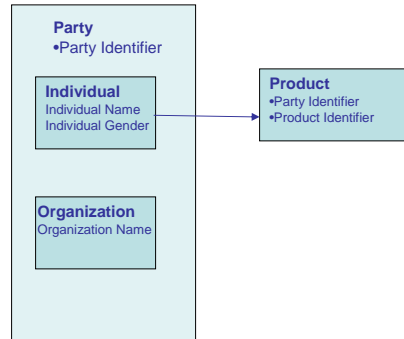


The MIT 2008 Information Quality Industry Symposium



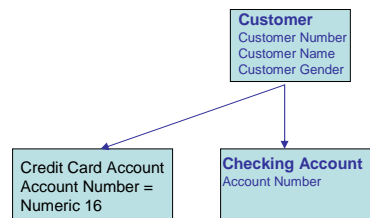
## Example 2 Inconsistencies across files

### Top Down Findings



Product Identifier  
Is a numeric 10  
E.G 0056793214

### Bottom Up



Checking Account Number =  
Numeric 10  
E.G. 0123456789

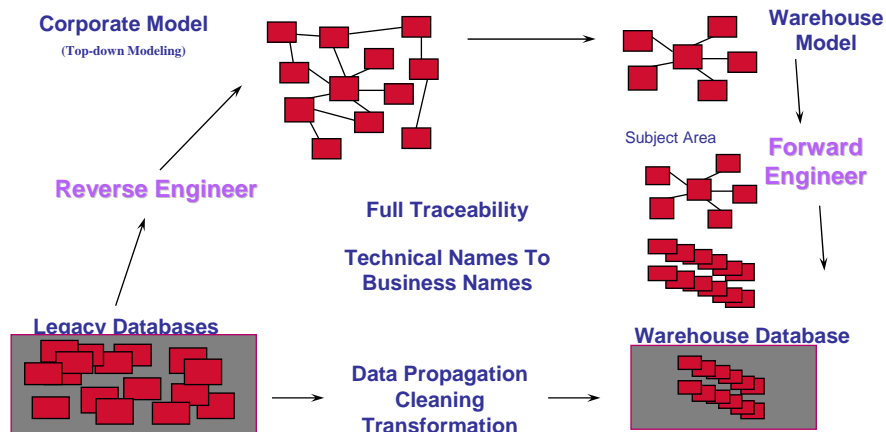
Credit Card Account Number =  
Numeric 16  
E.G 0987 6543 2112 3456

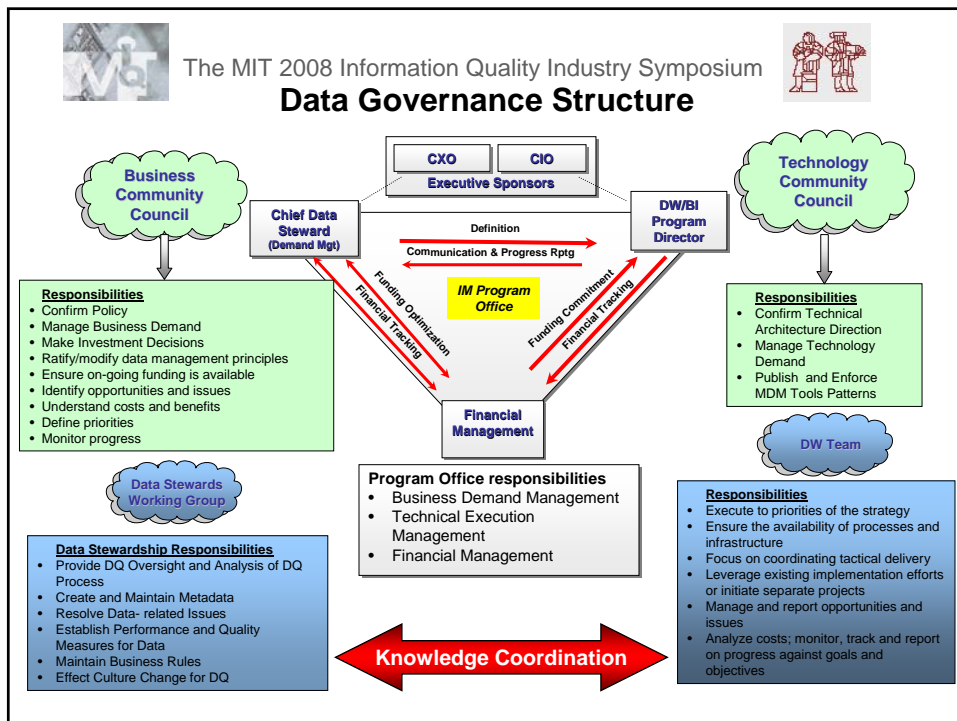
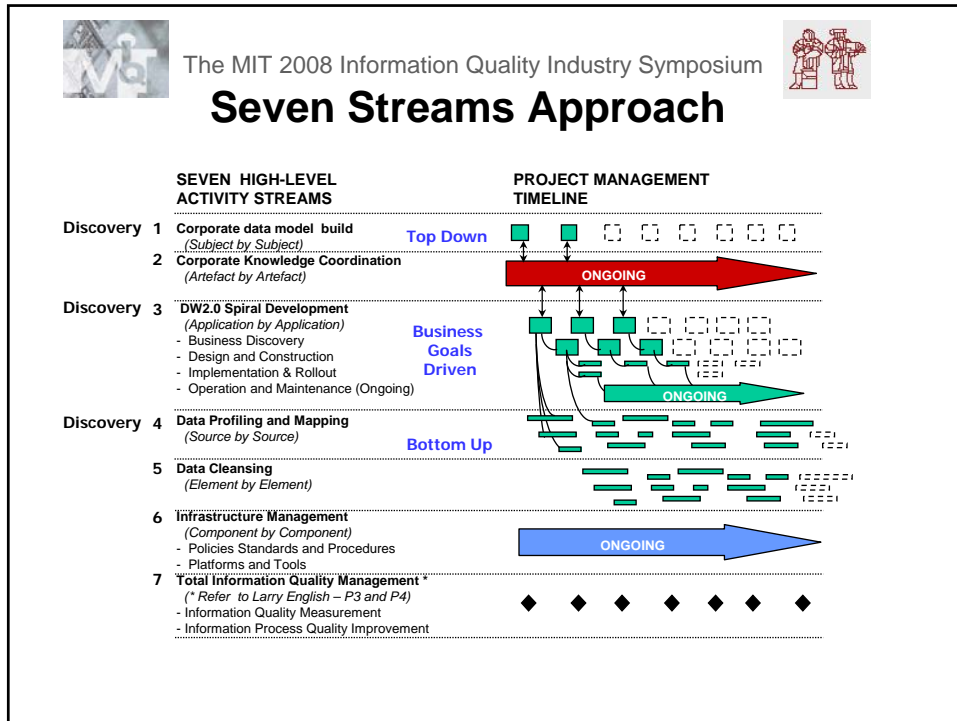


The MIT 2008 Information Quality Industry Symposium



## Legacy To Warehouse





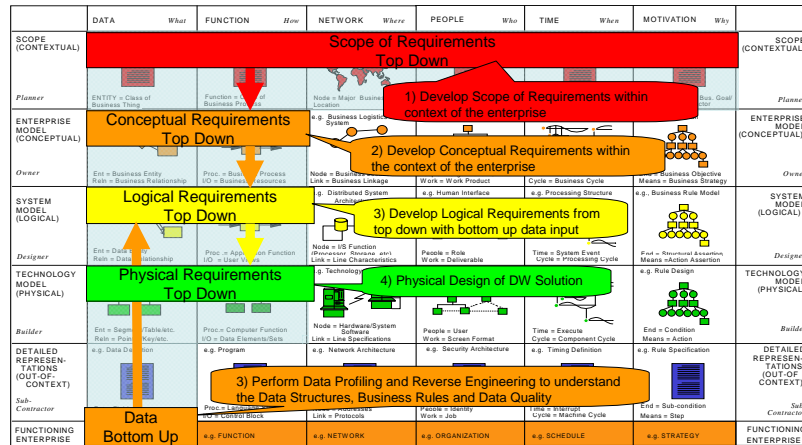


The MIT 2008 Information Quality Industry Symposium



# Knowledge Co-ordination Stream

ENTERPRISE ARCHITECTURE - A FRAMEWORK™



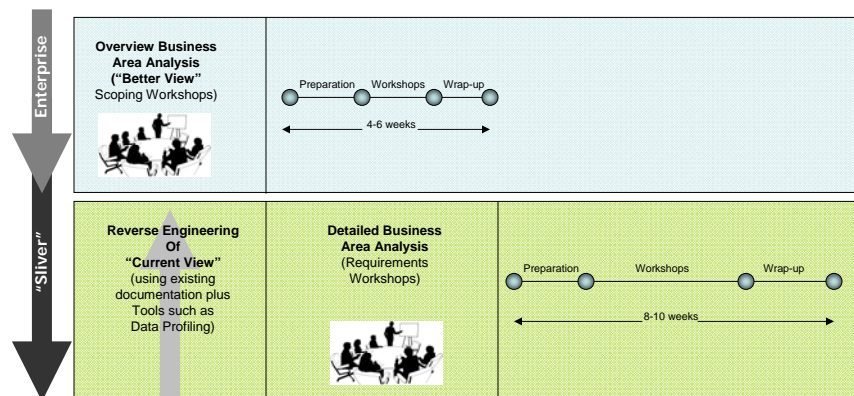
John A. Zachman, Zachman International (810) 231-0531

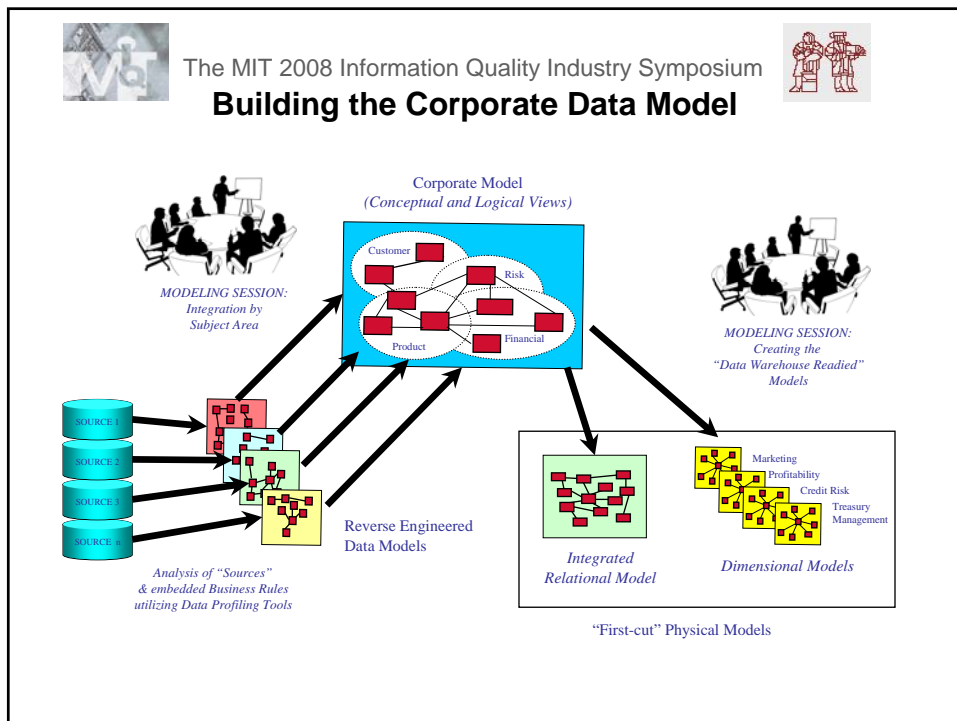
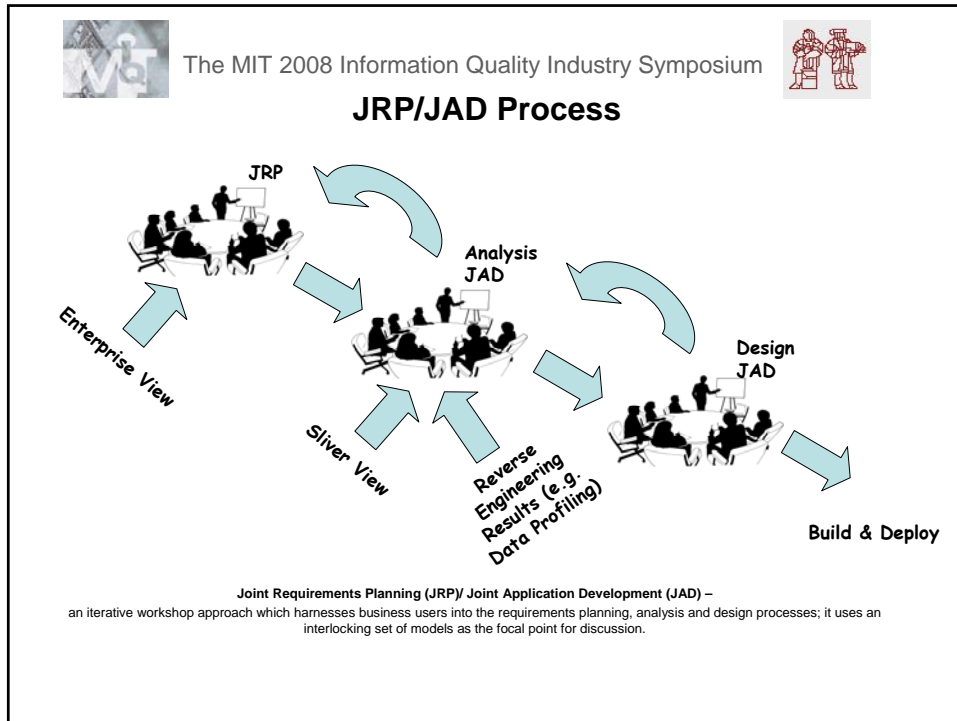


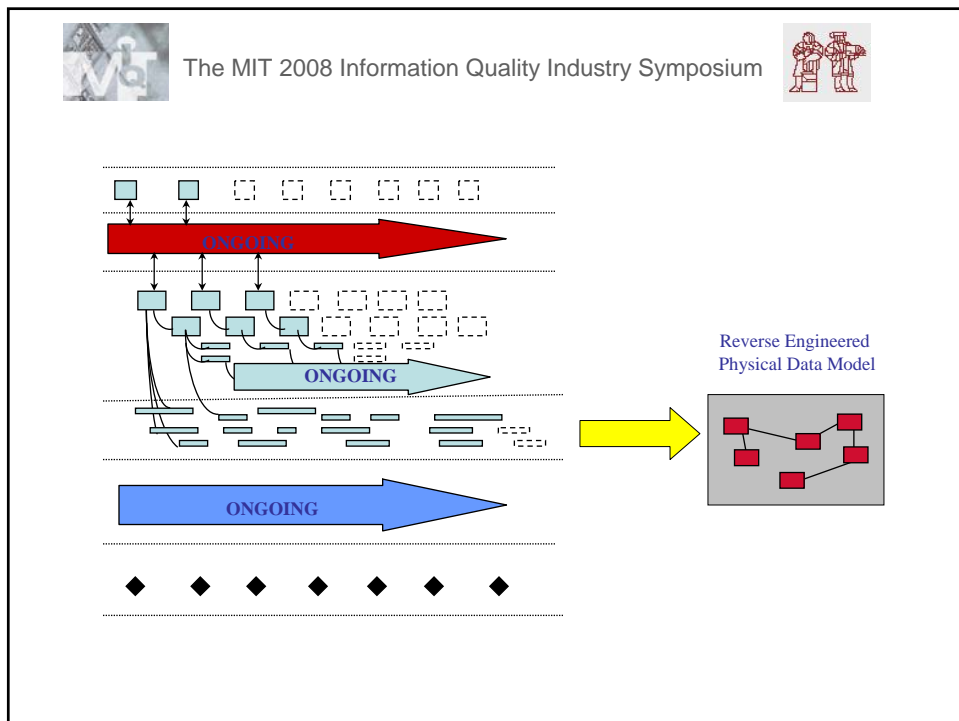
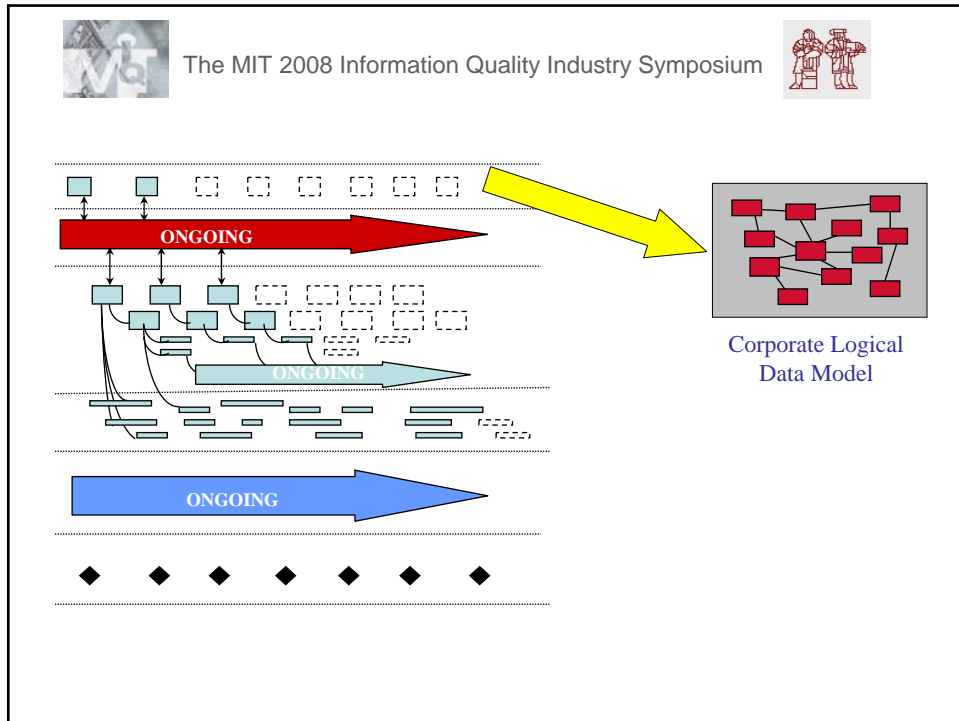
The MIT 2008 Information Quality Industry Symposium

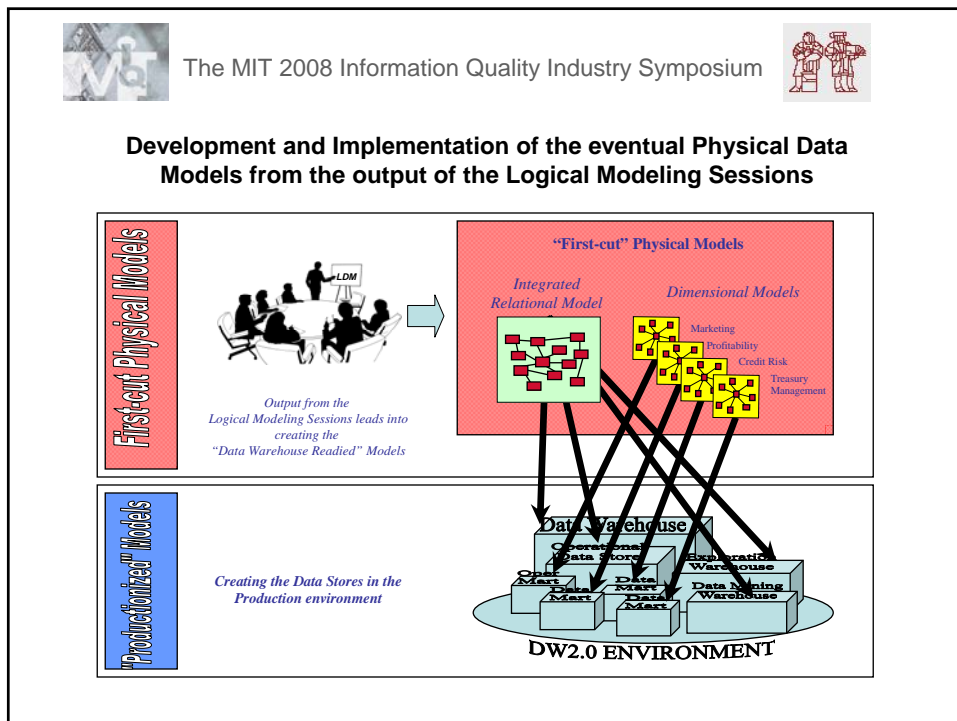
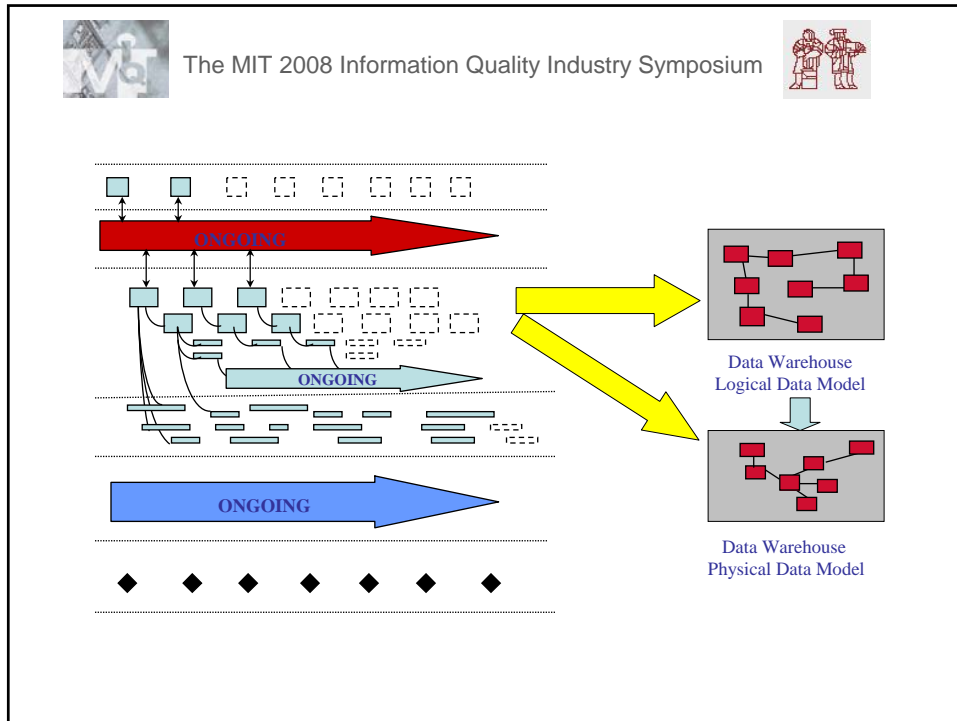


## “Ballpark” Timeline for each Architectural Initiative











The MIT 2008 Information Quality Industry Symposium

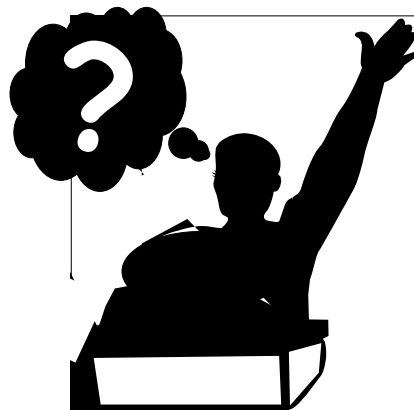


## Summary

- **Reusability is a critical success factor for second generation data warehousing and there is a much-needed focus on the quality of the Data Models which underpin the program.**
  - The models must accurately reflect the business and they must be reusable in all future releases of the program.
  - Sustained success of a DW/BI program requires a robust data architecture.
- **The foundational model is the Corporate Data Model.**
  - Traditionally, this model was derived using a top down approach and by utilizing Joint Requirements Planning and Joint Application Design techniques.
  - These techniques can deliver a good model relatively quickly.
  - The problem with models derived in this way is that they are based purely on business rules as perceived by management and senior analysts.
  - In reality, the systems that use the data may have a different set of rules. This is due to the fact that the systems are often 20 years old (and sometimes older). Undocumented changes have been made to the data and in the majority of cases the people that made the changes are no longer with the organization.
- **The only way to uncover what the data actually looks like is to reverse engineer the data into an abstracted logical data model.**
  - First generation data warehouse initiatives attempted this in the past but the tools available to help were limited.
  - Today a new set of tools has evolved – data profiling tools. These tools are an ideal aid to reverse engineer data and build a data model from the bottom up.
  - When a model is built in this way it is based on actual data content and the chance for errors and omissions in the data modeling process is reduced.
  - This “bottom-up” model is used as an input into the creation of the model that results from the “top-down” approach; in effect the former is used to challenge the latter model being drawn up by the business.



The MIT 2008 Information Quality Industry Symposium





# ***EWSolutions***

## **Implementing a Successful Enterprise Data Quality Initiative**

By **David Marco**  
*President*  
**EWSolutions**

[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 1

*Strategic Partner & Systems Integrator*  
*Intelligent Business Intelligence<sup>sm</sup>*



## ***EWSolutions' Background***

**EWSolutions** is a Chicago-headquartered strategic partner and full life-cycle systems integrator providing both **award winning** strategic consulting and **full-service implementation services**. This combination affords our clients a full range of services for any size enterprise information management, managed meta data environment, and/or data warehouse/business intelligence initiative. Our notable client projects have been featured in the Chicago Tribune, Federal Computer Weekly, Crain's Chicago Business, and won the 2004 Intelligent Enterprise's RealWare award, 2007 Excellence in Information Integrity Award nomination and DM Review's 2005 World Class Solutions Award.



**Information Integrity Coalition**  
2007 Excellence in Information Integrity Award Nomination



**Chicago Tribune**





**DM Review**  
2005  
World Class Solutions Award Data Management



**Federal Computer Week**  
GADGET



**intelligent Enterprise REAL Ware TRANSFORM**  
**2004 WINNER**  
Best Business Intelligence Application Information Integration  
Client: Department of Defense

For more information on our Strategic Consulting Services, Implementation Services, or World-Class Training, call toll free at 866.EWS.1100, 866.397.1100, main number 630.920.0005 or email us at [Info@EWSolutions.com](mailto:Info@EWSolutions.com)

[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 2

*Strategic Partner & Systems Integrator*  
*Intelligent Business Intelligence<sup>sm</sup>*





## EWSolutions' Partial Client List

<p>Arizona Supreme Court Bank of Montreal BankUnited Basic American Foods Becton, Dickinson and Company Blue Cross Blue Shield companies Branch Banking &amp; Trust (BB&amp;T) British Petroleum (BP) California DMV College Board Corning Cable Systems Countrywide Financial Defense Logistics Agency (DLA) Delta Dental Department of Defense (DoD) Driehaus Capital Management Eli Lilly and Company Federal Aviation Administration Federal Bureau of Investigation (FBI) Fidelity Information Services</p>	<p>Ford Motor Company GlaxoSmithKline Harris Bank The Hartford Harvard Pilgrim HealthCare Health Care Services Corporation Hewitt Associates HP (Hewlett-Packard) Information Resources Inc. International Paper Janus Mutual Funds Johnson Controls Key Bank LiquidNet Loyola Medical Center Manulife Financial Mayo Clinic Microsoft National City Bank Nationwide</p>	<p>Neighborhood Health Plan NORC Physicians Mutual Insurance Pillsbury Quintiles Sallie Mae Schneider National Secretary of Defense/Logistics South Orange County Community College SunTrust Bank Target Corporation The Regence Group Thomson Multimedia (RCA) United Health Group United States Air Force United States Navy United States Transportation Command USAA Wells Fargo Wisconsin Department of Transportation Zurich Cantonal Bank</p>
--	--	--




**Schedule**  
Contract GS-35F-0453M



For more information on our Strategic Consulting Services, Implementation Services, or World-Class Training, call toll free at 866.EWS.1100, 866.397.1100, main number 630.920.0005 or email us at [Info@EWSolutions.com](mailto:Info@EWSolutions.com)

[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 3


**Strategic Partner & Systems Integrator**  
Intelligent Business Intelligence<sup>sm</sup>



## Professional Profile/Contact Information

Mr. Marco is an internationally recognized expert in the field of enterprise information management, data warehousing, Capability Maturity Model (CMM), business intelligence, and is the **world's foremost authority on meta data management**. Mr. Marco has authored several books including the widely acclaimed **"Universal Meta Data Models"** (Wiley, 2004) and the classic **"Building and Managing the Meta Data Repository: A Full Life-Cycle Guide"** (Wiley, 2000). These groundbreaking books have been broadly endorsed by many of the largest software companies in the industry and by several major magazines.


- ☐ Selected to the prestigious **2004 Crain's Chicago Business "Top 40 Under 40"**
- ☐ Crain's Chicago Business anointed him the "Melvil Dewey of Metadata"
- ☐ **2008 DAMA Data Management Hall of Fame** (Professional Achievement Award)
- ☐ Chairman of the Enterprise Information Management Institute (EIMInstitute.ORG)
- ☐ **2007 DePaul University** named him one of their **"Top 14 Alumni Under 40"**
- ☐ Presented hundreds of keynotes/seminars across four continents
- ☐ Published hundreds of articles on information technology
- ☐ Author of several best selling information technology books
- ☐ Taught at the **University of Chicago** and **DePaul University**
- ☐ Judged dozens of various industry awards in meta data management and data warehousing



**Email:** [DMarco@EWSolutions.com](mailto:DMarco@EWSolutions.com)

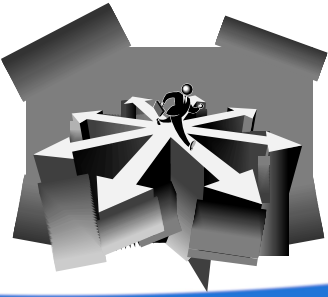
[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 4

**Strategic Partner & Systems Integrator**  
Intelligent Business Intelligence<sup>sm</sup>



## Agenda

- ❑ The State of Data Quality
- ❑ Data Quality Challenges
- ❑ Data Quality Foundations
- ❑ Data Quality Solutions
- ❑ Real-World Data Quality Case Studies



[www.ESolutions.com](http://www.ESolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 5

*Strategic Partner & Systems Integrator*  
Intelligent Business Intelligence<sup>sm</sup>



# The State of Data Quality

[www.ESolutions.com](http://www.ESolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 6

*Strategic Partner & Systems Integrator*  
Intelligent Business Intelligence<sup>sm</sup>



## The State of Data Quality

- ❑ Most organizations data quality can be best described as abysmal
- ❑ Large bank's Executive VP stated: ("The numbers (data in their data warehouse) are the best that we have,) and (as long as they are calculated the same way every time,) (they give us a directionally correct view of the business.")

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 7

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



## The State of Data Quality

- ❑ When data is presented to a business analyst or executive typically the person does not know:
  - if the data accurate?
  - what percentage of the data is not accurate or in doubt
  - the origin of the data
  - the meaning of the data
  - where the data came from
- ❑ Our information technology (IT) system's architectures are not conducive to data quality

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 8

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™

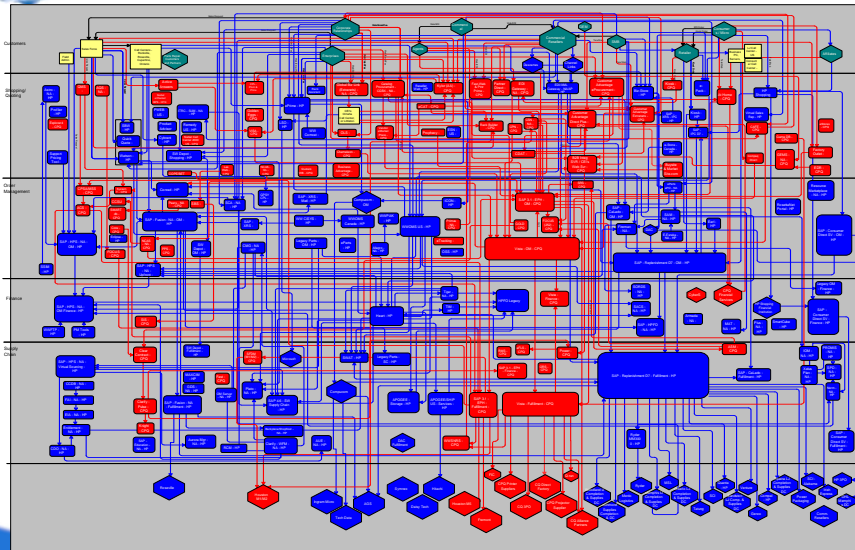
# What Does One Process Look Like for a Large Company?

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 9

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>

## Fortune 50 – One Process



[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 10

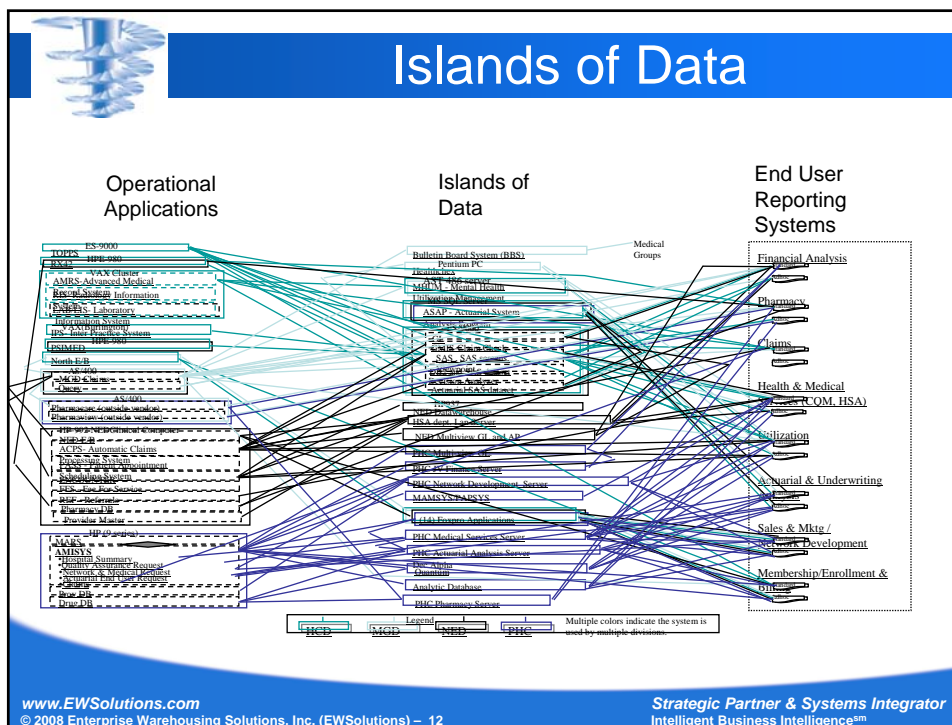
Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



# Maybe this Image Clears When you Look at One Application?

[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 11

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>





## How Did We Get Here

- ☐ Most companies did not plan their infrastructure...rather it just grew over time
- ☐ Previously companies focused on lines-of-business, as opposed to an enterprise view

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 13

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



## Why is Data Quality Growing as an Issue?

- ☐ Enterprise initiatives (e.g. data warehousing, customer relationship management, supply chain, etc.) are most significantly impacted by data quality issues
- ☐ Corporations are moving from a decentralized structure to a centralized structure
- ☐ Government regulations require data quality (SOX, 21 CFR Part 11, BASIL II, HIPPA, DoDAF, ISO 11179, etc.)

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 14


Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



# Data Quality Solutions

[www.ESolutions.com](http://www.ESolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 15

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



## Data Quality Solutions

- ☐ ***Data Quality Falls Under the Larger Enterprise Information Management Umbrella***
- ☐ ***Meta Data Management is the Technical Enabler of Data Quality***
- ☐ ***Data Stewardship defines the Business Processes of Data Quality***

[www.ESolutions.com](http://www.ESolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 16


Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



# Data Quality & Enterprise Information Management

[www.ESolutions.com](http://www.ESolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 17

*Strategic Partner & Systems Integrator*  
Intelligent Business Intelligence<sup>sm</sup>



## Data Quality & EIM

- ☐ Many companies rush to implement data quality “silo” solutions without realizing that they need to understand the “larger picture”
- ☐ Data quality falls under the larger Enterprise Information Management (EIM) umbrella

[www.ESolutions.com](http://www.ESolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 18

*Strategic Partner & Systems Integrator*  
Intelligent Business Intelligence<sup>sm</sup>





## What is EIM?

- ❑ **Enterprise Information Management (EIM):** The systematic processes and governance procedures for applications, processes, data, and technology at a holistic enterprise perspective
- ❑ The purpose of enterprise information management is to bring enterprise order, purpose, structure, efficiency, and performance to applications, processes, data, meta data and technology
- ❑ EIM is not a single technology or component, but a coordinated framework of disciplines for managing data, meta data and information assets throughout the organization
- ❑ ***Data Does Not Manage Itself!!***

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 19

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



## Data Quality & EIM

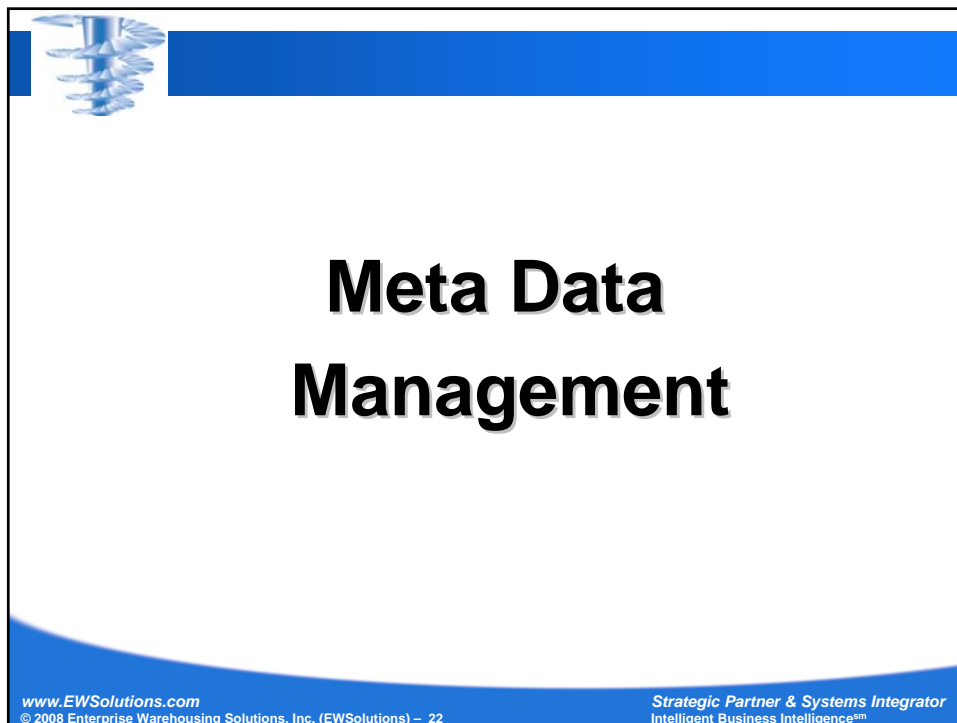
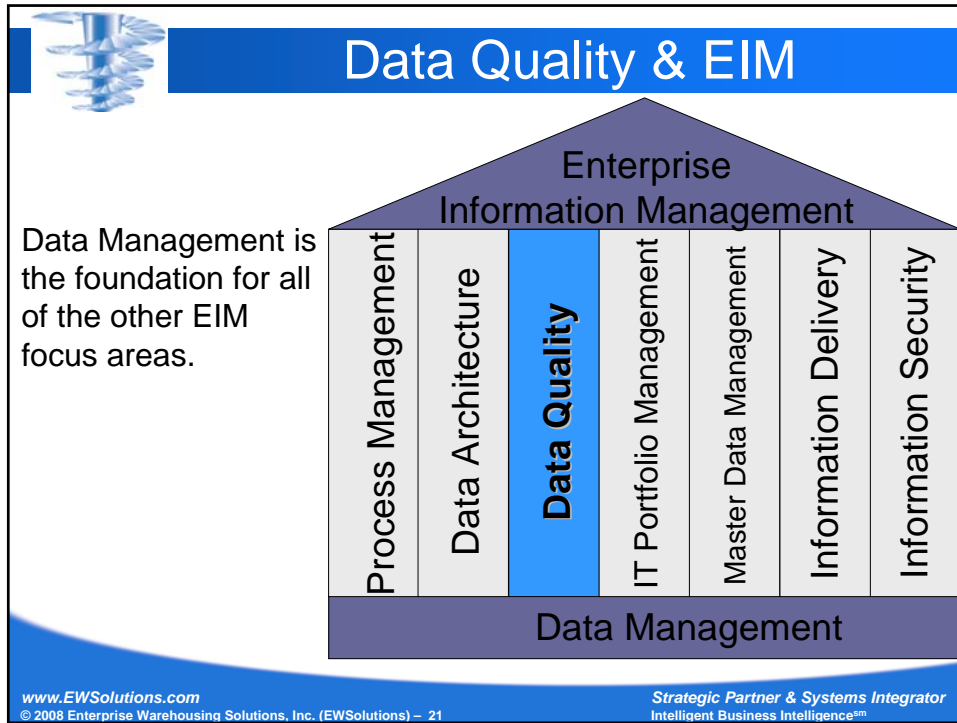
- ❑ There are three foundational elements that span each area of EIM
  - Meta Data Management
  - Data Governance
  - Data Management
- ❑ No matter what focus area of EIM you are targeting you will need to address each of these elements



[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 20

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>





## Meta Data vs. Data

- ❑ **Meta Data:** Meta data contains the knowledge that a **1)** field is called “Customer\_Name”, is 40 characters in length, and exists in systems A, B, and C; **2)** that our company has 3 systems which contain customer master data. These systems are...
- ❑ **Data:** Data would be a specific instance of “Customer\_Name” equaling “John Doe”
- ❑ **Information:** Data that is meaningful to a business user. They understand it and they know what to do with it



[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 23

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



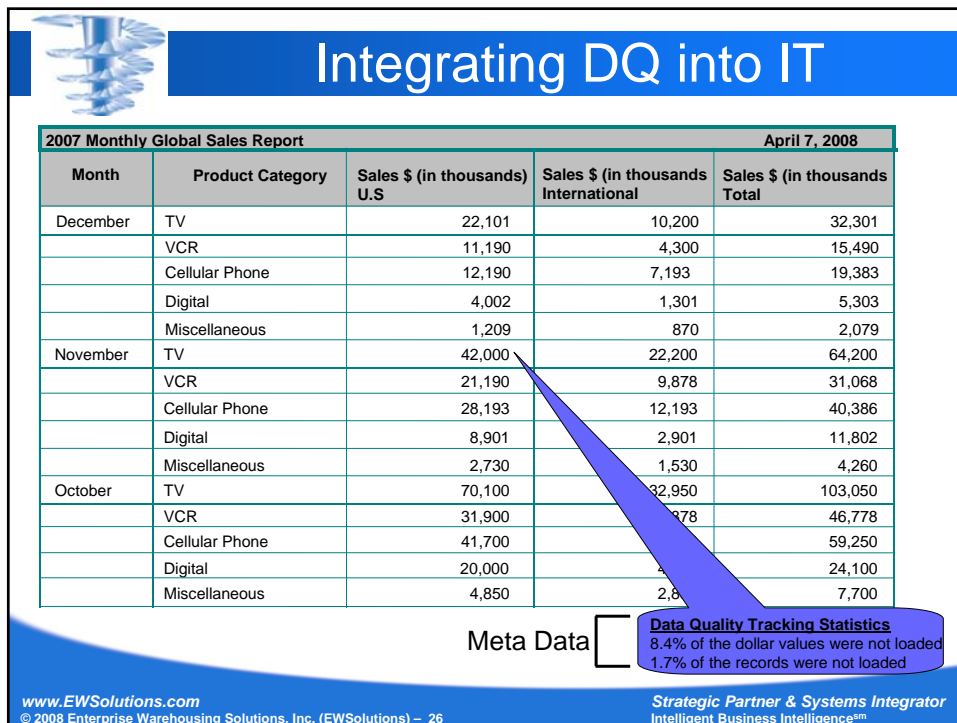
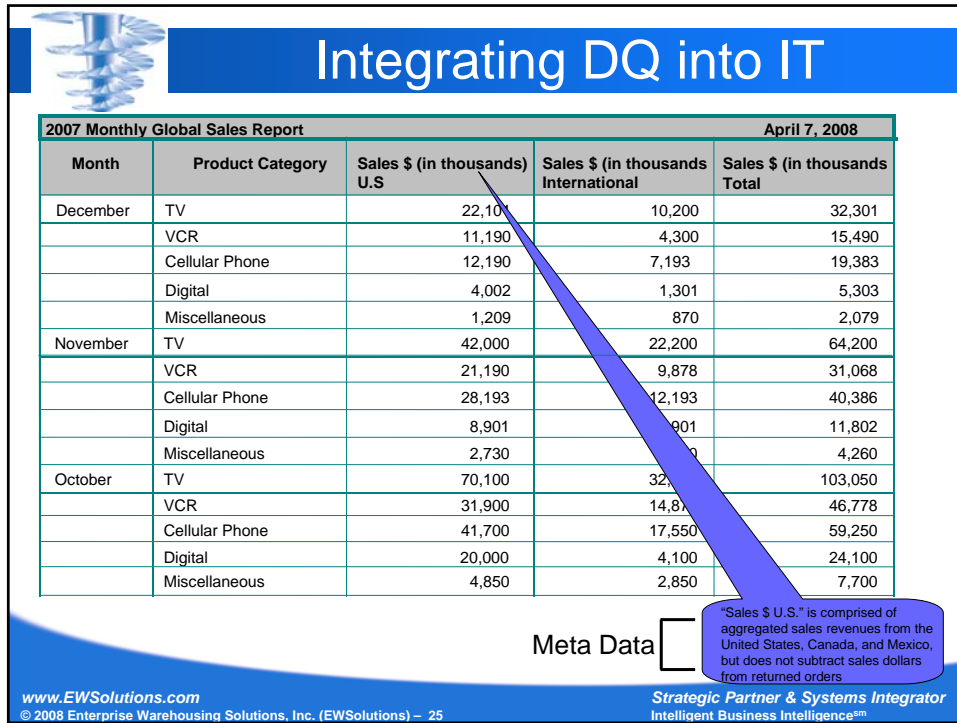
## EIM Fundamentals

**Information = Data + Meta Data**  
(content) (context)

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 24

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>





# Does Understanding Your Data Lead To Better Decisions?

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 27

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



## NASA Example – Mars Orbiter

- ❑ December 11, 1998 the Mars Climate Orbiter was launched
- ❑ Engineers calculated rocket firing using feet-per-second, the orbiter was programmed in meters-per-second (metric system) of thrust
- ❑ The difference was 4.4 feet per second
- ❑ “Each time there was a burn (rocket firing) the error built up,” said Art Stephenson, Director of the Marshall Spaceflight Center and Head of the NASA Investigation Team
- ❑ “We entered the Mars atmosphere at a much lower altitude (than planned),” said Ed Weiler, NASA's chief scientist. “It (the spacecraft) either burned up in the Martian atmosphere or sped out (into space). We're not sure which happened.”
- ❑ The cost of this mission was \$250 - \$300 million

Source: Associated Press, Paul Recer, <http://www.anomalous-images.com/news/news537.html>

[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 28

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™

# Integrating DQ into IT

Campaign Analysis Query

FirstRateMotors

January 20, 2008

Campaign Dates	Campaign Name	Campaign Description	Auto Sales	Auto Type
01/01/2003 - 03/31/2003	Buy Now, Pay Later	The "Buy Now, Pay Later" campaign was a nation-wide campaign. This campaign offered customers the option not to begin payments for a new car purchase until January 1, 2004. This offer is only applicable for those customers with better than standard credit ratings.	7,045	ECON – Smaller, cheaper, economy car line
			9,500	MID – Mid-size, sedan automobiles
			22,010	SPTS – Sports/performance cars
			14,700	SUV – Sport Utility Vehicles
			<b>Grand Total</b>	<b>53,255</b>
04/01/2003 - 06/30/2003	Buy One, Get A Free Scooter	The "Buy One, Get A Free Scooter" campaign was a nation-wide campaign. This campaign offered customers a free scooter for the purchase of a new car. Dealer can only discount cars, up to \$1,000	5,205	ECON – Smaller, cheaper, economy car line
			7,250	MID – Mid-size, sedan automobiles
			17,888	SPTS – Sports/performance cars
			10,900	SUV – Sport Utility Vehicles
			<b>Grand Total</b>	<b>41,243</b>
07/01/2003 - 09/30/2003	\$1 Trade-In	The "Buy One, Get A Free Scooter" campaign was a nation-wide campaign. This campaign offered customers a free scooter for the purchase of a new car. Dealer can only discount cars, up to \$1,000	6,102	ECON – Smaller, cheaper, economy car line
			8,330	MID – Mid-size, sedan automobiles
			19,750	SPTS – Sports/performance cars
			12,400	SUV – Sport Utility Vehicles
			<b>Grand Total</b>	<b>46,582</b>
10/01/2003 - 12/31/2003	0 down and 0% Interest	The "0 down and 0% Interest" campaign was a nation-wide campaign. This campaign offered customers standard car discounts, with 0% and 0% interest for only those customers with standard or better credit ratings.	6,700	ECON – Smaller, cheaper, economy car line
			8,925	MID – Mid-size, sedan automobiles
			20,820	SPTS – Sports/performance cars
			13,220	SUV – Sport Utility Vehicles
			<b>Grand Total</b>	<b>49,665</b>

## Data Quality Statistics

2.3% of the records were not loaded in the data warehouse batch runs.  
Skew percentage equals 2.3%

www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 29

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



# Data Quality and Meta Data Management

www.EWSolutions.com  
 © 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 30

Strategic Partner & Systems Integrator  
 Intelligent Business Intelligence<sup>sm</sup>



## Data Quality and Meta Data Management

- ❑ Good data quality professionals understand the that meta data management is their **key technical enabler**
- ❑ Where do business rules for data quality go?
  - a spreadsheet?
  - a document?
  - maybe we can just commit them to memory?
  - in a Managed Meta Data Environment (MME)
- ❑ What is a data quality rule? Its meta data!
- ❑ Let's look at some MME implementations that have strong focuses on data quality

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 31

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



## Data Stewardship

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 32

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



## Data Stewardship

- ❑ **Data Stewardship:** The process of having data stewards work with the data and meta data of an organization to ensure its quality, accuracy, formats, domain values, and that it is properly defined and understood across the enterprise
- ❑ **Data Steward:** A person(s) responsible for working with the data and meta data. There are different types of data stewards
- ❑ The data steward acts as the conduit between IT and the business. The data steward (often not just one person, but a collection of people) aligns the IT systems (both decision support and operational) with the business' requirements. The data steward has the challenge of guaranteeing that one of the corporation's most critical assets--its data--is used to its fullest capacity

[www.ESolutions.com](http://www.ESolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 33

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



## Data Stewardship

- ❑ Data is one of the most important assets in a corporation
- ❑ Data has value when it is delivered timely, properly formatted, concise, accurate and understood
- ❑ Business ownership of the data and active participation are critical
- ❑ The role of the data steward has grown considerably over the years

[www.ESolutions.com](http://www.ESolutions.com)


© 2008 Enterprise Warehousing Solutions, Inc. (ESolutions) – 34

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



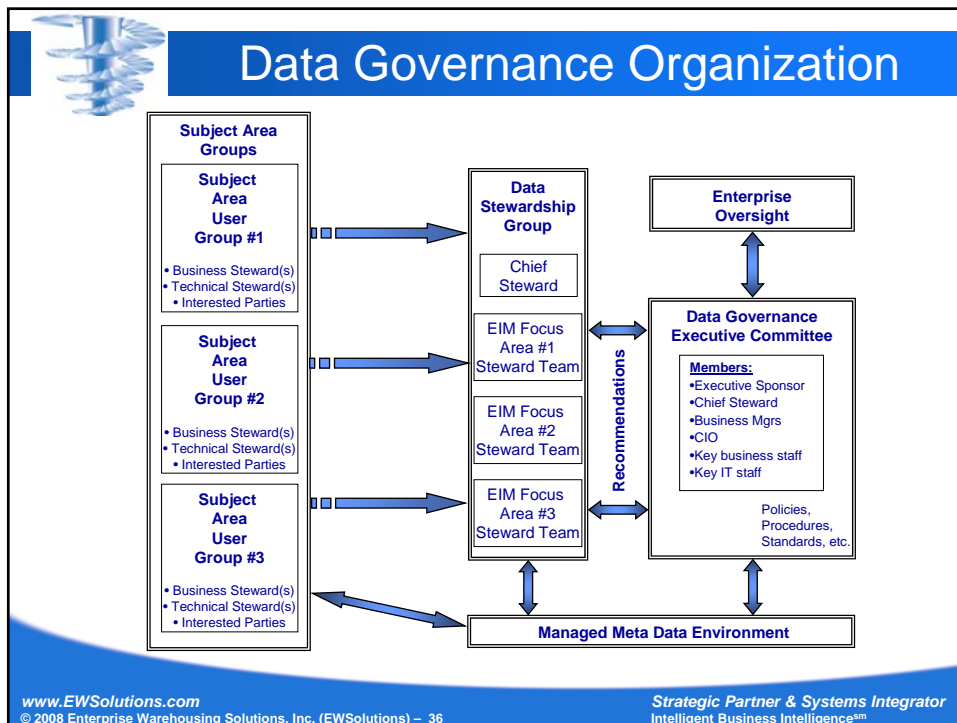
## Data Governance Organization

- ❑ Every organization forms their data governance organization a little differently
- ❑ Some have a more or less complex organization
- ❑ What is critical is that the organization:
  - is actively using the MME
  - has clear lines of communication
  - has a defined and well understood decision making process
  - well defined feedback loop



www.EWSolutions.com  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 35

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>







# Real-World MME Implementations

[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 37

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>




# Allstate Corporation



*The following slides are derived from the book "Universal Meta Data Models", David Marco & Michael Jennings, Wiley 2004*


[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 38

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>




## Allstate Corporation

- ❑ The nation's largest publicly held personal lines insurer
- ❑ Provides insurance products to more than 16 million households and has approximately 12,300 exclusive agents and financial specialists in the United States and Canada
- ❑ Multi-channel organization: customers can access products and services through Allstate agents, or in select states at allstate.com and 1-800-Allstate®
- ❑ Allstate Financial Group includes the businesses that provide life and supplemental insurance, retirement, banking, and investment products through distribution channels that include Allstate agents, independent agents, and banks, and securities firms
- ❑ ***The MME was part of a larger Data Asset Management effort***



[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 39

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™




## Allstate Corporation

### Challenges/Opportunity

- ❑ Early 1990s Allstate, like most large corporations, found itself challenged with managing disparate systems to satisfy its IT needs
- ❑ Needed applications to talk across platforms
- ❑ Systems that have different coding schemes for common codes and mismatches in field types and sizes cannot interchange data easily
- ❑ Dramatically reduce the number of point-to-point interfaces
- ❑ Needed a precise understanding and knowledge of the data that the analysts of the data warehouse would utilize

[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 40

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™




## Allstate Corporation

### Challenges/Opportunity

- ❑ Need to Manage Code Data
  - Very “code-driven” environment because every state has its own regulatory statutes to which Allstate must adhere
  - Highly time-consuming and difficult to integrate systems because projects would take longer to deliver and be too costly
  - Sound meta data management techniques was expected to reduce or eliminate IT rework, speed up projects, and lower their overall costs
- ❑ Enable Data Warehousing Applications
  - Data warehousing made data quality more important than ever
  - Previously subject matter experts in an application area had to know what the data represented, what it actually meant, and how to use it
  - Now this data was going to be presented to actual end users, and Allstate cannot afford to have user “interpretations” of what the data *may* mean be used to make “live” business decisions.
  - A strong, centralized data management environment would be the basis for consistent data driving high-quality decisions by end users

[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 41

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™



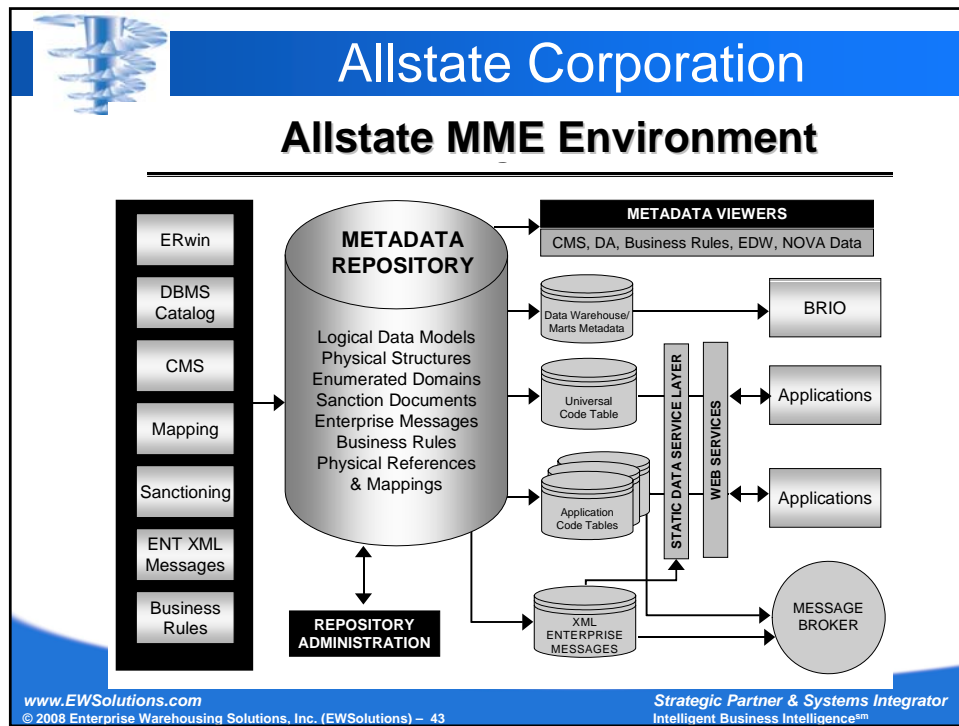
## Allstate Corporation

### MME Overview

- ❑ Custom built MME
- ❑ Initially focused on the management of the codes that permeated their systems
- ❑ A portion of the MME (Codes Management System (CMS)) allows for the identification of enumerated domains (those for which a set list of values can be listed) and define the various coding schemes that were found in the different applications and the associated business values
- ❑ CMS allowed Allstate’s Codes Analysts group to do their job more effectively
- ❑ Codes analysts, along with a group of data administrators, then became the nucleus of the Enterprise Data Management group
- ❑ Kept an enterprise perspective by documenting each unique domain they encountered and storing it in the MME
- ❑ As they worked with subsequent projects, they were able to see where the same data had been encoded differently between applications

[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 42

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™




## Allstate Corporation

### Data Stewardship


- ❑ January 1997 formed a Data Stewardship Council (DSC)
- ❑ DSC is a cross-business-unit team focused on the business aspects of managing data as a valued enterprise asset
- ❑ DSC is a part-time, virtual team of Allstate employees who have strong business knowledge, vision, and the ability to look horizontally across the enterprise
- ❑ Data stewards are focused on addressing the business issues behind key data resource management objectives: managing data redundancy, implementing data shareability and standardization, and managing and improving data integrity
- ❑ Stewards follow several basic principles for managing data resources of any type; these include the following:
  - Requirements for the resource must be anticipated and fulfilled proactively
  - Allstate cannot afford an infinite amount of the data resource; therefore, the amount must be optimized
  - Data resource should be shared and leveraged in as many ways as possible, in order to maximize its value while diminishing its overall costs
  - Data resource must be carefully managed to ensure that its use in the business is prudent, efficient, effective, and secure

www.EWSolutions.com  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 44

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>




# Department of Defense




[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 45

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



## How Large is the DoD Supply Chain?

- ☐ \$480 Billion Dollars Budget
- ☐ \$80 Billion Supply Chain
- ☐ 3 Million People
- ☐ 8 Million Parts
- ☐ Global – 150 Countries
- ☐ Dynamic Supply Chain



[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 46

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



## What's the Challenge

- ☐ Can't Account for Several Billion Dollars a Year
- ☐ Have Data Quality Issues
- ☐ Fix or Refine a Process and Break Several Other Processes
- ☐ Never Planned the Enterprise
- ☐ The Enterprise "Just Grew"

[www.EWSolutions.com](http://www.EWSolutions.com)

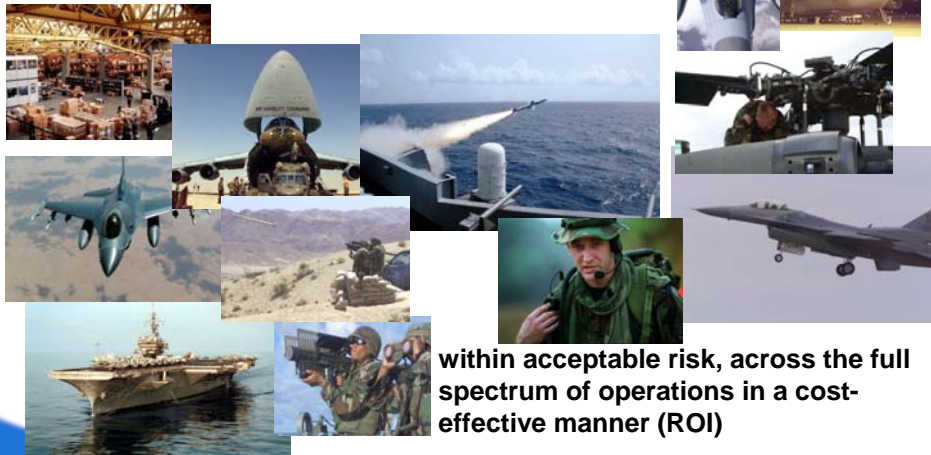
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 47

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



## DoD Enterprise Architecture is Based on:

**Support rapid, agile deployment, employment, sustainment and reset/reconstitution Total Force,**

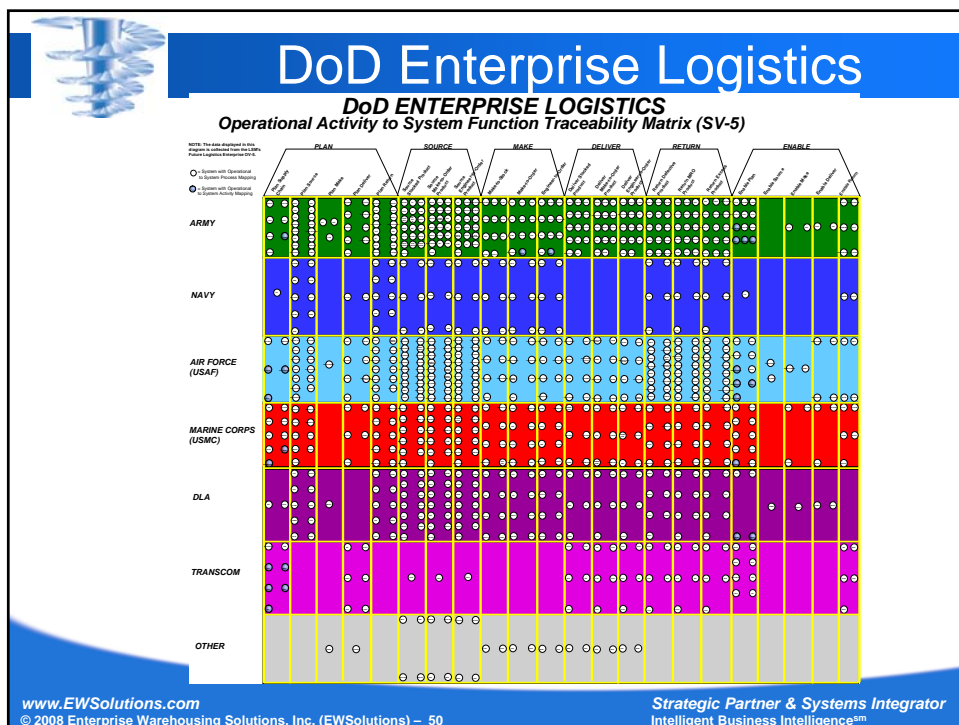
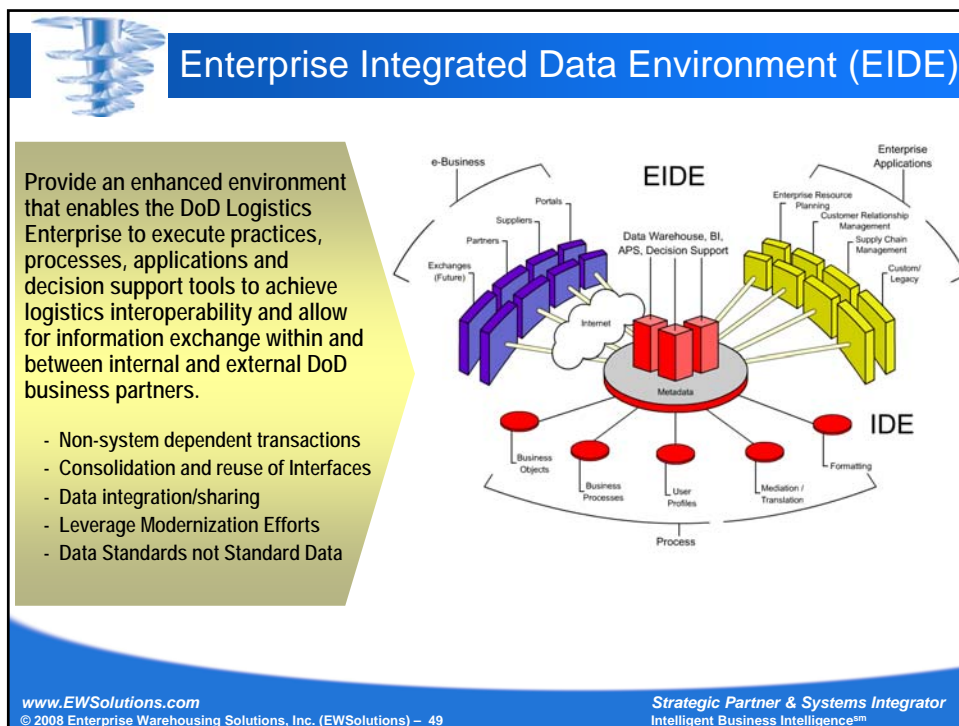


**within acceptable risk, across the full spectrum of operations in a cost-effective manner (ROI)**

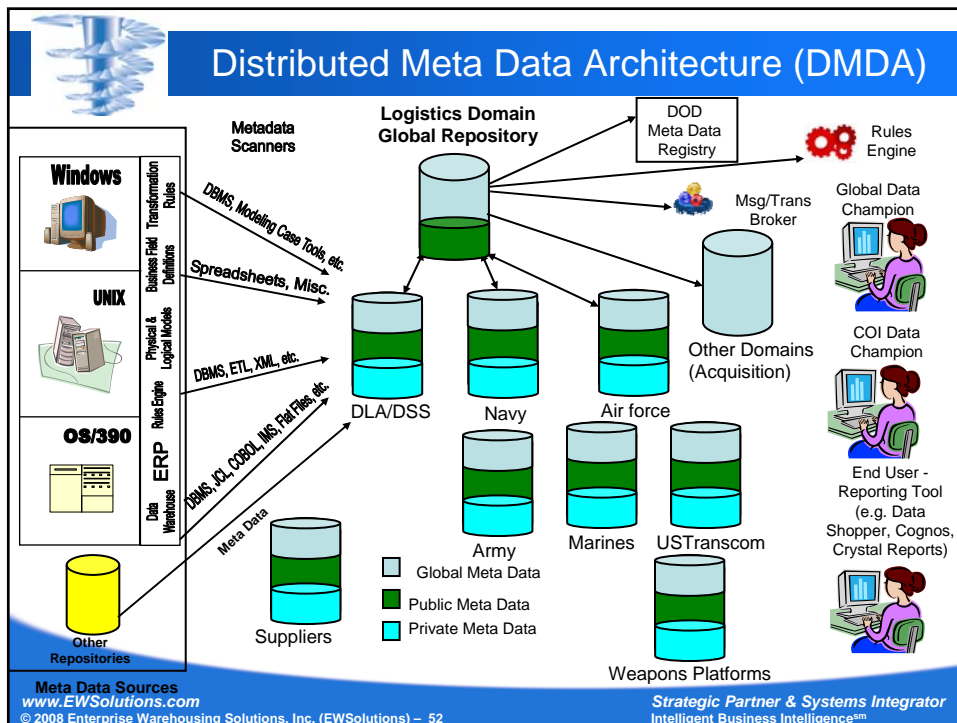
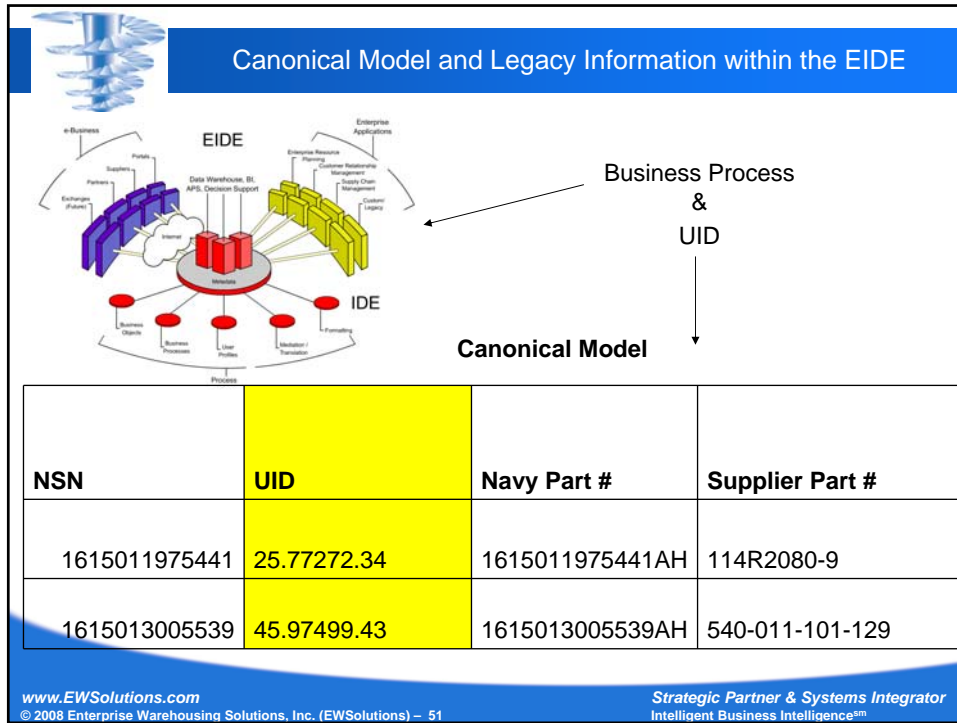
[www.EWSolutions.com](http://www.EWSolutions.com)

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 48

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>







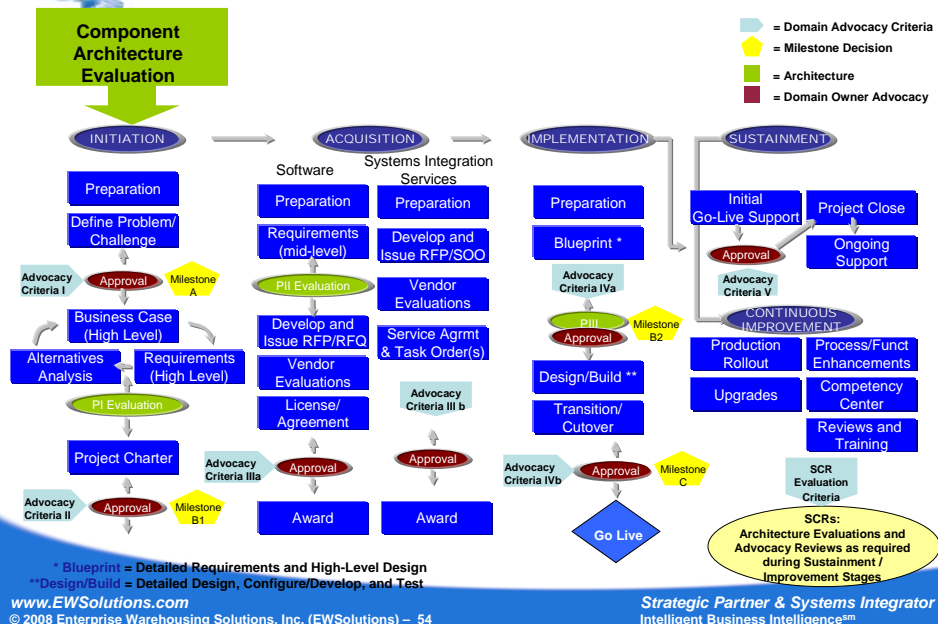
# Applying Compliance Architecture and Programs The Toolkit™


www.EWSolutions.com

© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 53

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence™


## Architecture and Program Compliance






## Summary

- ❑ Plan Your Enterprise Around the “To Be Business Processes” not your “As Is” (80-20 Rule)
- ❑ Data Strategy is the Key to Data Interoperability Across the Enterprise
- ❑ No Meta Data Management, No Data Strategy
- ❑ Compliance is an “On-Going Process”
- ❑ Your Business Processes Should Not Be Unique




[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 55

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>




## Don't Limit Yourself




[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 56

Strategic Partner & Systems Integrator  
Intelligent Business Intelligence<sup>sm</sup>



## Questions



[www.EWSolutions.com](http://www.EWSolutions.com)  
© 2008 Enterprise Warehousing Solutions, Inc. (EWSolutions) – 57

*Strategic Partner & Systems Integrator*  
Intelligent Business Intelligence<sup>sm</sup>



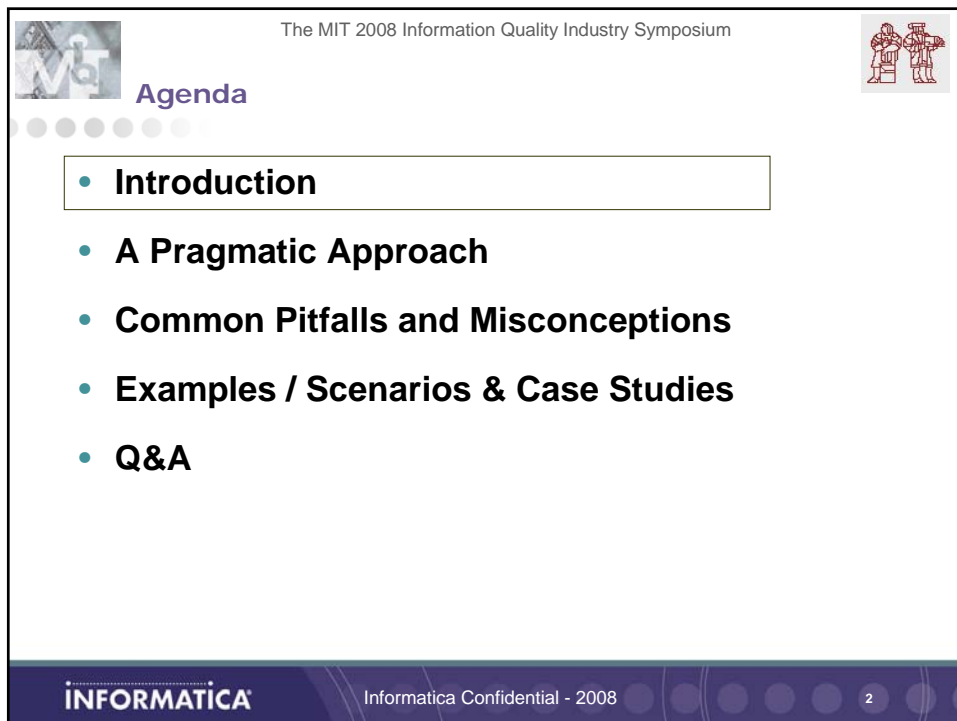
**Data Quality for  
Successful Data Governance**

**INFORMATICA**



*Ivan Chong  
EVP and General Manager  
Data Quality Business Unit  
Informatica Corporation*

Informatica Confidential - 2008

1

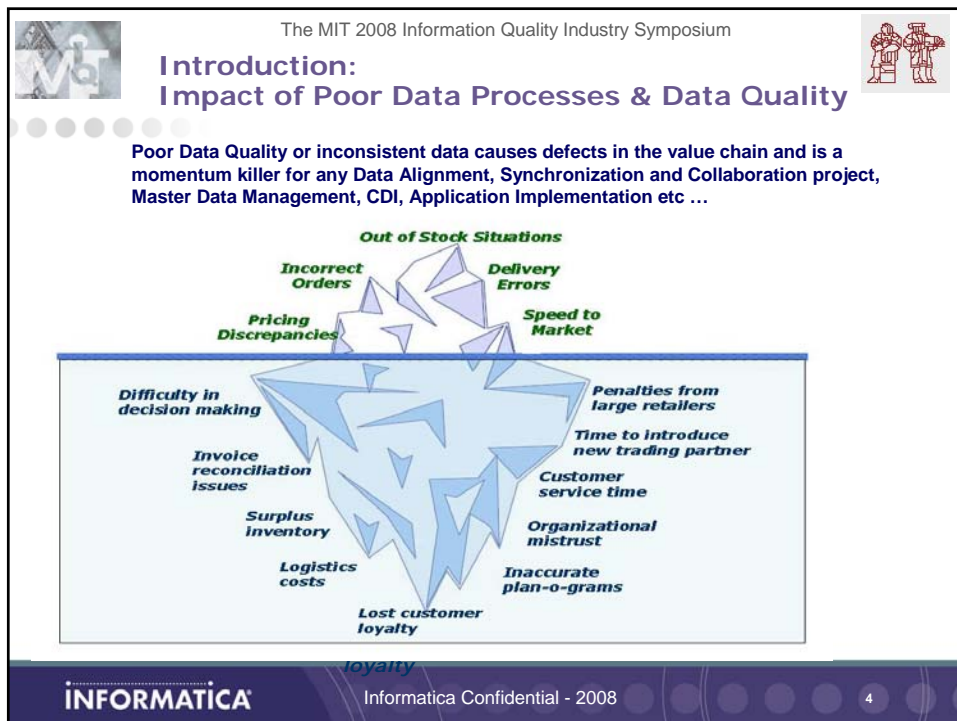
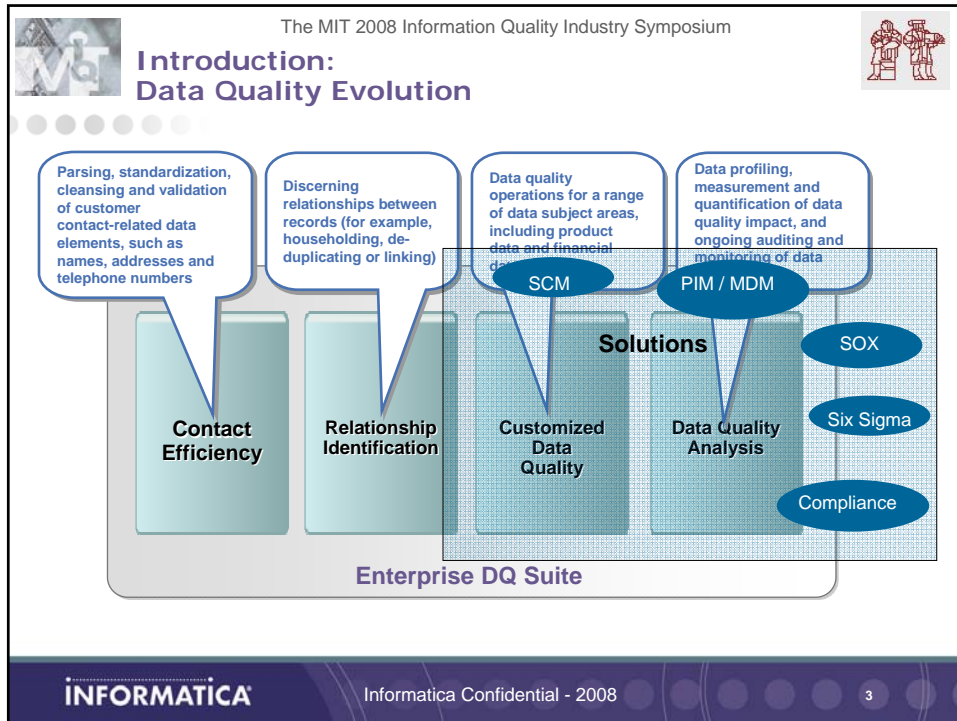



The MIT 2008 Information Quality Industry Symposium

 **Agenda** 

- **Introduction**
- **A Pragmatic Approach**
- **Common Pitfalls and Misconceptions**
- **Examples / Scenarios & Case Studies**
- **Q&A**

**INFORMATICA** Informatica Confidential - 2008 2





The MIT 2008 Information Quality Industry Symposium


## Introduction: One Perspective to Governance, top down



**Governing corporate data may be slightly easier than governing nations, but companies can benefit from adopting similar strategies, according to a Gartner analyst.**


- Executive level.** sponsorship, strategic direction, funding, advocacy and oversight
- Judicial level.** planning activities and to enforce governance activities or corporate policies. Mediating disagreements
- Legislative level.** chaired by a senior business leader designated by the executive team and may include business and technology leaders from Finance, IT, Data Management and Operations
- Administrative level.** Implement data governance on a day-to-day basis; responsible for developing data models and corporate data vocabularies, implementing master data management best practices, organizing content

**Top Down – often initiated by Exec initiative – Compliance / Audit / BPR / Six Sigma**




Informatica Confidential - 2008

5



The MIT 2008 Information Quality Industry Symposium

## Introduction: Data Quality Perspective to Governance



Lack of a comprehensive **Data Quality Management Strategy** and poor **Data Quality** is causing **Master Data Management (MDM)** initiatives to **fail** or be **significantly delayed**.

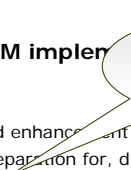
**key elements of effective Data Quality Management:**

- capable data processes
- ownership and accountability
- standards and metrics
- measurement and control


**key stages in any migration or MDM implementation**

- discovery and assessment
- planning and preparation
- cleansing, alignment, enrichment and enhancement
- operational impact and support in preparation for, during and after implementation
- Transition management**
- post implementation measurement and control

**Operational  
Risk**





**Bottom Up – often initiated by project – DQ / Migration / App Implement / Six Sigma**



Informatica Confidential - 2008


6



The MIT 2008 Information Quality Industry Symposium



## Agenda

- Introduction
- **A Pragmatic Approach**
- Common Pitfalls and Misconceptions
- Examples / Scenarios & Case Studies
- Q&A



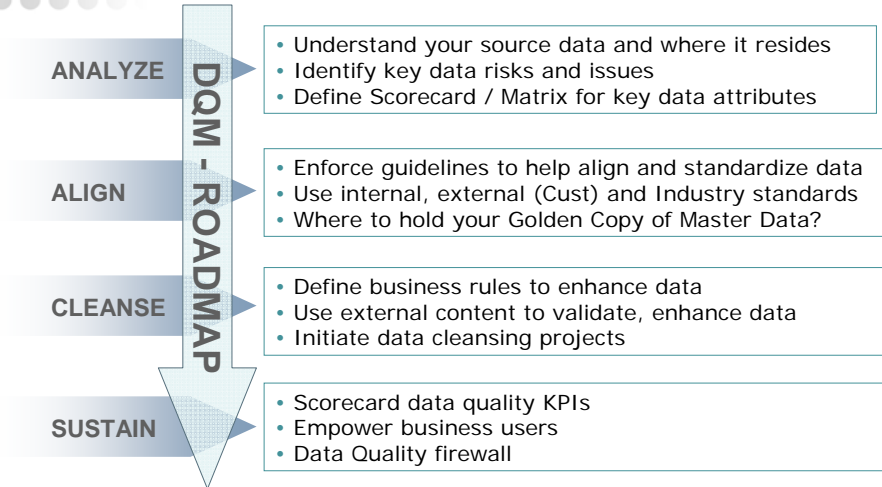
Informatica Confidential - 2008

7



The MIT 2008 Information Quality Industry Symposium

## Data Quality Management – A Good Starting Point



**ANALYZE**

- Understand your source data and where it resides
- Identify key data risks and issues
- Define Scorecard / Matrix for key data attributes

**ALIGN**


- Enforce guidelines to help align and standardize data
- Use internal, external (Cust) and Industry standards
- Where to hold your Golden Copy of Master Data?

**CLEANSE**

- Define business rules to enhance data
- Use external content to validate, enhance data
- Initiate data cleansing projects

**SUSTAIN**


- Scorecard data quality KPIs
- Empower business users
- Data Quality firewall




Informatica Confidential - 2008

8






The MIT 2008 Information Quality Industry Symposium




## Agenda

- Introduction
- A Pragmatic Approach
- **Common Pitfalls and Misconceptions**
- Examples / Scenarios & Case Studies
- Q&A




Informatica Confidential - 2008

9

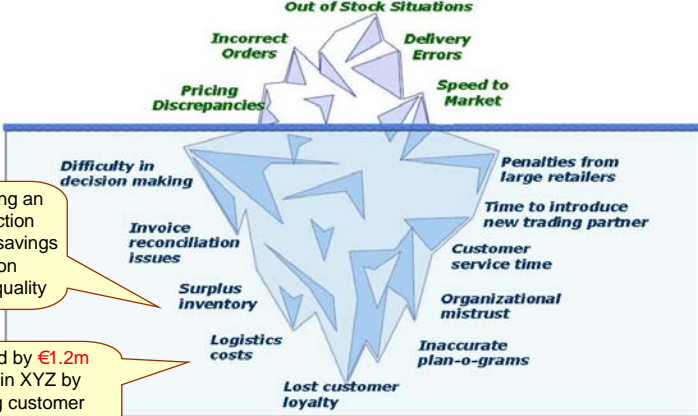


The MIT 2008 Information Quality Industry Symposium




## Background: Focus on value early

The impacts of poor quality data are not always highly visible. Using Supply Chain as an example here, many impacts are hidden 'below the waterline'. The visible impacts represent the 'tip of the iceberg'



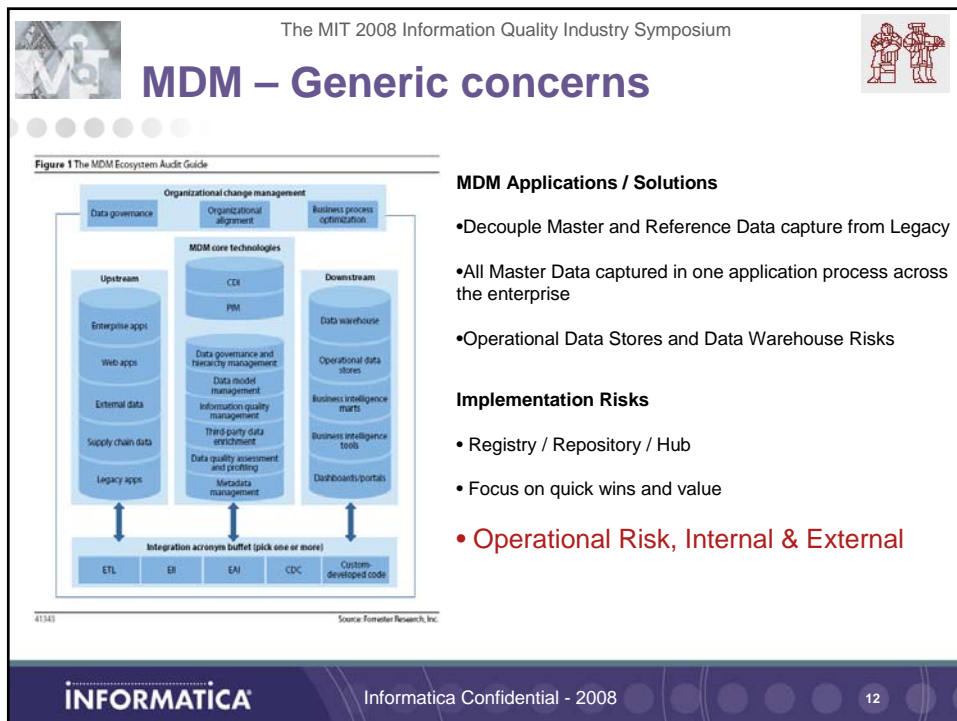
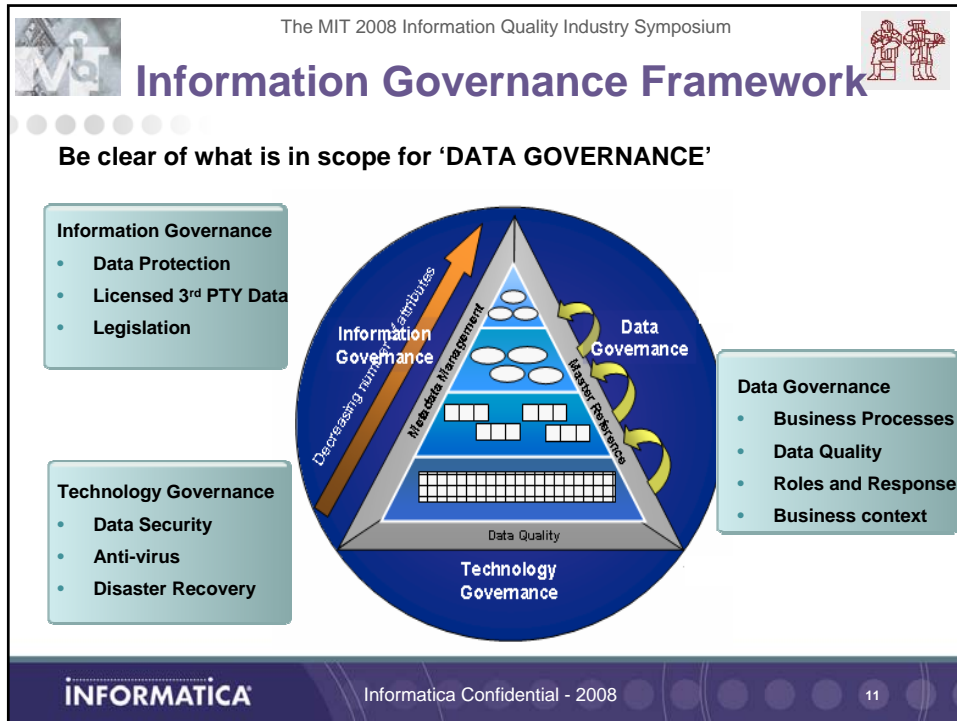
When considering an inventory reduction initiative, **13%** of savings were based on improving data quality

Decreased by **€1.2m** annually in XYZ by improving customer delivery time data




Informatica Confidential - 2008

10

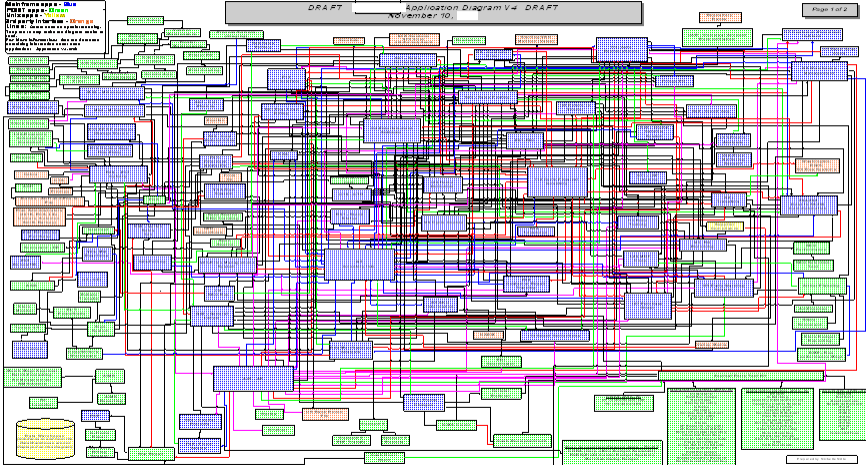


The MIT 2008 Information Quality Industry Symposium



## The Challenge


### Data Integration Complexity



Application Diagram V4 - DRAFT  
November 10, 2007  
Page 1 of 2

**INFORMATICA** Informatica Confidential - 2008 13


The MIT 2008 Information Quality Industry Symposium



## The Challenge

### Early Analysis is critical – Obsolescence / Duplicates

**Data Alignment means focus on Data Quality**





Results of worldwide Data Cleansing:

	Obsolete	Duplicate	Active	Cleansing rate
Material	724'804	29'529	544'383	58%
Customer	1'728'298	105'033	1'733'809	51%
Vendor	1'079'660	23'266	632'952	64%
<b>Total</b>	<b>3'532'762</b>	<b>157'828</b>	<b>2'911'144</b>	<b>56%</b>


Cleansing includes completeness and correctness checks on active data records

**INFORMATICA** Informatica Confidential - 2008 14




The MIT 2008 Information Quality Industry Symposium

## One Customers Experience - Large US Bank




### Lessons Learned

- Lessons Learned #1
  - Too much, too fast. Don't try to solve all the problems at once
  - Give the stakeholders enough to start with, but do not overload them
  - Only give the stakeholders the defects that apply to them & separate defects by type:
    - Data Entry defects → Front line stakeholders
    - Data Movement defects → Extract/Load Technology teams
    - Data Enrichment → Transformation Technology teams
- Lessons Learned #2
  - Create data quality business rules that align with Policies and Procedures
  - Don't create rules that "seem to make sense". Create rules that are "Actionable"



Informatica Confidential - 2008

15



The MIT 2008 Information Quality Industry Symposium

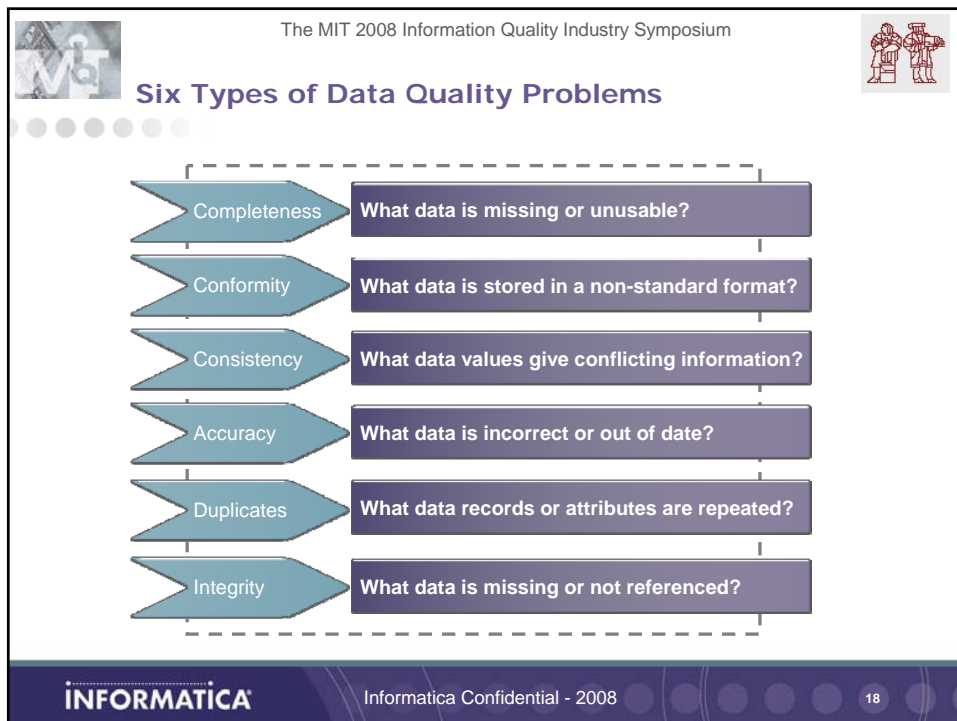
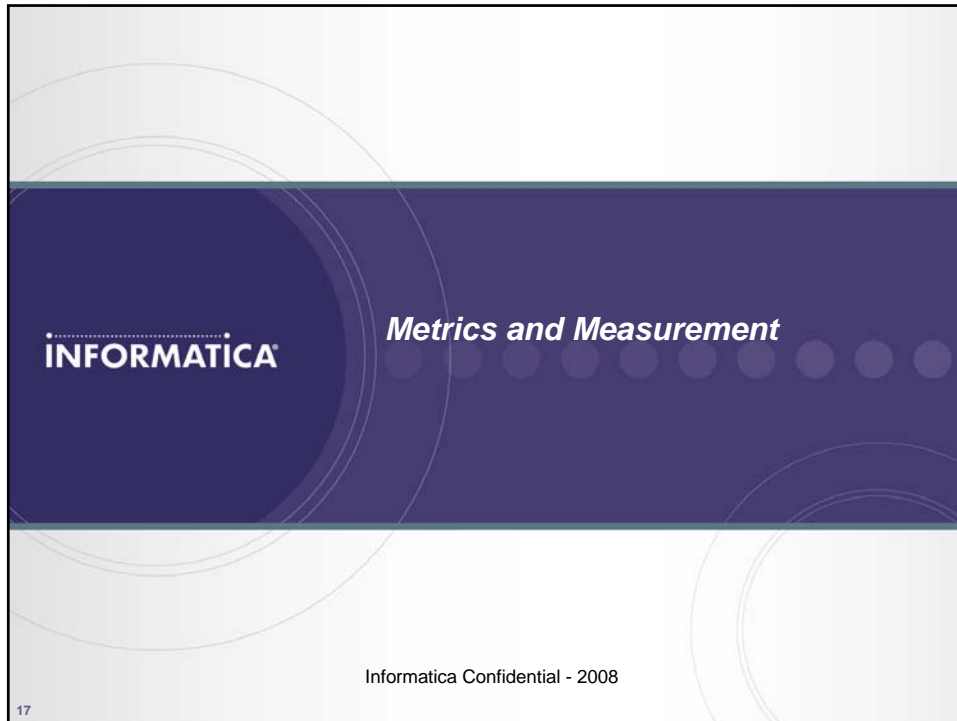
## Agenda

- Introduction
- A Pragmatic Approach
- Common Pitfalls and Misconceptions
- Examples / Scenarios & Case Studies
- Q&A



Informatica Confidential - 2008

16



The MIT 2008 Information Quality Industry Symposium

## Data Quality Certification Scorecard Attributes Within Object with context

**Scorecard**

**Bad Record Reports  
Issue Reports**

Completeness  
Conformity  
Consistency

**INFORMATICA** Informatica Confidential - 2008 19

The MIT 2008 Information Quality Industry Symposium


## Data Quality Certification - Scorecard per Object

**Customer Data Scorecard**


**Intranet Monitor**

72

**INFORMATICA** Informatica Confidential - 2008 20




The MIT 2008 Information Quality Industry Symposium



## Example – CPG - What is Data Quality?


<b>Completeness</b>	Required values electronically recorded
<b>Standards Based</b>	Data conforms to industry standards
<b>Consistency</b>	Data values aligned across systems
<b>Accuracy</b>	Data values are right, at the right time
<b>Time Stamped</b>	Validity timeframe of data is clear

Data Quality




Informatica Confidential - 2008


21



The MIT 2008 Information Quality Industry Symposium




## Customer Example: AML



### CIP/KYC- Reporting Definitions

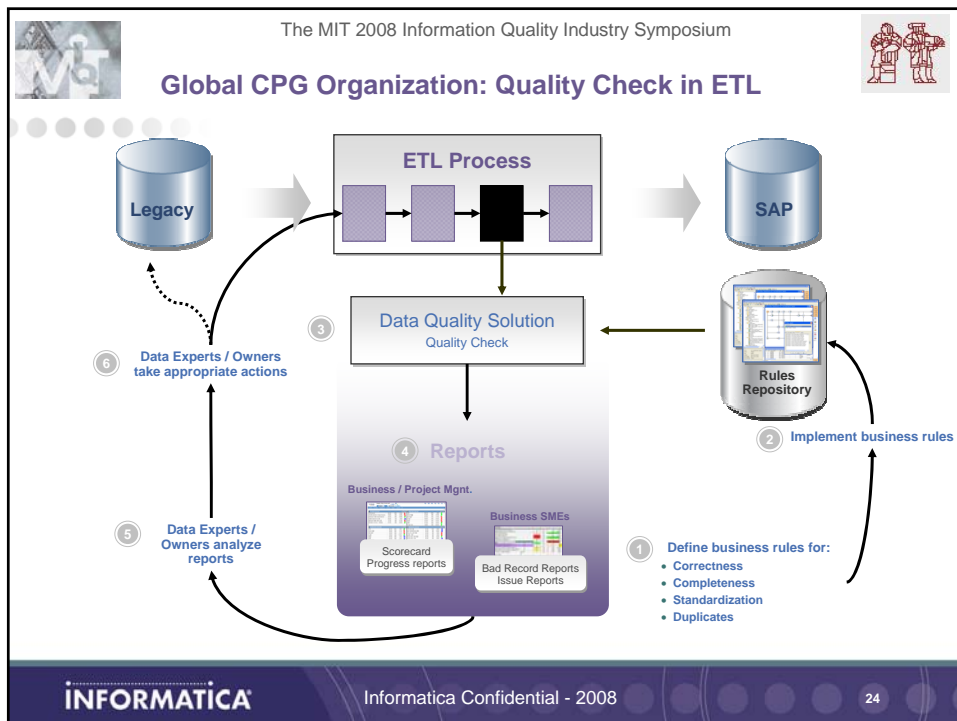
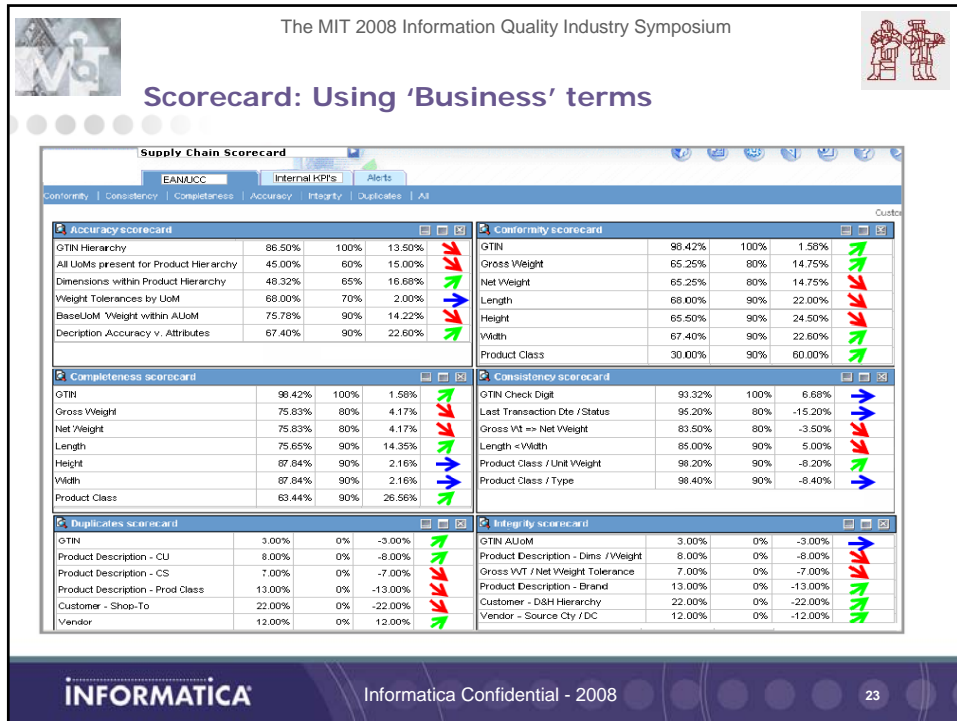
- Completeness** - A measure of the degree to which the requested information field contains data and is non-blank.
- Logical** - A measure of the degree to which the requested information field contains data that has been determined to be sound and reasonable within parameters defined by the responsible AML Process team.
- Verification** – A measure of the degree to which the requested information field contains data that agrees with an appropriate authoritative source as determined by the responsible AML Process Team.
- Accuracy** – A measure of the degree to which the requested information field contains data that agrees with the source or is consistent with fact or reality.



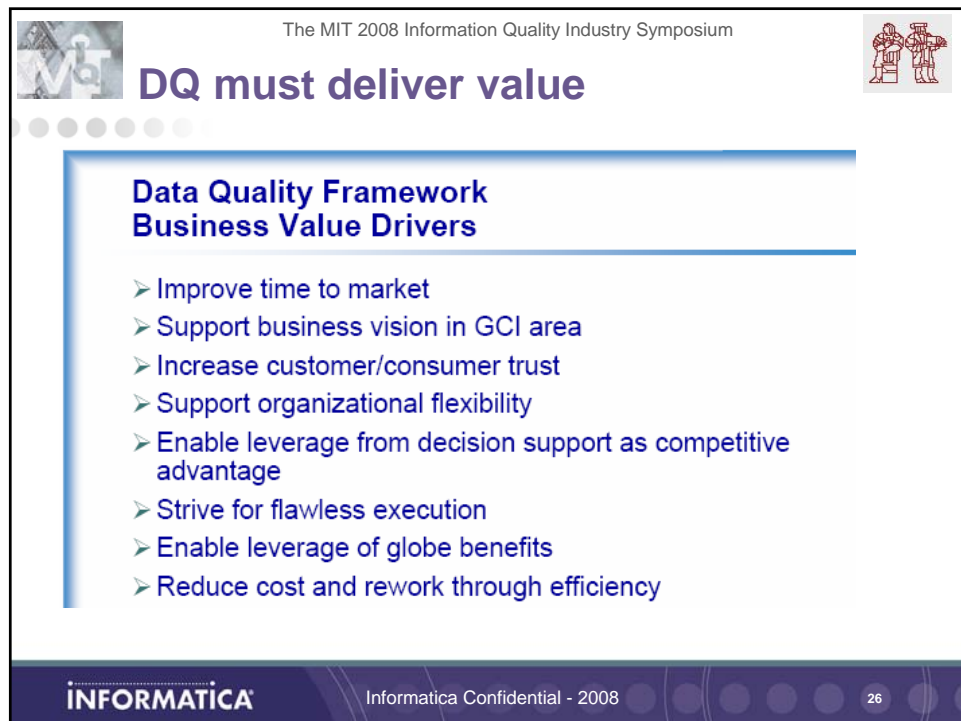
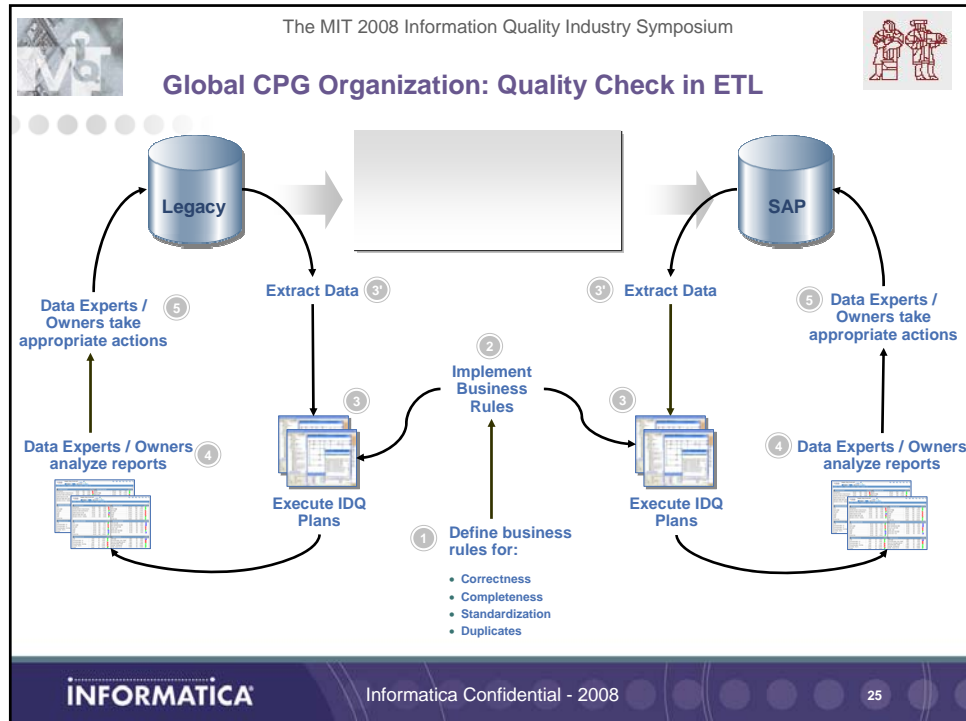
Informatica Confidential - 2008



22












The MIT 2008 Information Quality Industry Symposium

## DQ underpins Data Governance

What we are doing at A.N.Other CPG Org.



**Achieving Data Governance thru the monitoring of the critical business rules.**

- Defining business rules to support data quality per business object (Material, Customer, Vendor, Employee, Banks etc.)
- Defining and implementing Global Data Quality and Data Management KPIs and exception reports based on the pre-Defined business rules.
- Providing Global visibility of Data Quality to the whole organisation at all levels:
  - Publishing the Data KPIs monthly on the operation site on the intranet accessible by A.N.Other organisation at all levels.
  - Presenting the Data KPIs at the Operational Steering Committee (Top Management Level) and at the market management level each month.
- Monitoring of the KPIs and the exception reports by the markets to take corrective actions.



Informatica Confidential - 2008


27



The MIT 2008 Information Quality Industry Symposium

## Agenda

- Introduction
- A Pragmatic Approach
- Common Pitfalls and Misconceptions
- Examples / Scenarios & Case Studies
- Q&A



Informatica Confidential - 2008

28



**INFORMATICA**

*Questions ?*

Thank You

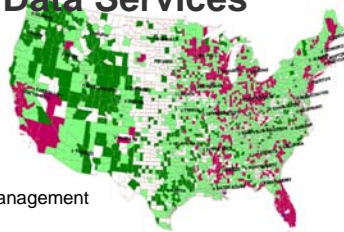
Informatica Confidential - 2008

29

The slide features a dark blue horizontal band across the center. The Informatica logo is on the left, and the text 'Questions ?' is in a large, white, italicized font. Below it, 'Thank You' is written in a smaller white font. A series of small white dots is positioned between the logo and the text. The background of the slide is light gray with faint circular patterns. The footer contains the text 'Informatica Confidential - 2008' and a small number '29' in the bottom left corner.



## Geography Reference Data Services



Justin Magruder, Vice President  
Enterprise Information Strategy & Management

Diane Schmidt, Senior Director  
Enterprise Information Strategy & Management

Xinhua Chris Deng, Senior Information Architect  
Enterprise Architecture

### Agenda



- Introduction (Justin)
- Geography Business Case (Justin)
- RDS IQ Architecture (Chris)
- Geography Approach, Status and Results (Diane)
- Conclusion (All)



## Introduction

- Freddie Mac is a stockholder-owned corporation chartered by Congress in 1970 to keep money flowing to mortgage lenders in support of homeownership and rental housing.
- The company's Information Strategy is driven by operational and technical organizations that have been established to:
  - » Integrate and rationalize functions across the company
  - » Develop and manage standards and procedures
  - » Develop and manage shared information repositories

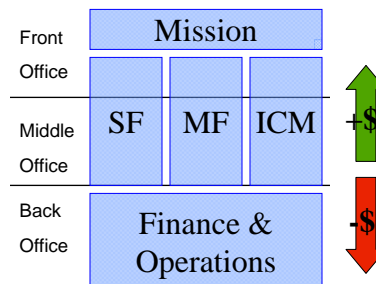
© Freddie Mac 2008

3



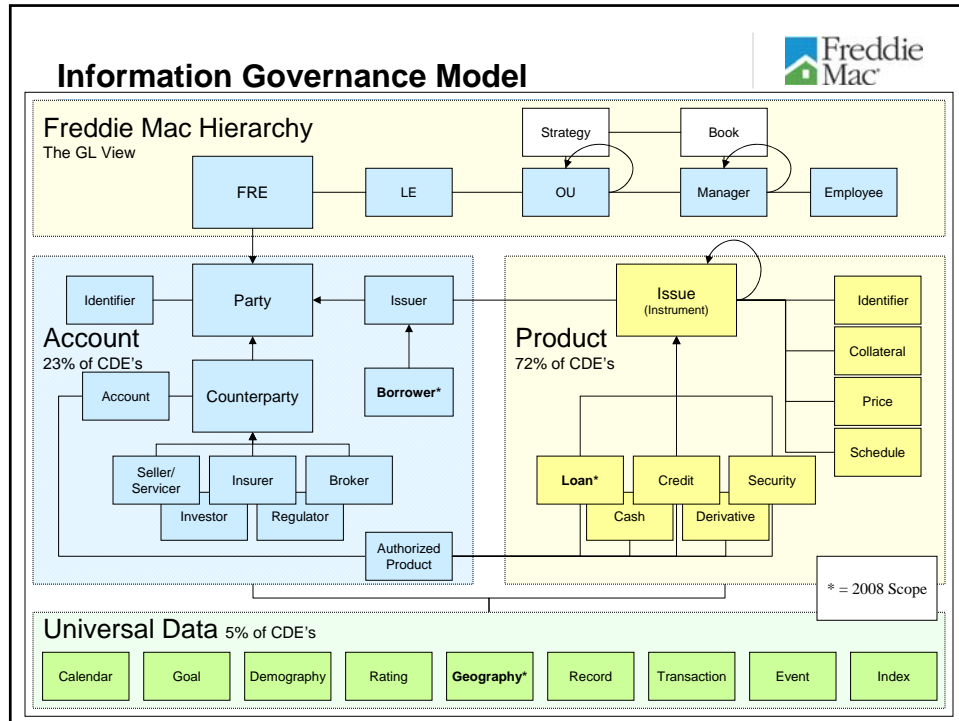
## Information Strategy


- Shift from remediation to strategic priorities
- Develop "reference data services" to improve product development, simplify business & operational processes, and technology
  - » Manage information as we do financial assets
  - » Capture, enrich and master data where it is originated, and distribute with data quality controls
- Reuse, extend, improve: leverage existing infrastructure & analyses
- Define incremental plan with periodic measurable benefits
- Build self-funding business cases
- Establish enterprise governance program & information architecture standards



© Freddie Mac 2008

4





We make home possible®

# Geography Business Case

© Freddie Mac 2007

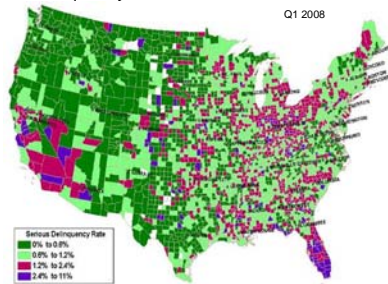
## Why Geography?



- Real estate is mortgage collateral
  - »Where is the property?
  - »What kind of property is it?
  - »What is nearby that drives value?
  - »What is happening in local markets?
- Gain insights into the market:
  - »Where are opportunities?
  - »Where are the risks & problems?
  - »What are root causes?
  - »What is the outlook?

Delinquency Rates in Prime Markets

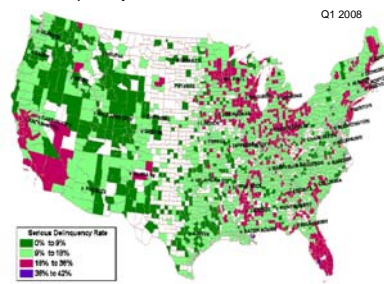
Q1 2008



© Freddie Mac 2008

Delinquency Rates in SubPrime Markets

Q1 2008



7

## Problem Statement



- Currently, Freddie Mac creates and manages geography data in at least 14 different business areas using multiple applications and redundant interfaces.
- Business areas have similar or identical geography data business and technical requirements
- Multiple interfaces lead to data integrity, control and consistency issues
- Increased costs to enhance and maintain numerous applications and interfaces

© Freddie Mac 2008

8



## RDS Information Quality (IQ) Architecture

© Freddie Mac 2007

### Information Architecture for IQ: Principles



- Establish a corporate governance process, including the development of shared Information definitions, common quality measurement standards and procedures, threshold and metrics for management, and decision processes.
- Understand volume, scalability and product development strategies, and align technology initiatives with business strategies.
- Define an integration architecture that can support new product development while isolating legacy platforms from financial systems, such that keep running the cash flow engines while developing new products and platforms.
- Establish common sense metrics and measure Information quality often and with rigor.

© Freddie Mac 2008

10

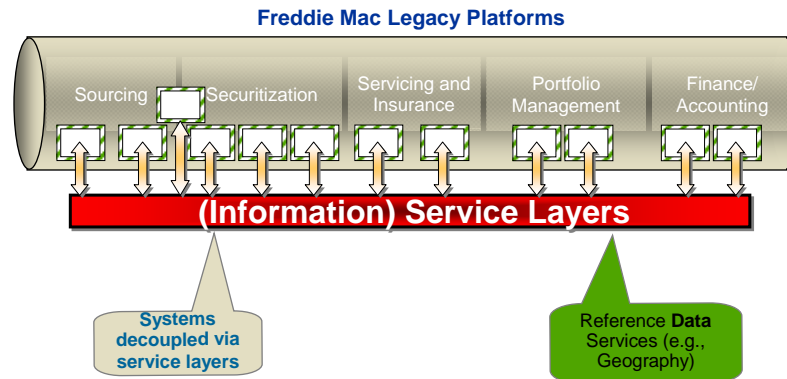


## Service Layer Integration Architecture



### Highlights:

- Decouple FRE legacy systems and platforms via Service Layers
- Reference Data Services is a key to Service Layer Integration Architecture



© Freddie Mac 2008

11

## Information Quality & Control Architecture



### Quality from Data Modeling:

Information is modeled to ensure understanding, and the data is well structured and data quality is included

### Quality from System of Records:

Best-Breed Authoritative list of SORs is developed and maintained, and its use across corporate is enforced

- Data Quality is included into Canonical data modeling
- Develop and maintain an authoritative list of SORs to enforce Information Quality across the enterprise
- Owners will need to create their information in a controlled and quality way, register and provide sufficient metadata for future consumers.

© Freddie Mac 2008

12

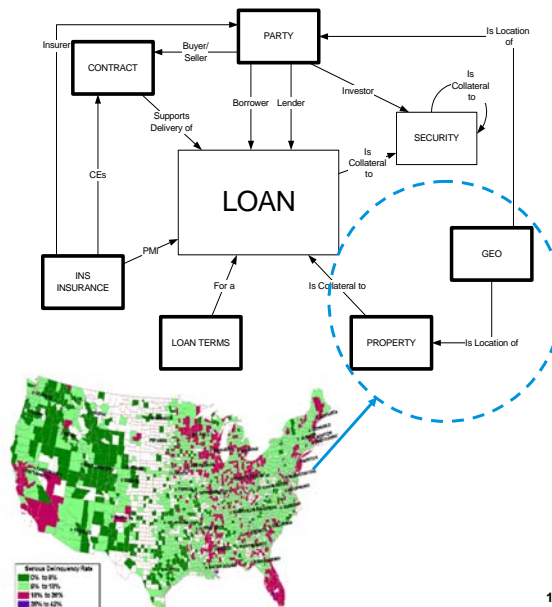


## Geography RDS Approach

© Freddie Mac 2007

### Why Geography?

- Property and geography data are critical to managing our affordable housing goals
  - Collateral to a loan
  - Used for national efforts: Katrina, conforming loan limits
- Supports multiple initiatives across the enterprise and domains: loan (property), counterparty (entity location, market area), regulatory reporting, risk management.
- Data is relatively static
- Concepts are well defined across various domains



© Freddie Mac 2008

14

## Loan.Geography Requirements

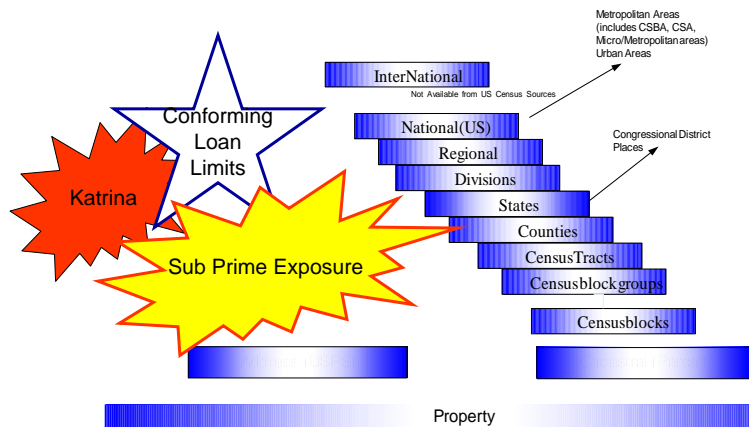


- Geographical information is used by Freddie to manage risk across its inventory of 15 million mortgages
  - » Mission Goals
  - » Loan Sourcing & Securitization
  - » Guarantees & Investment Management
  - » Risk Management
- Geocodes identify assets in the mortgage market
  - » Emergency legislation in February 2008 identified 43 unique metropolitan areas with new conforming loan limits based upon average home price
  - » During Hurricane Katrina, portfolio risk managers needed to identify affected assets in near real time
  - » When combined with demographic or market data, geographic data provides dramatic and illustrative decision criteria

© Freddie Mac 2008

15

## Geographical Hierarchy



FIPS State Code	FIPS County Code	Census Tract Code	Census Block Group	Census Block ID	CSBA Code	CSBA Metro Flag	CSBA Code	CSA Code	MSA Code	MCD / CCD Code	County Name	GCP Census Code Names	CSBA Name	CSBAID
-----------------	------------------	-------------------	--------------------	-----------------	-----------	-----------------	-----------	----------	----------	----------------	-------------	-----------------------	-----------	--------

RDS-GEO Output Fields

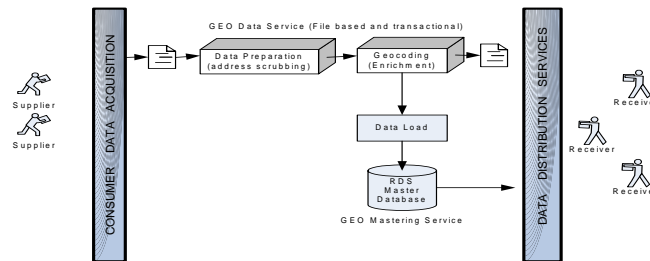
© Freddie Mac 2008

16

## Loan.Geography Requirements



- The Loan.Geography Reference Data Service provides:
  - » Scrubbed, mastered, geocoded-property data for the Master Data Service
  - » Data services to support a wide range of users with different requirements
  - » Common interface through which applications get access to MDS
- The service produces a cleansed and validated address for subscribers
- These capabilities represent the basic level 2 capabilities of Freddie's Data Management Maturity Model



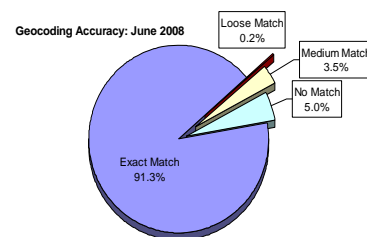
© Freddie Mac 2008

17

## Loan.Geography Status and Results



- In April 2008, the first iteration of the RDS Geography service was implemented into production.
- A customer test environment (CTE) was also established to support non-production related geography needs.
- Enterprise recognized operational process to include:
  - » Monthly Group 1 file updates
  - » Business and technical support model
  - » Subscriber procedures for both the production and CTE environments
- Cost savings will result from reduction in the following:
  - » Number of resources necessary to support business functions
  - » Software license reduction
  - » Cost avoidance related to requirements of new systems
- Year to date Results:
  - » Address scrubbing and geo-coding service level agreements met
  - » Significantly improved performance by ~7 to 10 times faster than legacy process
  - » Provided additional geo-coded data elements for subscribers
  - » Enhanced corporate enterprise geography model to include data required for Mission goal counting
- Management information and metrics are leveraged to continually monitor results



© Freddie Mac 2008

18

## Conclusion



- Freddie can improve the way it manages data
- The key to managing core data assets is through a reference data strategy
- Understanding Geography data is critical to our corporate mission of affordable housing goals

## Q&A



- ...



*“Build to Share”*

# *Federal Data Quality Guide: A Framework for Better Information Sharing*

*July 2008*

---

U.S. Federal Data Architecture Subcommittee



## Agenda

- ◆ Document Purpose and Intended Outcome
- ◆ Federal Data Quality Guide Overview
- ◆ Examples of Federal Agency Data Quality Practices
- ◆ About the Data Architecture Subcommittee (DAS)





## Purpose

- ◆ Few agencies practice data quality at the enterprise and extended enterprise levels
- ◆ The Federal Data Quality Guide advises agencies on the key components needed for an effective enterprise-wide data quality improvement program



3



## Intended Outcome

- ◆ Data quality programs among Federal agencies and Communities of Interest (COIs) align to a common description of data quality improvement practices
- ◆ Information that is shared improves in quality
- ◆ Decision support in agencies and COIs improve

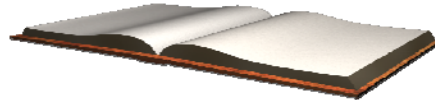


4



## Federal Data Quality Guide Overview

- ◆ Build a data quality framework using EA
- ◆ The business case for data quality
- ◆ Value proposition using the reference models
- ◆ Data Quality Improvement implementation
- ◆ Advice on data quality tools
- ◆ Suggested additional reference material

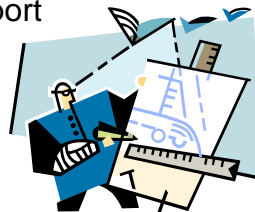


5



## Key Advice Use Existing EA Program

- ◆ Establish data quality procedures and practices into existing agency and community of interest business processes that are part of their Enterprise Architecture (EA)
  - ◆ Provides a framework for improved information sharing and decision support



6





## Data Quality Improvement: *The Challenge*

- ◆ Federal agencies and COIs have struggled with coordinated approaches to the quality of disseminated information due to:
  - ◆ Complexities of size and scope
  - ◆ Need to standardize and modernize technology and information technology (IT) processes
  - ◆ Internal management shortcomings



7

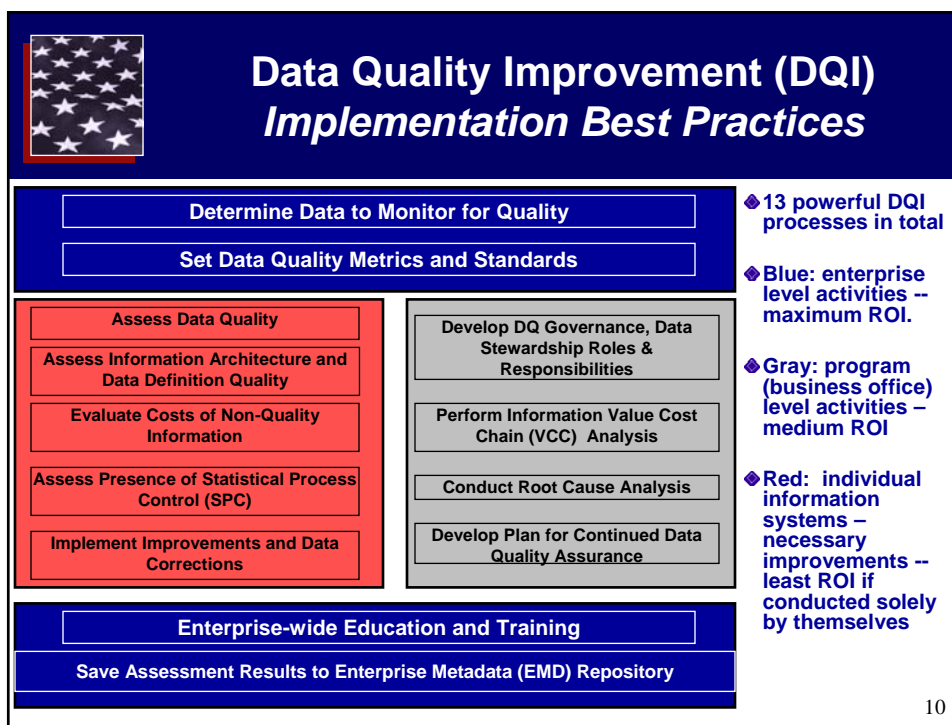
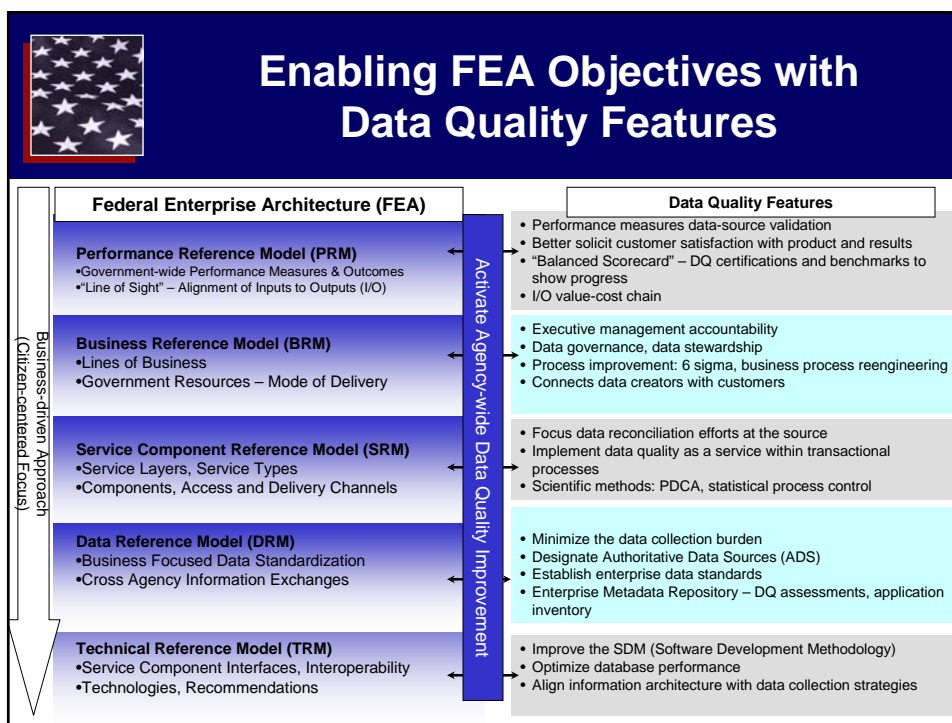


## Business Case for Enterprise-wide Data Quality Improvement

- ◆ Data Quality Improvement (DQI) provides agencies and COIs with repeatable processes for:
  - ◆ detecting faulty data,
  - ◆ establishing data quality benchmarks,
  - ◆ certifying (statistically measuring) their quality, and
  - ◆ continuously monitoring their quality compliance



8





## Some Agency Examples

- ◆ Agencies that have strong data quality programs at the enterprise level



- ◆ Defense Logistics Agency



- ◆ Housing and Urban Development

11

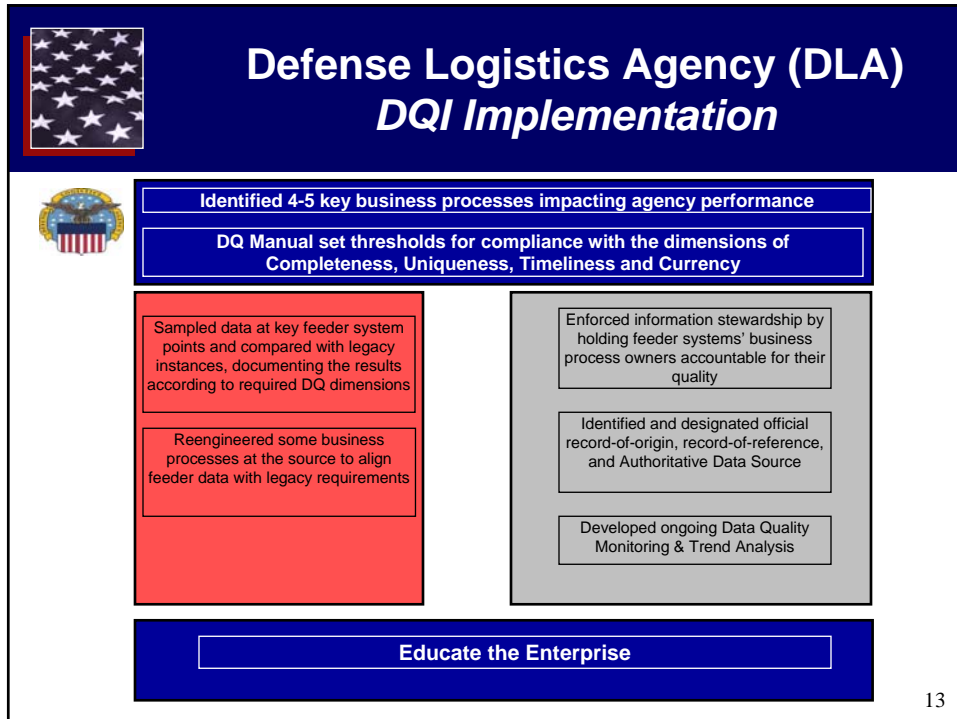


## Defense Logistics Agency (DLA) *Data Quality Challenges*



- ◆ Building understanding of data and functional process flows of four feeder data systems into a DLA portal
- ◆ Analyzing multiple data entry points of the same classes of mission-critical data
- ◆ Determining authoritative source for multiple data “instances”
- ◆ Determining data stewardship responsibilities

12



**Defense Logistics Agency (DLA)  
Internal DQI Scorecard**

	Enterprise Level (minimal DQI impact felt here)	Program Level (most DQI impact felt here)	System Level (modest DQI impact here)
<b>Successes</b>	<ol style="list-style-type: none"> <li>Some key business processes and their sequencing (operational "racetrack") developed for first time</li> <li>DQ Manual developed with metrics and standards</li> </ol>	<ol style="list-style-type: none"> <li>Data Integrity Branch (DIB), program area stewardship defined</li> <li>Data Quality Monitoring &amp; Trend Analysis program taken up by DIB</li> </ol>	<ol style="list-style-type: none"> <li>Assessment points for sampling feeder data developed strategically</li> <li>Reengineered some business processes to decrease data redundancy</li> </ol>
<b>Challenges remaining</b>	<ol style="list-style-type: none"> <li>EMD Repository solution required</li> <li>Training required across the enterprise</li> </ol>	<ol style="list-style-type: none"> <li>Authoritative Data Source (ADS) analysis completed, but full information Value Cost Chain from feeders to legacy not understood</li> </ol>	<ol style="list-style-type: none"> <li>Refining Statistical Process Control methodology</li> <li>Determining ROI for DQ improvement</li> <li>Defining investment threshold for reaching point of diminishing return</li> </ol>

14



## Housing & Urban Development (HUD) *Data Quality Challenges*



Information Architecture required redesign to better support accuracy and quality of information exchange

### Legacy Grants Monitoring System

#### Business Goal:

- Support job creation in underprivileged areas

#### Reporting Method:

- Data from multiple collection points aggregated to report on job creation statistics in HUD's Annual Performance Plan

#### Challenge:

- Allowable data entry points did not use common method to convert jobs data

15



## Housing & Urban Development (HUD) *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 EDM staging area

16



## Housing & Urban Development Internal DQI Scorecard

	Enterprise Level (some DQI impact felt here)	Program Level (modest DQI impact felt here)	System Level (most DQI impact felt here)
<b>Successes</b>	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
<b>Challenges remaining</b>	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



## Data Quality Tools Advice

Enabling tools for data quality at minimum:

- ◆ Data Profiling (Business Rule Discovery)
- ◆ Data Defect Prevention
- ◆ Metadata Management
- ◆ Data Re-engineering and Correction





## Current Status

- ◆ The Federal Data Quality Guide is in draft form undergoing review.
- ◆ A copy of the draft is available on the Data Architecture Subcommittee collab site on Core.gov.
- ◆ A copy of the draft can also be obtained via e-mail request: [suzanne\\_acar@ios.doi.gov](mailto:suzanne_acar@ios.doi.gov)

19



### About the Federal Data Quality Guide Authors:

**Federal Data Architecture Subcommittee  
(DAS)**

20



## Data Architecture Subcommittee

### ◆ Federal Data Architecture Subcommittee (DAS) Facts

- Chartered by the Federal CIO Council
- 2 appointed Co-chairs
  - Suzanne Acar, DOI
  - Adrian Gardner, NWS
- Membership Federal CIO representation + contributors (135)
- Eight work groups



### ◆ Key FY08/09 Activities/Deliverables

1. Federal Data Quality Guide
2. Final Draft Person Framework Standard
3. DRM Implementation Guide

21



## Summary

22





## Summary

- ◆ The Federal Data Quality Guide informs agencies on features of an enterprise-wide data quality program.
- ◆ The key advice is to leverage existing EA programs.
- ◆ The outcome is improved information sharing, interoperability, and decision support.
- ◆ Supports key principle to manage information as a national asset.

23




## Questions



*Contact info:*

Mark Amspoker  
Lead, Federal DQ Guide Working Group  
Citizant, Inc.  
[mampoker@citizant.com](mailto:mampoker@citizant.com)

24






Joe Bugajski  
*Senior Analyst*  
Burton Group  
[jbugajski@burtongroup.com](mailto:jbugajski@burtongroup.com)

Bob Grossman  
*Managing Director*  
Open Data Group  
[rlg@opendatagroup.com](mailto:rlg@opendatagroup.com)

## Data Governance for Improved Information Quality

MIT IQ Industry Symposium  
Cambridge Massachusetts USA  
16-17 July 2008

All Contents © 2008 Burton Group and Open Data Group. All rights reserved.





## Data Governance for Improved IQ

2

### Thesis

- Good Information Quality (IQ) needed to comply with Sarbanes Oxley, EU Privacy, and Anti-Terrorism Acts
- Audits, business process controls and spot checks are necessary but insufficient to assure good IQ
- Data field reuse changes meaning of information
  - May result in sensitive data stored in unprotected data fields
  - Increases transaction failure risks that reduce profitability
- Data Authority Reference Model lowered liabilities and recovered USD \$2 billion annual sales
  - Measure and improve semantic reliability
  - Provides effective data governance for high IQ



## Data Governance for Improved IQ

3

### Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions





## Data Governance for Improved IQ

4

### Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions



## Authors



5

Joseph M. Bugajski: Senior Analyst, Burton Group Inc.

- Research data and technology governance, architecture, information quality and standards. Formerly Chief Data Officer at Visa Inc. responsible worldwide for information architecture and quality. Previously, Visa chief enterprise architect and resident entrepreneur for information strategy, risk systems, and data warehousing. Before Visa, CEO of Triada, a business intelligence software firm.

Robert L. Grossman: Managing Partner, Open Data Group

- Deliver management consulting, outsourced analytic services, and analytic staffing for data mining; Director at the National Center for Data Mining and Prof. at University of Illinois at Chicago (UIC). Also, Chair of the Data Mining Group (DMG) and advisory board member of InfoBlox. Before Open Data Group, founder and CEO of Magnify a data mining software provider to financial services. Before Magnify, was co-founder and co-director of the National Scalable Cluster Project, and co-founder and co-director of the PASS project.

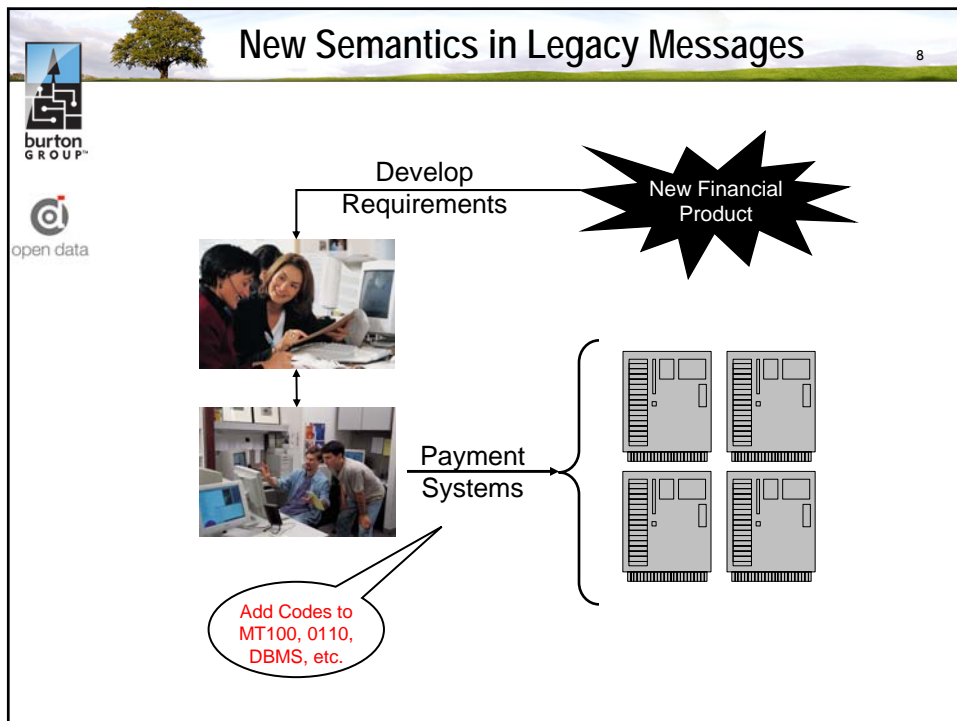
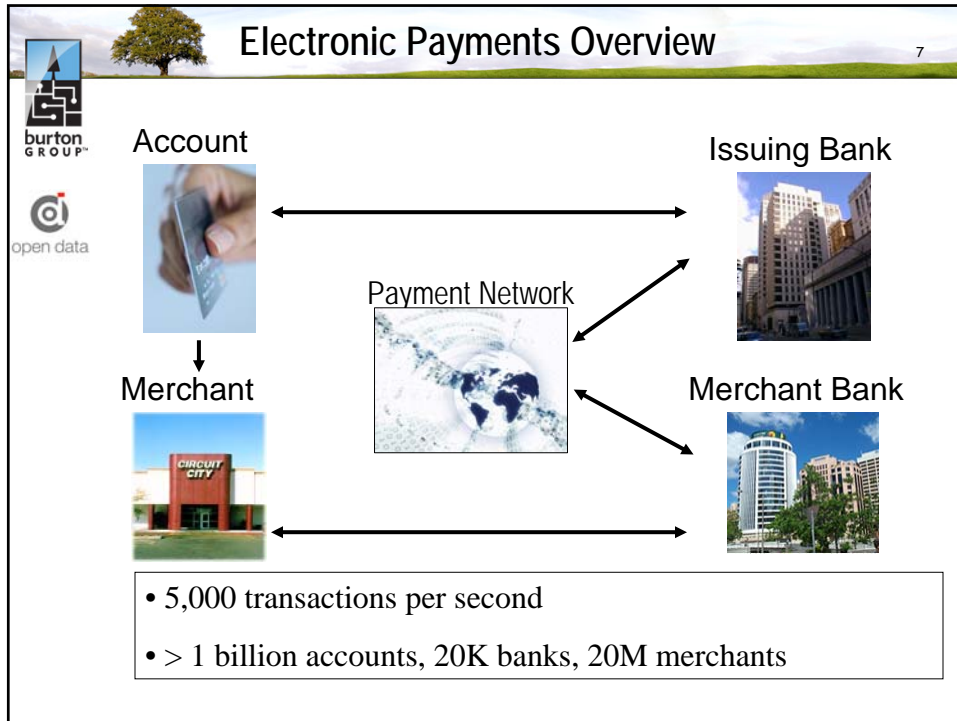


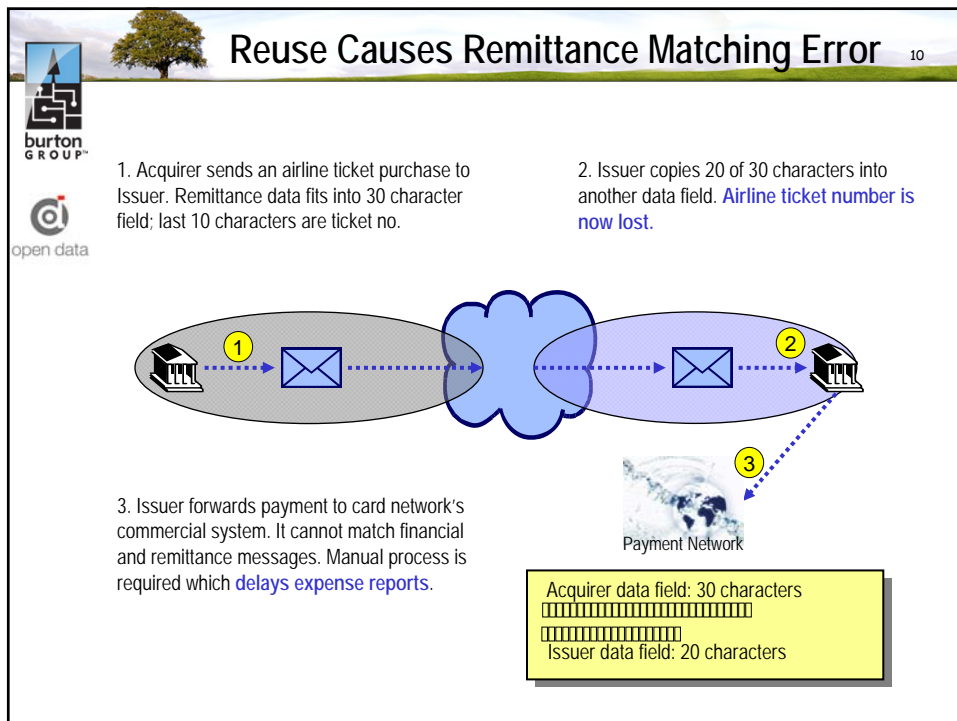
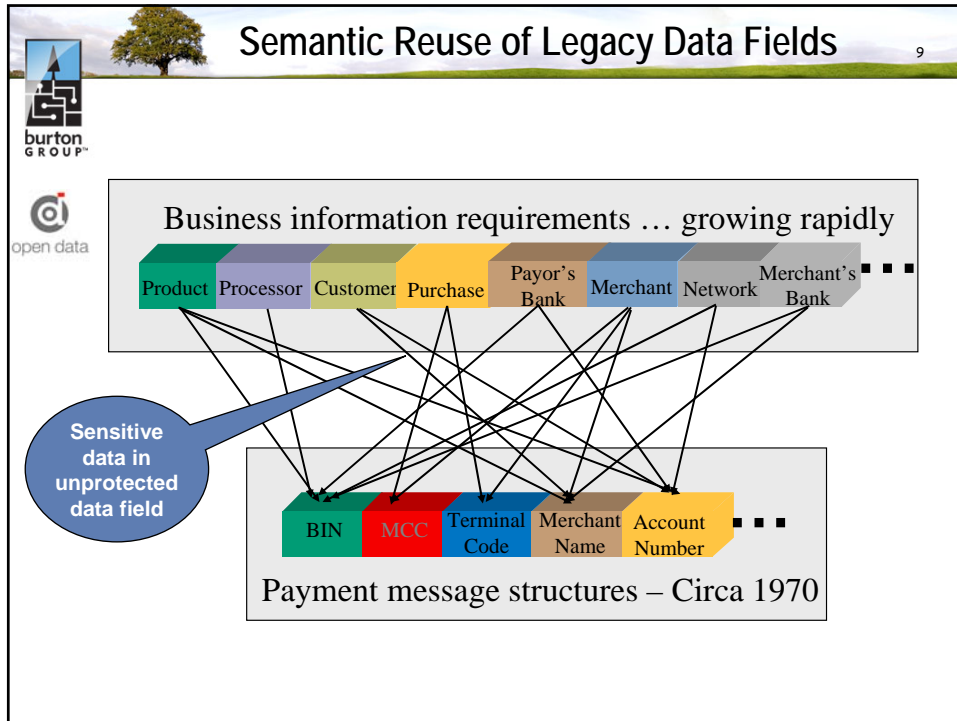
## Data Governance for Improved IQ



6

### Agenda

- Authors
- **Managing information in a global payment network**
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions







## Information Quality Challenges

11

### Problem Definition

- High costs and long lead times for adapting legacy systems for new products demands data field reuse
- Data field reuse creates confusing semantics and it may let sensitive data into unprotected data fields
- Multiple parties are involved in message networks, any of whom can modify data that heighten risks
- Faulty messages increase processing costs, result in lower sales revenue, and raise liability risks
- Data governance is required to lower risks





## Data Governance for Improved IQ

12

### Agenda

- Authors
- Managing information in a global payment network
- **Data Authority Reference Model**
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions

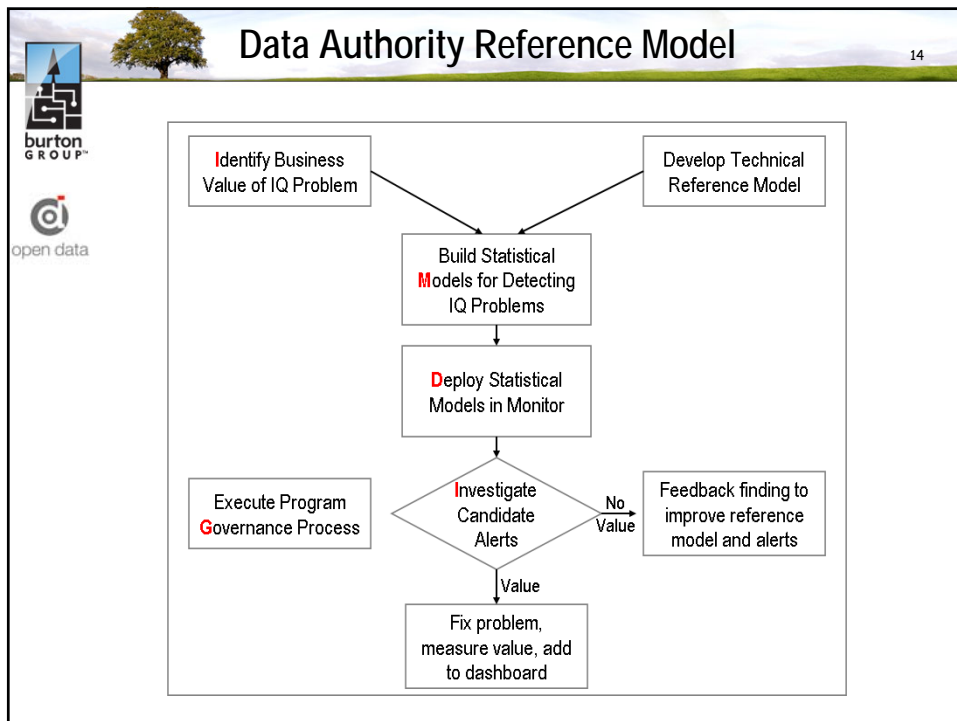



## Data Authority Reference Model




13

### Identify, Model, Deploy, Investigate and Govern (IMDIG)

- Identify business problem associated with IQ problem
- Model – build statistical and rule-based models to monitor data; develop data architecture practice
- Deploy – implement data monitor in production
- Investigate – problems highlighted during monitoring
- Govern – organize data governance team









## Data Authority Reference Model

15

### Organizing the Data Authority Reference Model

1. Review data architecture across all I.T. systems
2. Measure production data quality (and interoperability)
3. Establish priorities for correcting data problems
4. Win commitment of CIOs to correct problems
5. Form a data governance team of experts, business and technical – team reports to CIOs
6. Build consensus solutions to top priority data issues
7. Write and maintain a Technical Reference Model
8. Report measurable progress to CIOs at least quarterly



## Data Governance for Improved IQ

16

### Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- **Measurements of Change Detection Using Cubes of Models (CDCM)**
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions

## CDCM Data Measurements

17

Apply CDCM at “tap points” across processing systems

- Tap at points to get “real” data – not cleansed
- Build baseline CDCM to compare with current sample

$$f(\text{CDCM Monitor}) = \text{CDCM Monitor}$$

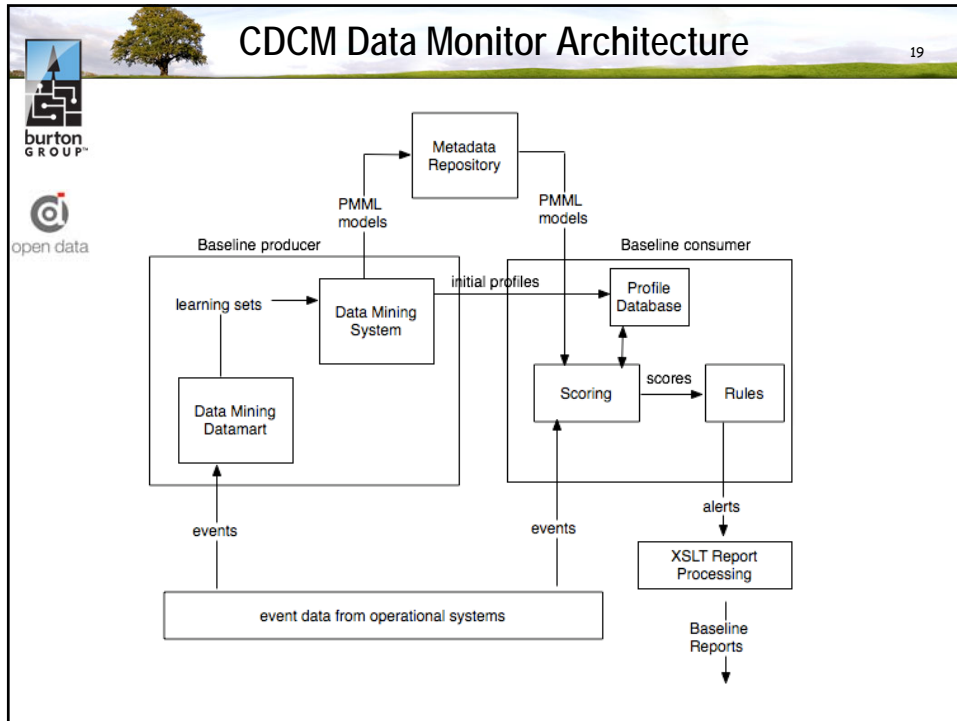
## CDCM Logical Design

18

Focus on Baselines & Changes

- Divide & conquer data (segment) using multidimensional data cubes
- For each cube, establish separate baselines by data quality dimension (e.g. completeness, validity, ...)
- Detect changes from baselines

Separate baselines for completeness, validity, etc.





### Example: Bivariate Distribution Change

The example shows a change in the bivariate distribution of values. The 'Baseline Model' and 'Observed Model' are compared.

Baseline Model	
Value	Percentage
90, -	0.13
90, blank	0.21
05, -	0.01
05, blank	0.01
etc.	etc.
<b>Total</b>	<b>100.00</b>

Observed Model	
Value	Percentage
90, -	0.13
90, blank	0.63
05, -	0.01
00, blank	0.11
etc.	etc.
<b>Total</b>	<b>100.00</b>

Logos for 'burton GROUP' and 'open data' are visible in the top left corner.



## CDCM Launch Summary

21

### Change Detection using Cubes of Models implementation

1. Select "best" data tap point(s) into production systems
2. Build baseline development and production monitor
3. Develop scoring process by quality measures: completeness, validity, etc.
4. Build baseline CDCM models and add to monitor
5. Score data from tap points – highlight serious problems
6. Generate trouble alerts then start improvement process
7. Track improvements and report results in CIO dashboard
8. Update Data Authority Reference Model with findings





## Data Governance for Improved IQ

22

### Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- **Building a Global Data Governance Program**
- Recapturing lost revenue through improved IQ
- Additional work required
- Conclusions



## Create Data Governance Program

23

Data governance program designed to deliver business value

- Use data architecture study and baseline data quality measurements to focus program on fixable problems
- Define measurement and correction program and build support for it among all data experts
- Ask finance for estimate of monetary value of fixes
  - Build a financial model; e.g., a P&L for program; and live with it!
  - Set financial objectives for top problems list
- Earn support for program at “C” level; start with CIOs
  - Set realistic monetary value recovery objectives
  - Make business case using simple and real examples of problems
  - Assure proper level of staff support to actually fix problems




## Organize Data Governance Team

24

Team members are data experts with “C” level credentials

- Data experts are from business units and I.T. groups
- Choose members with knowledge and collaboration skills
- Have members approved by their “C” level executives
- Host a “kick-off meeting of data governance team
  - Agree upon financial objectives and dashboard reporting
  - Agree upon data quality measurement rules
  - Approve initial version of Data Authority Reference Model
- Meet telephonically at least bi-weekly
- Meet in person at least quarterly





## Lead the Data Governance Team

25

Team members are have real work to do!

- Write problem “alerts” as individual business cases
- Assign each alert to a governance team member
- “C” level granted authority to team members to deliver business value by fixing data problems
- Team members report progress to financial objectives
- Leader solves common problems and meets with CIOs
- Feedback from team members regarding alerts and quality improvements provides foundation which drive
  - Updates to Data Authority Reference Model
  - Improvements to CDCM and Alerting process



## Data Governance Program Review

26

Support data quality improvement with appropriate governance

- Align corporate objectives, program oversight, and alert investigation and improvement processes
- Report to CIOs on dashboard
  - Top priority concerns as agreed by the governance team
  - Strategic objectives met through program operations
  - Total value recovered versus agreed financial objectives
- Adjust alert workflow by refining quality rules, number of CDCM statistical models, and alert value thresholds
- Support the governance team members at all times
- Maintain the Data Authority Reference Model





## Data Governance for Improved IQ

27

### Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- **Recapturing lost revenue through improved IQ**
- Additional work required
- Conclusions

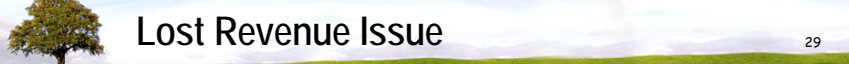




## Customer Satisfaction Problem

28

### Chip Card Terminal Coding Error

- Cards with "contact chips" common outside USA
- Terminals in USA read "contactless chips"
- Data field codes not well defined in technical manuals
- CDCM monitor notes contact chip activity in USA
- VIP customer shortly thereafter receives payment decline
- Business case for solution is clear
  - Update technical manuals with clearer definitions of chip codes
  - Send technical letter terminal suppliers and banks with correct codes
- Problem resolved within weeks

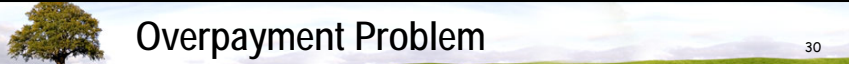




## Lost Revenue Issue

29

### Incorrect Country Code Error

- Payments messages record location of merchant
- CDCM finds mismatch between country and city
- Problem results in higher than normal rate of declines
- Inappropriately declined payments reduces revenue
- Governance team member worked with merchant's bank to define data problem and a business case for fixing it
- Problem resolved within a few weeks with corresponding increase in payment revenue





## Overpayment Problem

30

### Incorrectly Coded Sales Channel

- Payments messages record sales channel and terminal type; e.g., face-to-face and card present or e-commerce
- Channel and terminal data helps with risk assessment and also can be used to set payment processing fees
- CDCM finds mismatch between channel and terminal
- Problem leads to higher than expected payment fees
- Governance team member worked with bank to define data problem and the business case for fixing it
- Problem resolved within a few weeks with corresponding reduction in payments followed by additional business





## Sensitive Data Values in Wrong Field

31

Data field reuse commonly applied for new products

- Payment messages designed to 1987 ISO standard
- These legacy messages used in thousands of systems
- Infrastructure costs for changing systems is HUGE!
- Reuse (overloading) of data fields is a best practice for recording data needed for new products
- Certain types of payment products required more information about cardholders than message held
- Decision to reuse data field for personally identifiable, non-public information reversed by governance team





## Data Governance for Improved IQ

32

Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- **Additional work required**
- Conclusions





## Data Semantics Wrong from the Start

33

70+ case studies show problems start with data definitions

- Confusion about business definitions for data values
- Confusion about ways to correctly code data values
- Confusion about meaning of messages expressed by combinations of data values
- Reuse of legacy data fields for new semantic value adds to confusion and admits potential for sensitive data entries into fields where controls might be less stringent
- Complicated, unrecorded, metrics for data analysis result in inconsistent risk analysis and reporting





## Financial Industry Tries to Fix Semantics

34

Work begun at ISO, UNCEFACT, OMG, other standards groups

- Solution to semantic variability problem is development of accurate, computer readable data models (e.g., UML)
- ISO 20022 (UNIFI\*) standard developed to create XML messages from computer data models
- UNCEFACT Core Components applies technology similar to UNIFI to update EDI specification
- OMG Conversion Models for Payment Messages (CM4PM) applies UNIFI to generate legacy messages
- XBRL, FIX, IFX and others intend to follow suit

UNiversal Financial Industry (UNIFI) message scheme



## Solutions Still Needed

35

High Value for solution to semantic variability problem

- Problem with defining elemental business terms
- Problem reusing elemental business terms in messages
- Modeling tools emit inconsistent XML form of models
- Software development tools today do not share models
- Metadata repositories do not support data governance
- Data models repositories incapable of global scale
- Full scale tests of technology stack remains incomplete
- Lingering belief that “yet another message format” will solve all these data semantics problems





## Data Governance for Improved IQ

36

Agenda

- Authors
- Managing information in a global payment network
- Data Authority Reference Model
- Measurements of Change Detection Using Cubes of Models (CDCM)
- Building a Global Data Governance Program
- Recapturing lost revenue through improved IQ
- Additional work required
- **Conclusions**



## Conclusions and Summary

37

### Data Authority Reference Model Program Delivers Value

- Program very effective for recovering business value and improving data quality
  - Authors recovered USD \$2billion for payment company over three years of program operation
- Identify the problem, Model the data, Deploy monitors, Investigate problems, and Govern program (IMDIG)
  - Build consensus about data problems and business value
  - Develop Change Detection using Cubes of Models (CDCM)
  - Monitor production data
  - Create alerts and track solutions
  - Governance team of "C" level delegates own dashboard reports about top problems, strategic accomplishments and value recovered



## Recommendations

38

### Build governance that returns value to your business

- Develop a Data Authority Reference Model program in your business or agency
- Report your results to this forum, ICIQ, DAMA
- Help design standards that address global problem with semantic variability
- Develop tools and technologies to solve open issues
- Continue research into root causes of data problems





## Data Governance for Improved IQ

39

### References

- Joseph Bugajski, Robert L. Grossman, An Alert Management Approach to Data Quality: Lessons Learned from the Visa Data Authority Program, 12<sup>th</sup> International Conference on Information Quality (ICIQ), 2007
- Joseph Bugajski and Philippe De Smedt, Assuring Data Interoperability Through the Use of Formal Models of Visa Payment Messages, 12<sup>th</sup> International Conference on Information Quality (ICIQ), 2007
- Joseph Bugajski, Robert L. Grossman, Chris Curry, David Locke, and Steve Vejcek, Data Quality Models for High Volume Transaction Streams: A Case Study, 13<sup>th</sup> International Conference on Knowledge Discovery and Data Mining, 2007
- Joseph Bugajski, Robert L. Grossman, Eric Sumner and Steve Vejcek, Monitoring Data Quality for Very High Volume Transaction Systems, Proceedings of the 11th International Conference on Information Quality, 2006
- Leo L. Pipino, Yang W. Lee and Richard Y. Wang, Data Quality Assessment, Communications of the ACM, Volume 45, 2002, pages 211-218
- The Predictive Model Markup Language (PMML), [www.dmg.org](http://www.dmg.org)
- The Augustus open source data mining system can be downloaded from [www.sourceforge.net/projects/augustus](http://www.sourceforge.net/projects/augustus)
- ISO 20022 (UNiversal Financial Industry message scheme), [www.iso20022.org](http://www.iso20022.org)
- Conversion Models for Payment Messages (CM4PM), approved and awaiting publication at [http://www.omg.org/technology/documents/spec\\_summary.htm](http://www.omg.org/technology/documents/spec_summary.htm)



## Data Governance for Improved IQ

40

# Thank You! & Questions?

<b>Joe Bugajski</b> <i>Senior Analyst</i> <i>Burton Group</i> <a href="mailto:jbugajski@burtongroup.com">jbugajski@burtongroup.com</a>	<b>Bob Grossman</b> <i>Managing Director</i> <i>Open Data Group</i> <a href="mailto:rlg@opendatagroup.com">rlg@opendatagroup.com</a>
---	---



The MIT 2008 Information Quality Industry Symposium



## Raising the Bar: DQ/IQ to “Enterprise IQ” Presentation at MIT IQ Industry Symposium July 17, 2008

Garry Darrer  
Getinge USA, Inc.  
Garry.Darrer@GetingeUSA.com

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



### Objectives of this presentation

- Traditional Data Quality and Information Quality:  
Example DQ/IQ Evolution with BI
- From DQ/IQ to “Enterprise IQ” Considerations
  - Organizational Context
    - Complexities
    - Business Strategy/Goals
    - Regulatory
    - Culture
- Summary

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Objectives of this presentation

- Traditional Data Quality and Information Quality: Example DQ/IQ Evolution with BI
- From DQ/IQ to “Enterprise IQ” Considerations
  - Organizational Context
    - Complexities
    - Business Strategy/Goals
    - Regulatory
    - Culture
- Summary

© Getinge USA, Inc. 2008

GETINGE

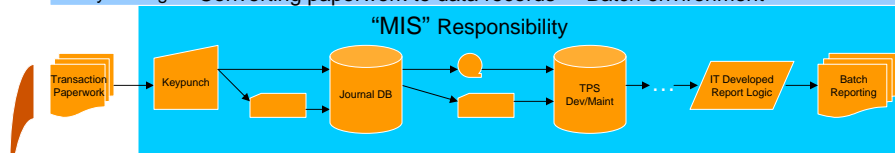


The MIT 2008 Information Quality Industry Symposium



## Example Systems DQ/IQ Evolution

20+ years ago: “Converting paperwork to data records” – Batch environment

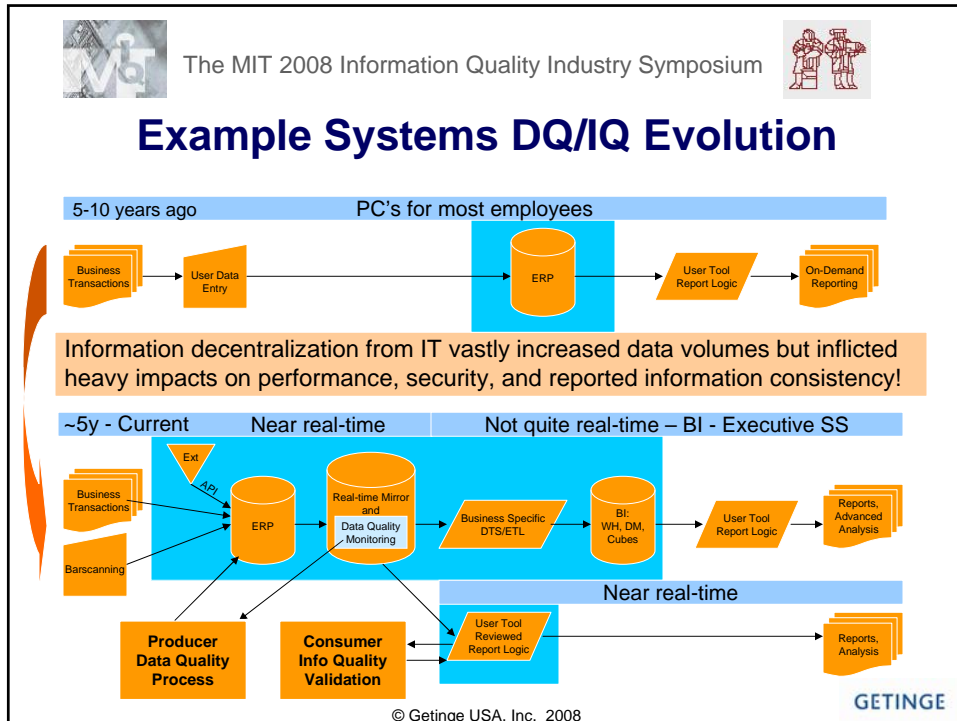


10-20 years ago: Terminals for most employees, PC's for power users



© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium

## Why BI?

- Some Drivers behind (Getinge USA's) BI:
  - Centralized consistent information source and toolset.
  - One analytic view of the customer
  - Enhanced analysis: strategic advantage
  - Ability to “close the books” in record time
- Management embraced BI as a springboard towards a data/information quality culture.

© Getinge USA, Inc. 2008

GETINGE





The MIT 2008 Information Quality Industry Symposium



## Why BI?

- Project significantly involved business management.
  - Required *analyzing* and *understanding* business processes.
  - Inter-relationships and complexities between information producers and consumers clarified.
- The BI project made it necessary: clean (scrubbed) source data...
  - Re-organize the business towards this goal.

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Why BI?

- Data Quality Review Board formed
  - Business Management (Data Stewards)
  - IT
- Continuous Improvement Process.
  - Identification or Hypothesizing of problems and areas for improvement.
  - Design and development of solution
  - Test then Implement
  - Monitor
  - Repeat

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Effective DQ/IQ Technical Mechanism

- Data Quality Monitoring: real-time alerts chosen instead of application edit modifications because of application re-validation requirements.
  - Alerts directly to data producer.
  - **Correction Deadline:** end of business day.
  - Data producers correct their own information
  - Management receives individual visibility and statistical summaries.
  - Managers assure compliance to alerts, devise corrective action such as training or discipline as necessary.

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Effective DQ/IQ Technical Mechanism

- For us, real-time alerts now robust
  - Mass “data scrubbing” is a thing of the past.
- For data consumer - there are two basic expectations:
  - “Realtime” – assume may not yet be scrubbed.
  - “BI” – assume information has been scrubbed.
- In the end: Data Quality is internalized as a normal daily business responsibility.

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Objectives of this presentation

- Traditional Data Quality and Information Quality: Example DQ/IQ Evolution with BI
- From DQ/IQ to “Enterprise IQ” Considerations
  - Organizational Context
    - Complexities
    - Business Strategy/Goals
    - Regulatory
    - Culture
- Summary

© Getinge USA, Inc. 2008

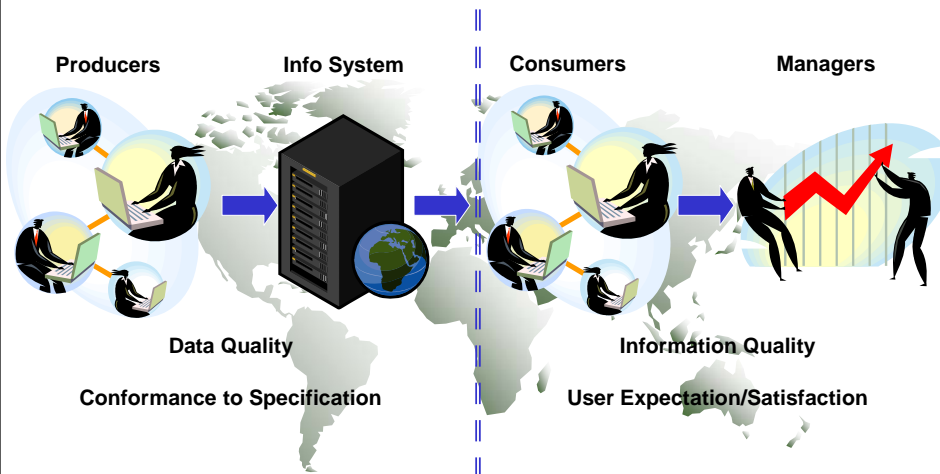
GETINGE



The MIT 2008 Information Quality Industry Symposium

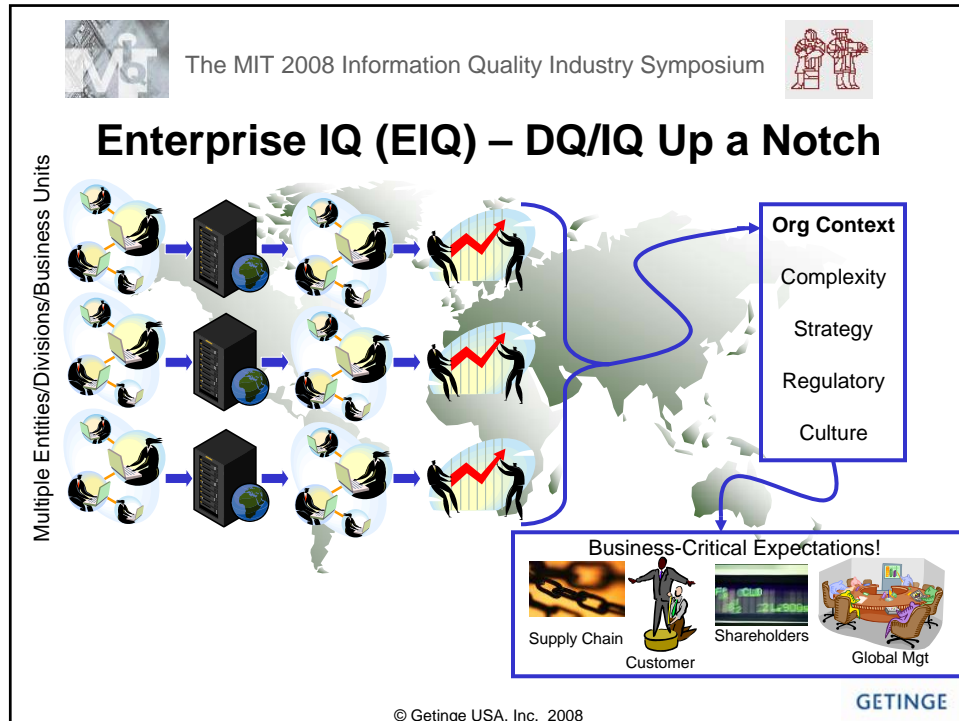


## Traditional DQ/IQ



© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium

## Enterprise IQ (EIQ)

- Especially for complex corporate environments, there are added demands to information, for example:
  - Multiple divisions/business units/sister companies
  - Multi-National Corporations (MNC's)
  - Multiple product families and shared customers/suppliers
  - Combinations of the above!
- Information Shareholders in EIQ are beyond end-users; they include stockholders, group management, and customers/suppliers...

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Objectives of this presentation

- Traditional Data Quality and Information Quality:  
Example: DQ/IQ evolution with BI
- From DQ/IQ to “Enterprise IQ” Considerations
  - Organizational Context
    - Complexities
    - Business Strategy/Goals
    - Regulatory
    - Culture
- Summary

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Complexities - Real-World Example



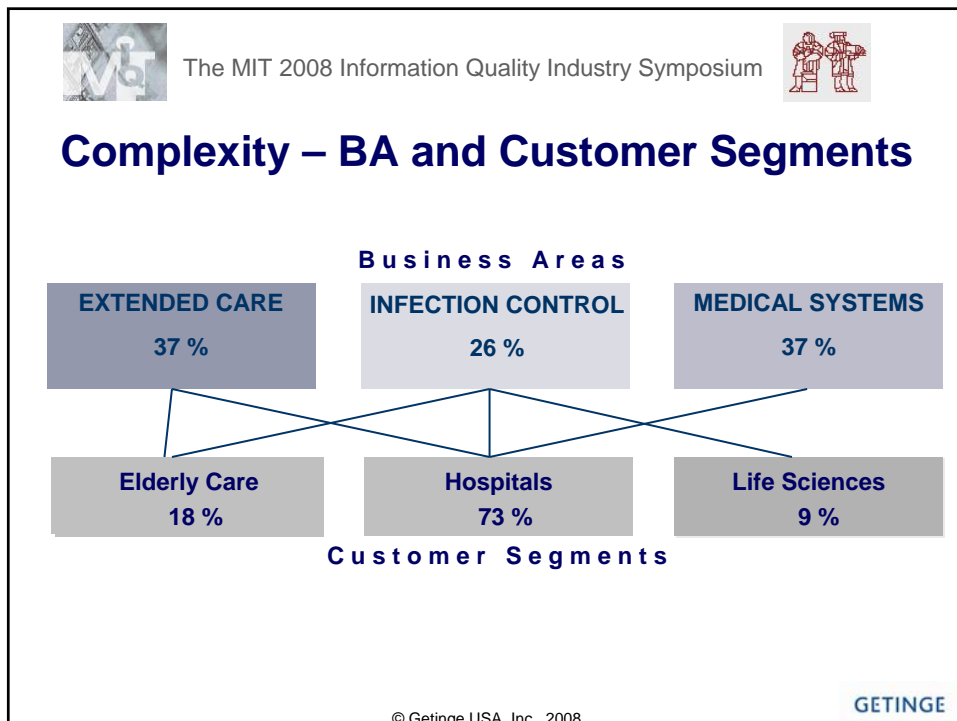
2007 Global Revenue: ~\$2.547B

(16,445 MSEK \* 0.154875 (Dec-31-2007 USD exch rate))

Employees: 10,358

© Getinge USA, Inc. 2008

GETINGE





The MIT 2008 Information Quality Industry Symposium



## Complexity – Infrastructure



© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Complexity - Multiple Applications

- ERP
- CRM
- Document Management
- HR
- QS
- BI/DW
- Budgeting
- Training Management
- CAPA
- NC
- Engineering
- Messaging/Collaboration
- Intra/Inter/Extranet
- eCommerce
- Legal



© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Objectives of this presentation

- Traditional Data Quality and Information Quality:  
Example: DQ/IQ Evolution with BI
- From DQ/IQ to “Enterprise IQ” Considerations
  - Organizational Context
    - Complexities
    - Business Strategy/Goals
    - Regulatory
    - Culture
- Summary

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Key: EIQ Alignment to Business Goals

Business Goals (examples):

- Top-quality deliverables to the customer, exceeding customer expectations?
- Increased Revenue?
- Increased Market Share?
- Reduce re-work and returns to minimize cost?
- Maximize return to investors?
- Increased Focus on Core Competencies?
- **Question:** where does information quality fit in with relation to your company's business goals?

© Getinge USA, Inc. 2008

GETINGE





The MIT 2008 Information Quality Industry Symposium



## IC Strategic Cornerstones: Customers

### COST LEADERSHIP

To utilize the business area's world-leading position to give cost-effective, good-value solutions to customers.

### INTEGRATED SOLUTIONS

To be the best complete solution provider, where Getinge's broad product range and expertise will benefit customers.

### SERVICE

To utilize Getinge's well-developed service network and the Getinge Academy to give customer superior service and optimal use of their investment.

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Strategy – Growth

### Infection Control

- Bolt-on acquisitions to add new technologies and to reach new geographies within existing product lines
- New product lines: Consumables

### Extended Care

- Bolt-on acquisitions to add new technologies or to build critical mass in existing product lines

### Medical Systems

- Bolt-on acquisitions to add new technologies and to reach new geographies within existing product lines
- New product lines: Cardiac surgery

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



		2006	2007
<b>Infection Control</b>	Sterilization	no 1	no 1
	Disinfection	no 1	no 1
<b>Extended Care</b>	Patient Handling	no 1	no 1
	Hygiene Systems	no 1	no 1
	Wound Care	no 4	no 2
	IPC / DVT	-	no 1
<b>Medical Systems</b>	Surgical Tables	no 1	no 1
	Surgical Lights	no 1	no 1
	Ceiling Pendants	no 2	no 2
	Cardiopulmonary	no 3	no 3
	Endoscopic vessel harvesting	-	no 1
	Beating heart surgery	-	no 2
	Anastomosis CABG	-	no 1
	Vascular grafts AAA, TAA	-	no 1
	Ventilation	no 1	no 1

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Leveraging Information for Growth

- Information for a customer-centric approach
  - Listen to customers
  - CRM approach: acquire/enhance/retain customers
    - Areas: sales, service/support, retention/loyalty, marketing, account/contact management.
    - Capture customer information at all contact points
    - Make a customer's information available for all who contact the customer.
  - Genuinely use customer feedback for improving products and services.
  - Leverage information across business units for a customer-centric experience.

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Objectives of this presentation

- Traditional Data Quality and Information Quality: Example - DQ/IQ Evolution with BI
- From DQ/IQ to “Enterprise IQ” Considerations
  - Organizational Context
    - Complexities
    - Business Strategy/Goals
  - Regulatory
  - Culture
- Summary

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Regulatory - Environment

- Multitude of regulatory/governance factors:
  - Various ISO: 9001:2000, 13485:2003, 14001:2004...
  - FDA: CFR820/QS, 21CFR-part11
  - GMP, TQM
  - Corporate Governance (example: SOX)
  - Internal and External Audits (Finance/Accounting, Quality, Corporate Governance)

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Regulatory - Governance → IQ

Corporate Governance internal controls:

### Objectives (CAVR):

- **Completeness**
- **Accuracy**
- **Validity**
- **Restricted Access (Security)**
- **Segregation of Duties...**

- **Preventative/Detective**
- **Automated/Manual**

### Fin Stmt Assertions:

- **Completeness**
- **Existence/Occurrence**
- **Validity/Accuracy**
- **Rights & Obligations**
- **Presentation/Disclosure...**

Note some overlap with IQ-related dimensions (Wang, et al, 1997).

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Objectives of this presentation

- Traditional Data Quality and Information Quality:  
Example DQ/IQ evolution with BI
- From DQ/IQ to “Enterprise IQ” Considerations
  - Organizational Context
    - Complexities
    - Business Strategy/Goals
    - Regulatory
  - Culture
- Summary

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Culture - Quality Maturity/Consistency

- Maturity level of Quality Culture influences where your company weighs:
  - **Balance** of Resources, Time, Cost, Risk, Urgency, ...
- A company undergoing frequent corporate combinations needs additional focus on cultural integration.
  - Quality Culture must prevail.
  - Especially for combinations with dissimilar maturities.
  - MNC: differing interpretations of same standards between countries.

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Culture - Quality Maturity/Consistency

- Overall Business Maturity and Strategy impacts EIQ:
  - Business Culture
  - Continual Improvement/Learning Organization
  - Measurements -> Accountability
  - Organizational Alignment with Quality
  - Education/Training
  - Individual understanding how roles contribute to the organization as a whole
  - Company-Unique Factors

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## QS/EIQ Positioning

Information Quality may be viewed as an essential element to organizational/operational success but...

Information Quality is a fundamental contributor to the larger goal of embracing business-wide Quality Systems and Principles.



© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Objectives of this presentation

- Traditional Data Quality and Information Quality: Example DQ/IQ evolution with BI
- From DQ/IQ to "Enterprise IQ" Considerations
  - Organizational Context
    - Complexities
    - Business Strategy/Goals
    - Regulatory
    - Culture
- Summary



© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Summary

### DQ/IQ:

- DQ/IQ can be reinforced in projects such as BI, via involvement of non-IT management.
- Re-organize around DQ/IQ and utilize stick and carrot.

### EIQ:

- DQ/IQ and EIQ Information Stakeholders are different.
  - It's about the customer, suppliers, shareholders, group management...
- Associate DQ/IQ/EIQ to the larger picture of Quality Systems.
  - Organizations in regulatory/governance environments by nature must internalize elemental concepts of DQ/IQ.
- There are complex contributors towards the end-goal of EIQ.
- IT's role is to be aligned to the business and contribute to progression towards strategic goals (value-add).

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## Thank You

### Questions?

Garry Darrer  
Getinge USA, Inc.  
[Garry.Darrer@GetingeUSA.com](mailto:Garry.Darrer@GetingeUSA.com)

© Getinge USA, Inc. 2008

GETINGE



The MIT 2008 Information Quality Industry Symposium



## **‘Fit For Use’ to a Fault**

Presented by

**Deborah Henderson, Tandum Lett,  
Anne Marie Smith and Cora Zeeman**

**July 2008**



The MIT 2008 Information Quality Industry Symposium



<b>Authors</b>	<b>Affiliations</b>
Deborah Henderson	CapGemini
Tandum Lett	Sullivan & Cromwell, LLP
Anne Marie Smith	EWSolutions
Cora Zeeman	University of Toronto





The MIT 2008 Information Quality Industry Symposium



## Presentation Objectives

- Introduce the concept of professional ethics in data management
- Inspect the behaviour of corporations in the drive for increased data quality and possible ethical implications
- Suggest a dimensional model of ethical challenges
- Present example ethical dilemmas
- Propose professional advocacy in the workplace around ethical handling of data.



The MIT 2008 Information Quality Industry Symposium



## Defining Ethics

- Compelled ethics - voluntary code, personal responsibility, signed code of ethics as a condition of professional affiliation
  - DAMA
  - ACM
- Imposed ethics – regulatory, cultural, legal framework
  - A comparison of approaches on privacy; Canada and the United States



The MIT 2008 Information Quality Industry Symposium



## US Privacy Law – Self Regulatory

US privacy law is self-regulatory

- Under the self regulatory regimes organizations design and implement their own privacy programs based on the criteria set down by the FTC :
- **Notice:** data collectors must disclose their information practices before collecting personal information from consumers;
- **Choice:** consumers must be given options with respect to whether and how personal information collected from them may be used for purposes beyond those for which the information was provided;
- **Access:** consumers should be able to view and contest the accuracy and completeness of data collected about them;
- **Security:** data collectors must take reasonable steps to assure that information collected from consumers is accurate and secure from unauthorized use; and
- **Enforcement:** the use of a reliable mechanism to impose sanctions for noncompliance with these fair information practices.

CZ



The MIT 2008 Information Quality Industry Symposium



## Canadian Privacy Law – the regulated approach

Canadian privacy law is a hybrid between a comprehensive regime of privacy protection with industry self regulation. **PIPEDA (personal information protection and electronic documents act)** covers all businesses who collect, use and disseminate personal information in the course of commercial activities and stipulates rules with exceptions that organizations must follow in the collection, use and dissemination thereof. The Act codifies the industry created privacy guidelines of the CSA Model Code for the Protection of Personal Information. The 10 guidelines set out therein are statutory obligations that all organizations that collect, use and disseminate personal information must follow.

- **Accountability:** An organization is responsible for personal information under its control and must designate an individual to be accountable for the organization's compliance with the principles;
- **Identifying Purposes:** An organization must identify the purposes for which personal information is collected at or before the time the information is collected;

CZ



The MIT 2008 Information Quality Industry Symposium



## Canadian Privacy Law – the regulated approach

- **Consent:** An organization must obtain the knowledge and consent of the individual for the collection, use or disclosure of personal information, except where inappropriate;
- **Limiting Collection, Use, Disclosure and Retention:** The collection of personal information must be limited to that which is necessary for the purposes identified by the organization. Information shall be collected by fair and lawful means. Personal information shall not be used or disclosed for purposes other than those for which it was collected, except with the consent of the individual or as required by law. Personal information shall be retained only as long as necessary for the fulfillment of those purposes;
- **Accuracy:** Personal information must be as accurate, complete, and up-to-date as is necessary for the purposes for which it is to be used;
- **Safeguards:** Personal information must be protected by security safeguards appropriate to the sensitivity of the information;

CZ



The MIT 2008 Information Quality Industry Symposium



## Canadian Privacy Law – the regulated approach

- **Openness:** An organization must make specific information about its policies and practices relating to the management of their personal information readily available to individuals;
- **Individual Access:** Upon request, an individual shall be informed of the existence, use and disclosure of his or her personal information and shall be given access to that information. An individual shall be able to challenge the accuracy and completeness of the information and have it amended as appropriate, and;
- **Challenging Compliance:** An individual shall be able to address a challenge concerning compliance with the above principles to the designated individual or individuals accountable for the organization's compliance.)

The federal privacy commissioner has the sole responsibility for handling privacy complaints against organizations. However, she fills an ombuds role wherein her decisions are recommendations and **not legally binding** and her decisions have **no precedential value, even within her own office.**

CZ



The MIT 2008 Information Quality Industry Symposium



## IT Compliance

Organizations often belong to privacy programs like **BBBonline** and **eTrust** which, along with providing organizations with privacy guidelines, monitor their compliance with the guidelines.

Complaints against any organization for violation of privacy can be brought to the FTC under their control over unfair and deceptive commercial practices.

CZ



The MIT 2008 Information Quality Industry Symposium



## Crossing the Line on Ethical Use

- Ubiquitous collection and use data
- Younger demographic don't see this as an issue
  - Facebook, YouTube, 15 minutes of fame
  - My privacy is not an issue ...as long as it doesn't affect me negatively
  - Understanding the issues

Ethical issues can only partially be monitored by automated means

CZ



The MIT 2008 Information Quality Industry Symposium



## Ethics and Data Management: an Analysis of Ethical Risk

Putting yourself in the hands of Corporate data and  
business intelligence analysts – what they do :

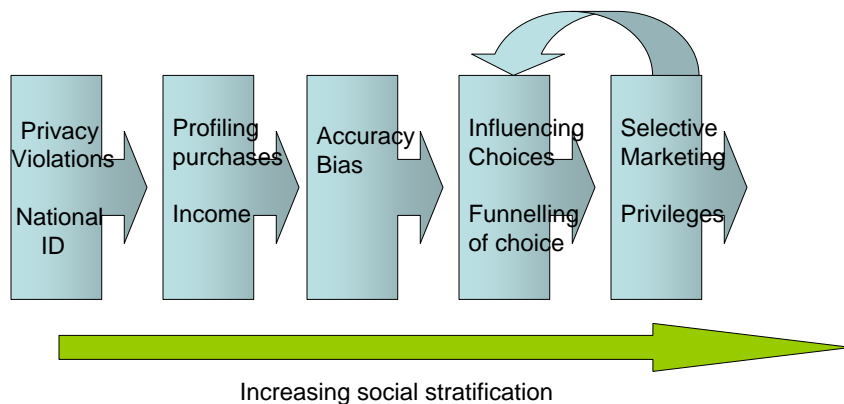
- WHO you are - Identity and theft, terrorist and criminal identification - **National Identity Schemes**
- WHAT you do – Profiling - Big Brother IS watching !
- WHEN you do it – Timing of data analysis, Bias, accuracy
- WHERE you do it – profiling and corralling your choices
- HOW are you treated - Credit scoring, preference tracking – ultimately privileged or not



The MIT 2008 Information Quality Industry Symposium



## Ethics and Data Management: an Analysis of Ethical Risk



**DO YOU WANT TO LIVE HERE?**



The MIT 2008 Information Quality Industry Symposium



## Who is Collecting Information on You?

- Government – census, taxes, any funded program you participate in
- Anywhere you go on-line – cookie, re-selling
- How you spend your money – RFID tagging connected to your credit card information?
- Employers and any organization you belong to

AS



The MIT 2008 Information Quality Industry Symposium



## Protection We Like

**We will give up our privacy for these types of concerns:**

- **Personal protection** - Credit card real time tracking of purchases protects us from fraud
- **Greater good** - Money laundering analysis by the Treasury (reporting of all large monetary transactions)
- **Greater good** – “No Fly” list - Protection from terror

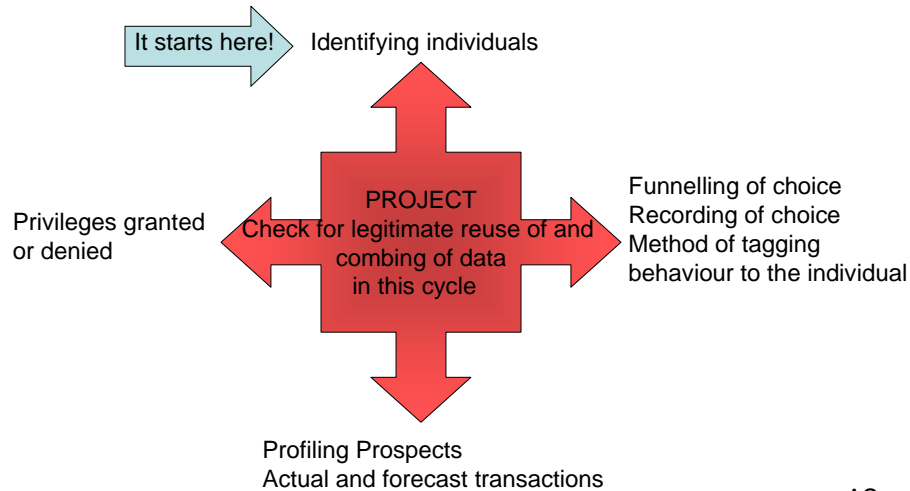
AS



The MIT 2008 Information Quality Industry Symposium



## Evaluating Ethical Risks – a Model



AS



The MIT 2008 Information Quality Industry Symposium



## Ethics and Data Management: some Examples of Ethical Risk

- Data Aggregation and Reconstitution
- Profiling - Categorizing into Strata
- Biased data collection and use
- Data Cleaning and standardization
- International movement of personal data
- 'Chinese wall' (non-porous) for information inside companies

AS



The MIT 2008 Information Quality Industry Symposium



## Categorizing into Strata

MarketMath in Canada has statistically analyzed the entire population into 52 psychographic segments.

- Gray-haired small town
- Executive burbs
- Francophone upwardly mobile

They have bought and combined datasets from the Census, private companies, liquor stores, pollsters

- combines with a geographic information system
- Sells this package for target marketing

***I get the offers for service, bargains, products  
YOU don't !***

AS



The MIT 2008 Information Quality Industry Symposium



## Bias Data Collection and Use

- Hunch and Search: Analyst has a hunch and wants to satisfy hunch; only uses data that satisfies their hunch
- Data Collection for Pre-defined result: Analyst is pressured to collect data and produce results based on pre-defined desires
- Biased use of data collected: Data is used to satisfy a chosen approach, data is manipulated for a chosen approach

AS





The MIT 2008 Information Quality Industry Symposium



## Data Cleaning & Standardization

- 30% of us move households every year
- Billions of transactions for analysis
- Organizations that are data cleaning – to what standards? And selling the latest and greatest information on you for profit
- Who assesses accuracy?

TQM Cop-out we have all heard:

***“We have no separate quality control, the quality is baked-in to the process”***



The MIT 2008 Information Quality Industry Symposium



## Data Cleaning & Standardization in Identity – We give up privacy for security – Right ?

### Steps Involved

- Registration
  - Biometric sample taken, stored and compared
  - ID token issued, based on existing records
- Data-matching and profiling, quality
  - Ongoing, behind the scenes
  - Data gathering + database linkages
- Authentication (at control points)
  - Identity match between body and ID token
  - Database checks (personal data, watch list)
  - Request denied or approved

AS



The MIT 2008 Information Quality Industry Symposium



## Securely and Reliably Identify Everyone? False Sense of Security

Everyone with a 'clean' record passes

- Most 9/11 attackers had NO record of suspicion
- Terrorist training manual: "fit in" as "normal"
- Terrorists can repeatedly test screening system, then only need to pass once!

***"The positive identification of individuals does not equate to trustworthiness or lack of criminal intent."***  
(emphasis in original)

(Ben Shneiderman, USACM testimony at the Congressional Hearings on National Identification Card Systems, Nov 2001)

AS



The MIT 2008 Information Quality Industry Symposium



## Movement of Data, internationally

- Privacy laws differ worldwide
- Corporations cannot move employee personal data (identification) across borders without reviewing laws of sending and receiving countries.
- May need to set expectations of 'single sign-on' in large multi-nationals



The MIT 2008 Information Quality Industry Symposium



## **‘Chinese Walls’ in Corporations**

- Legal firms may represent both the plaintive and the defendant
- IT policies need to be implemented and monitored to protect the privacy of both parties.

Recombination of information in the external business world may embarrass the firm

TL

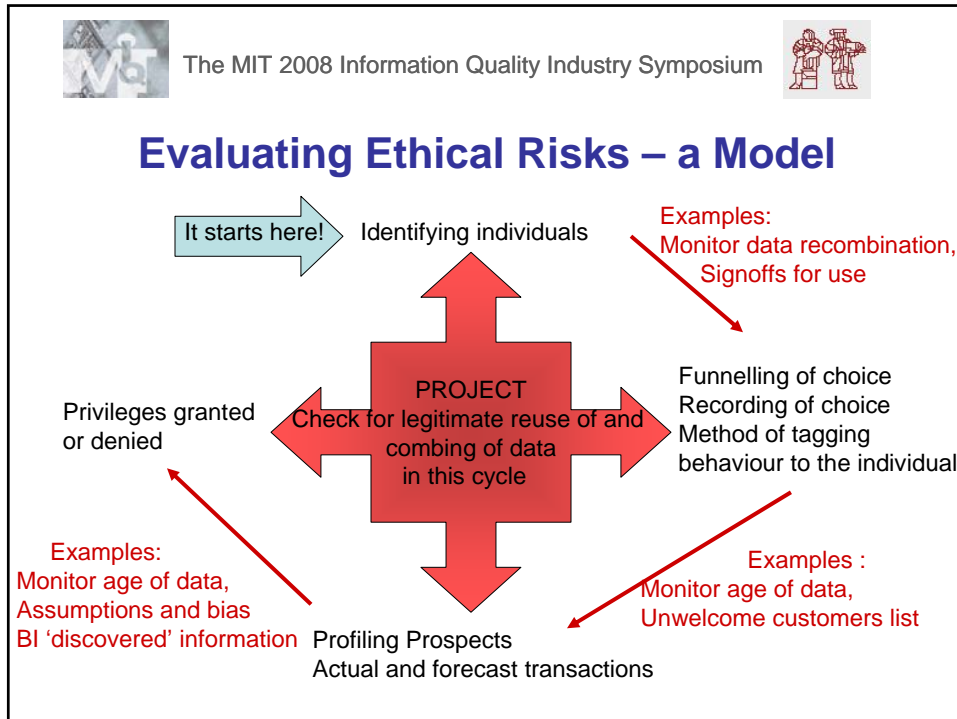


The MIT 2008 Information Quality Industry Symposium



## **Call to Action**

- Business is not aware of ‘where the data comes from’ and ethical issues may not be obvious to them
- Automated monitoring is not sufficient protection
- Cultural norms and ethics in the workplace influence corporate behavior
- Taking a Professional stand
- Acting ~! Evaluate the risk/benefit
  - Take and “index” of Ethical Risk based on our model



The MIT 2008 Information Quality Industry Symposium

## Conclusion - What do we\* do now?

**\* We = Professionals/Associations + citizens**

- Convene and participate in public data ethics forums
  - DAMA, MITIQ!.. Get more going!
- Resist emphasis on overly costly, unreliable, narrowly technological approaches
  - What are the purposes? Would it be effective?
  - Who is being served? Disadvantaged?
  - What are the alternatives?
- Demand social and political accountability

TL



The MIT 2008 Information Quality Industry Symposium



## Questions



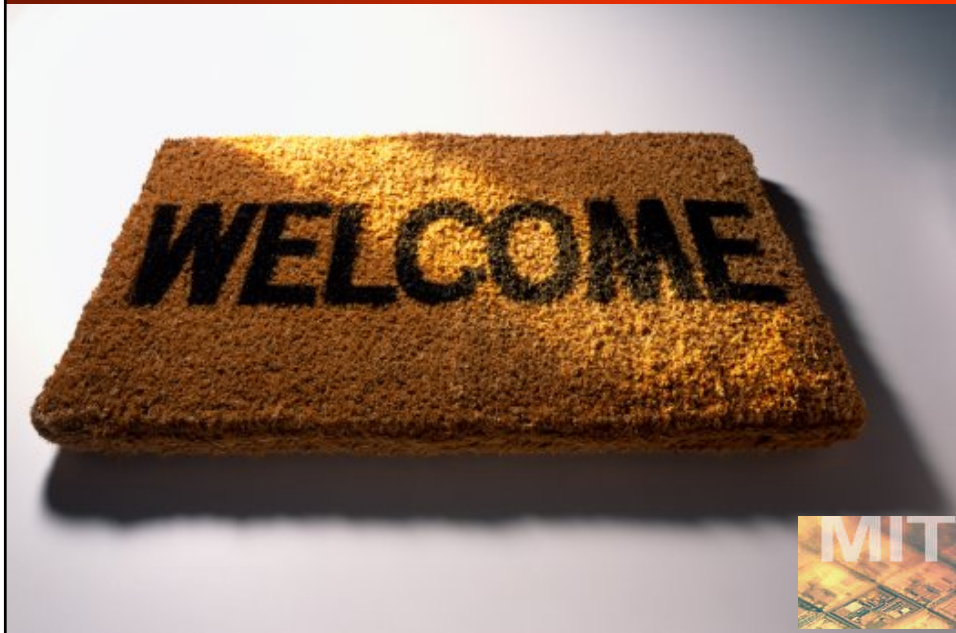
The MIT 2008 Information Quality Industry Symposium



## Contacts

- Deborah Henderson:  
[deborah.henderson@inergi.ca](mailto:deborah.henderson@inergi.ca)
- Tamdum Lett:  
[LettT@sullcrom.com](mailto:LettT@sullcrom.com)
- Anne Marie Smith:  
[AMSmith@ewsolutions.com](mailto:AMSmith@ewsolutions.com)
- Cora Zeeman:  
[kitytje@hotmail.com](mailto:kitytje@hotmail.com)

*Linda Kresl, BI Manager, Mentor Graphics  
MIT IQ Symposium, July 17-19, 2008 Boston, MA.*



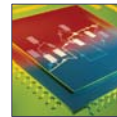
### *Quality Information as a Service for SOA*

- 1. How to relate the Service Component to your business entities.***
- 2. How to build a high quality data service that enables all other services.***
- 3. How to develop the Information Management Strategy that prevents wasting your SOA investment.***

## Who are we?



- Fastest growing of the top tier EDA companies
- Market leadership in key segments
  - ESL
  - Functional Verification
  - Analog-Mixed Signal
  - Design for Manufacturing
  - Integrated Systems Design
- Leader in Verification Standards
- Worldwide support, training and consulting



**Mentor  
Graphics**

## Mentor Graphics

Wilsonville, Oregon U.S.A. Headquarters

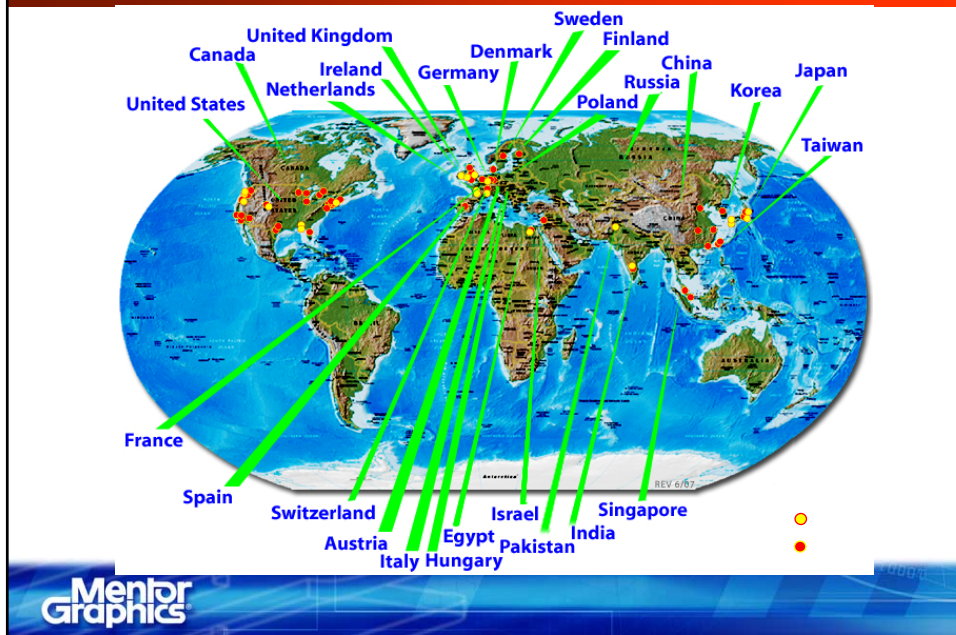


- 300,000 Square Feet of Office & Laboratory Space
- 4,350 Employees Worldwide
  - 1,000 at Wilsonville, Oregon Headquarters

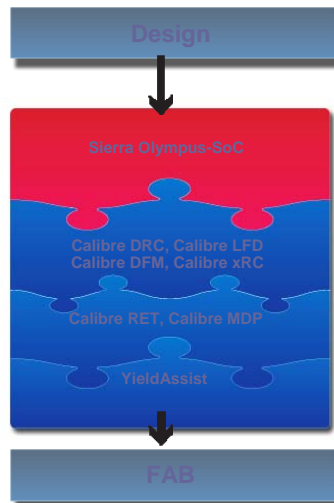
**Mentor  
Graphics**



## Mentor Graphics around the World



## Mentor Offers Integrated Design-to-Fab



- Manufacturing-driven closure  
P&R architecture is the  
discontinuity of the 65/45/32/22  
nm era
- Mentor Graphics now delivers an  
integrated design-to-fab solution  
for 65/45 nm and beyond



## Mentor's ESL Architectural Tools

### System Architect™

- Assess performance and power of TLM blocks
- Validate hardware against real software execution
- Quickly analyze architecture alternatives

### Vista™

- Easily debug SystemC transaction-level models
- IDE-style project creation and advanced debugging
- Intuitive transaction sequence viewer

### Visual Elite™ SD

- Design & assemble mixed SystemC / HDL models
- Intuitive SystemC / HDL text and graphical design
- Create TLM virtual prototypes for software development

A complete hardware design suite for architectural exploration and optimization

Mentor  
Graphics

## Data is the Foundation.....

*In the Gartner Report: Key Issues for Data Management and Integration, 2006 Ted Friedman writes:*

*What impact will enterprise information management have on approaches such as service oriented architecture? New approaches to architecture and implementation of applications, such as SOA, are creating pressure to increase the availability, timeliness of delivery, consistency and auditability of data. **Without a strong focus on data at the foundation of their initiatives, organizations will fail in capturing the benefits of speed and agility they seek from SOA.** Our research will analyze the key dependency points between EIM and SOA to expose the risks of failing to align the application and data points of view.*

Mentor  
Graphics

## *What is Service Oriented Architecture?*

- ✓ ***SOA is a method of conceptualizing, designing & building applications by assembling reusable building blocks, each of which is usually represented as a service.***

Mentor  
Graphics

## *Service Components*

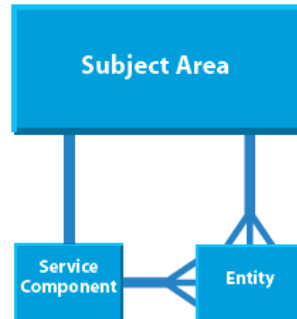
- ✓ ***Service Component (and interface)***
- ✓ ***main entities in the SOA Concept***
- ✓ ***specify them uniquely***



Mentor  
Graphics

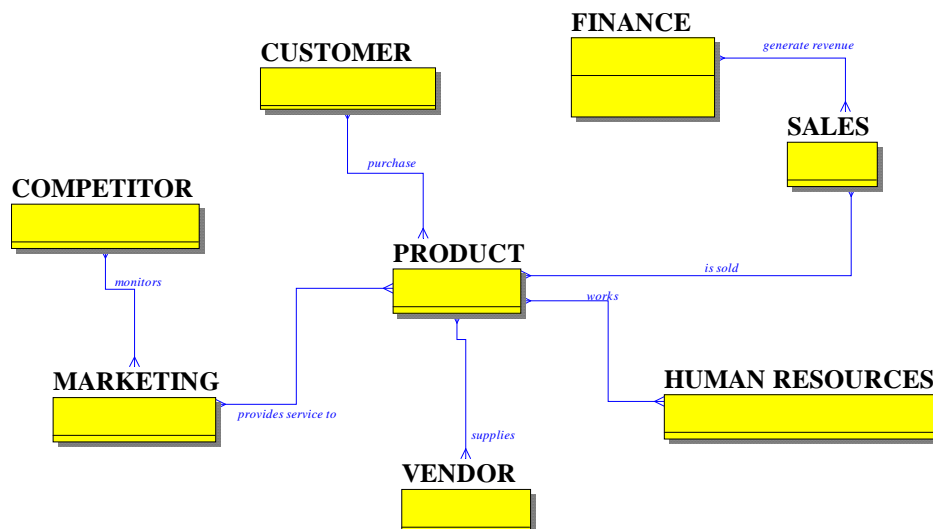
## Entities

- ✓ ***A person, place, or thing that the business cares enough about to store and track information.***

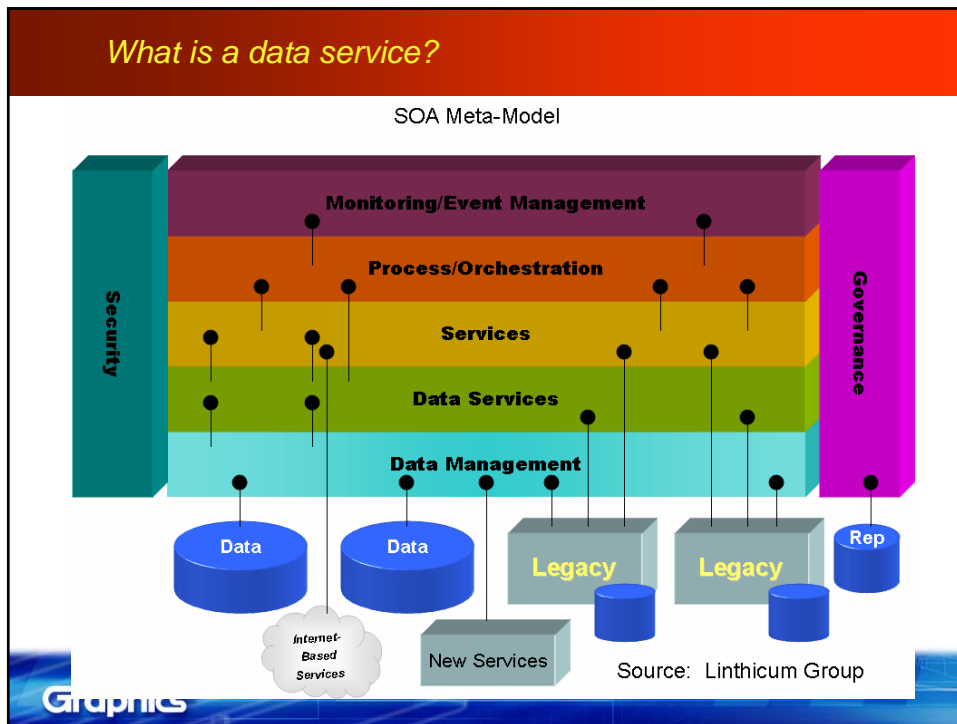


Mentor  
Graphics

## Data Model is a communication vehicle for the business



Mentor  
Graphics

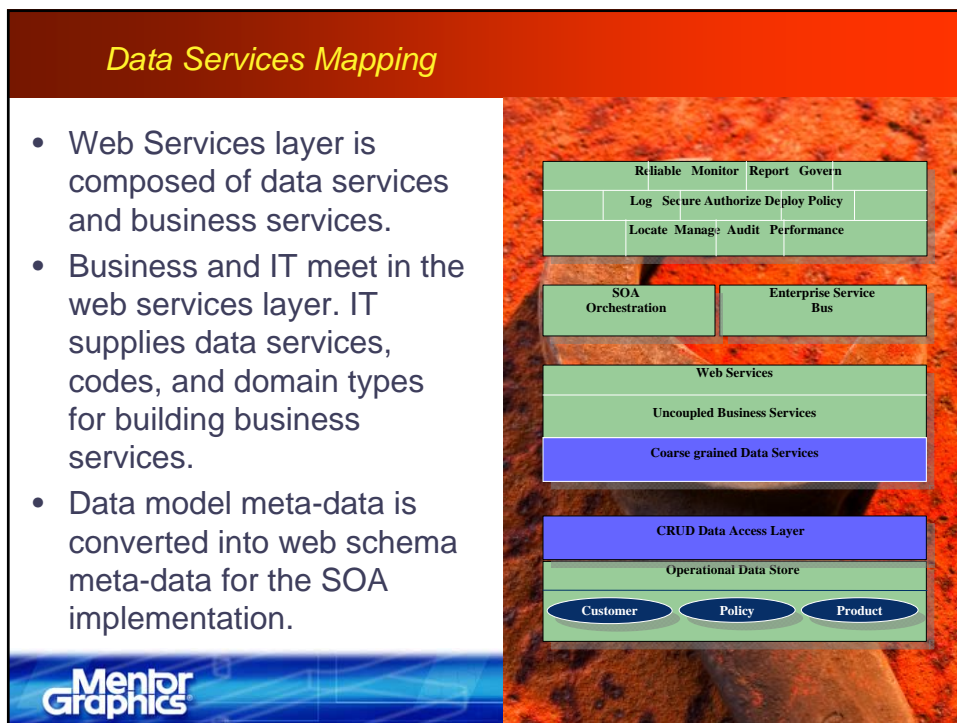
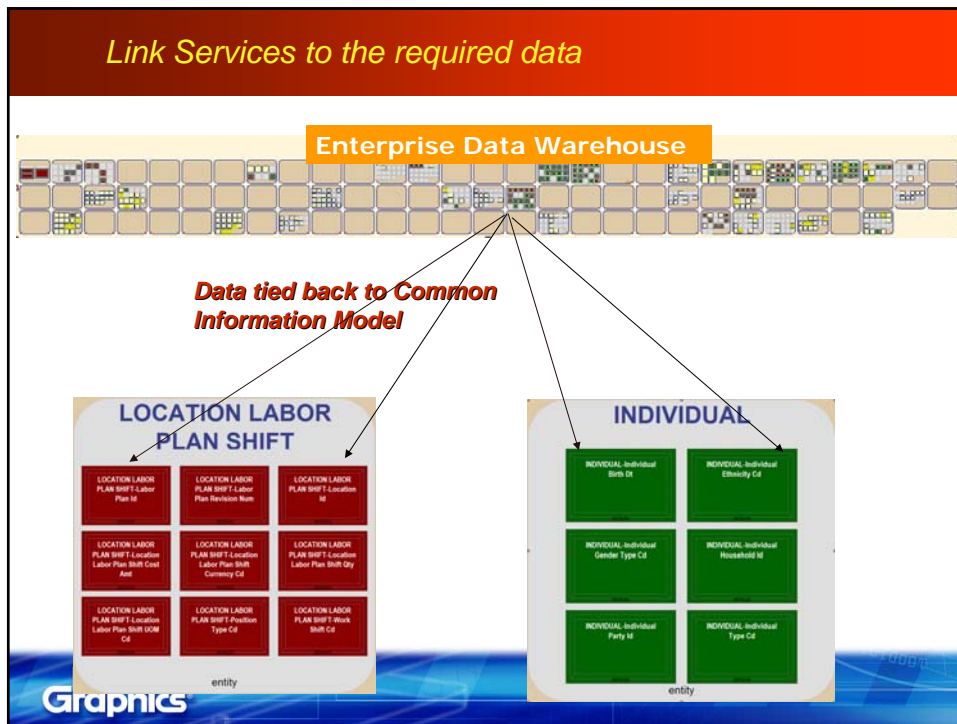


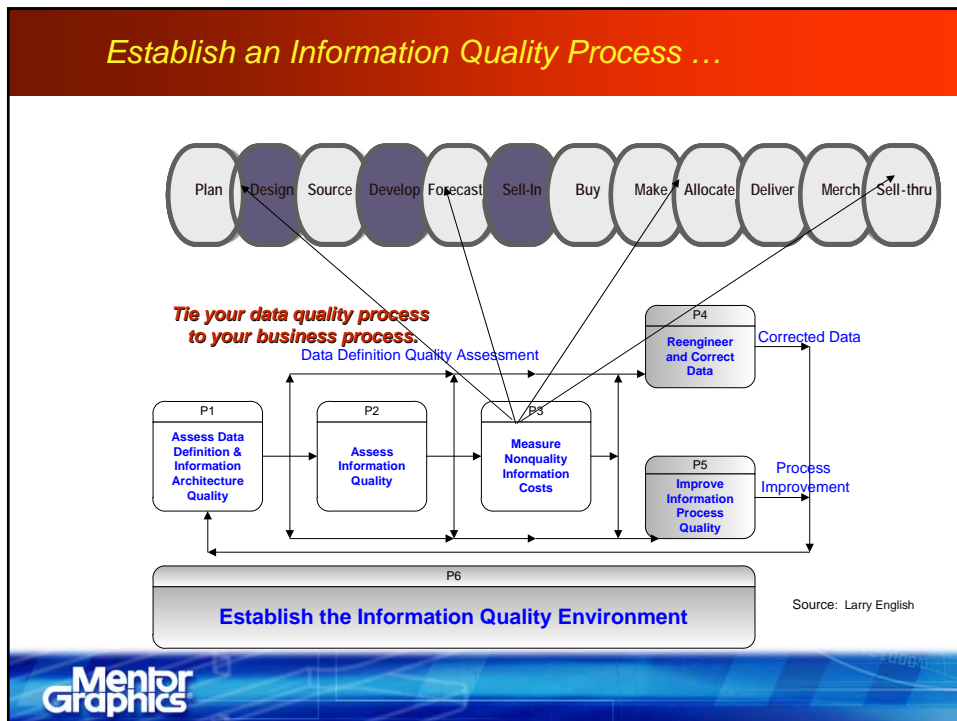
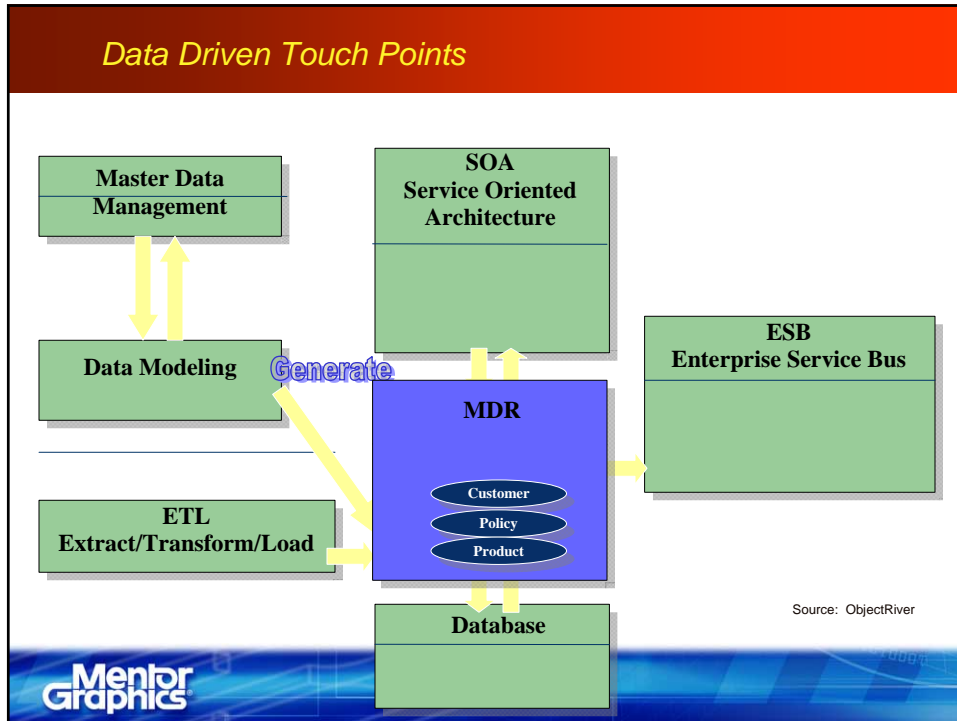
### Data Services . . . a necessity for SOA

- The advantage of a data service is the ability to bind many different data types into one unified **enterprise-wide data** model, including schema and content.
- Manages poorly designed and normalized **data** without having to force fixes to the back-end systems.

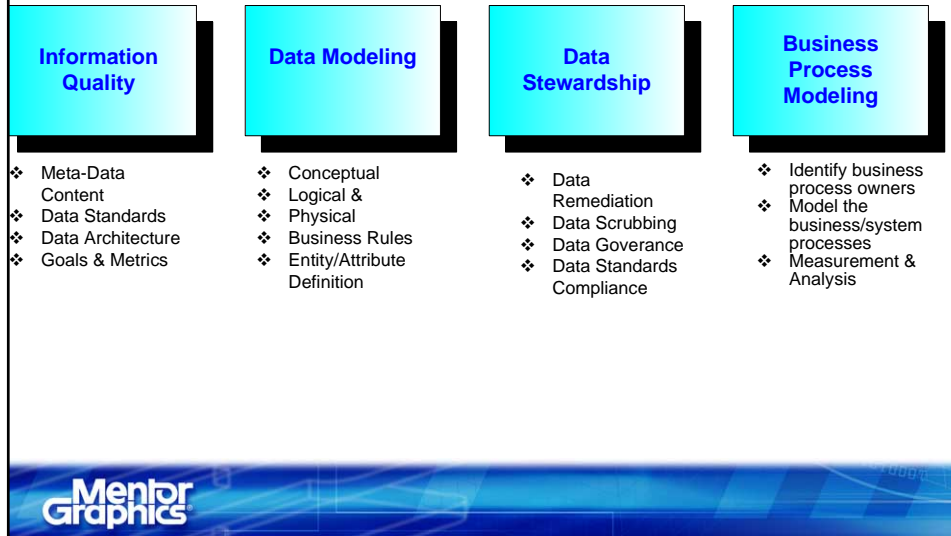
The slide features a list of two bullet points on the left, describing the advantages of data services in SOA. On the right, there is a photograph of a long, straight asphalt road with white dashed lines, receding into the distance under a clear blue sky. The Mentor Graphics logo is visible in the bottom left corner.

Mentor Graphics

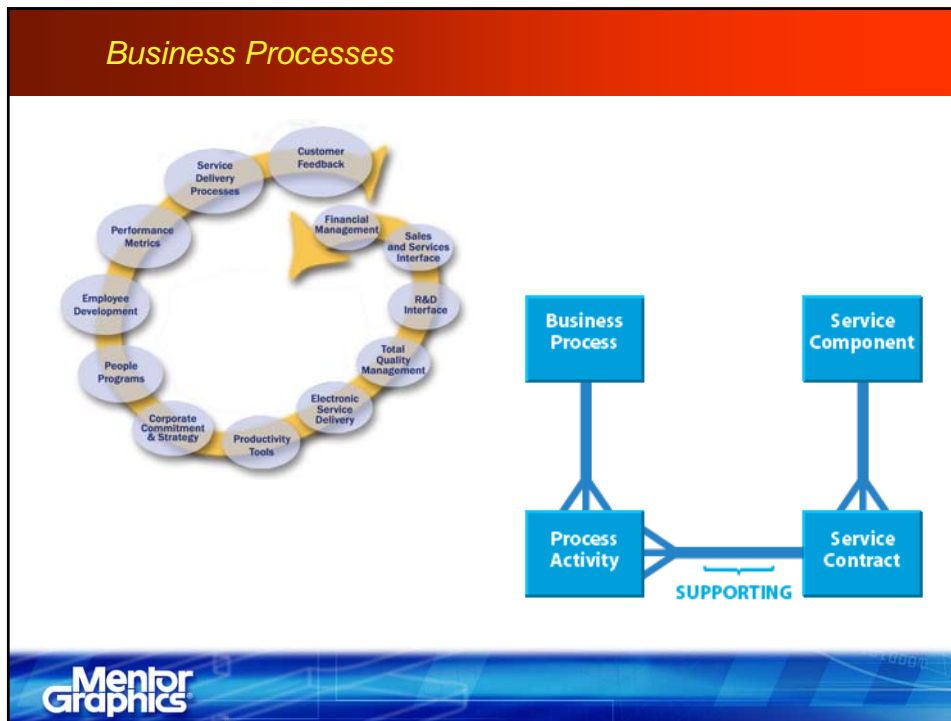




**Information Quality Management will focus on the following activities:**



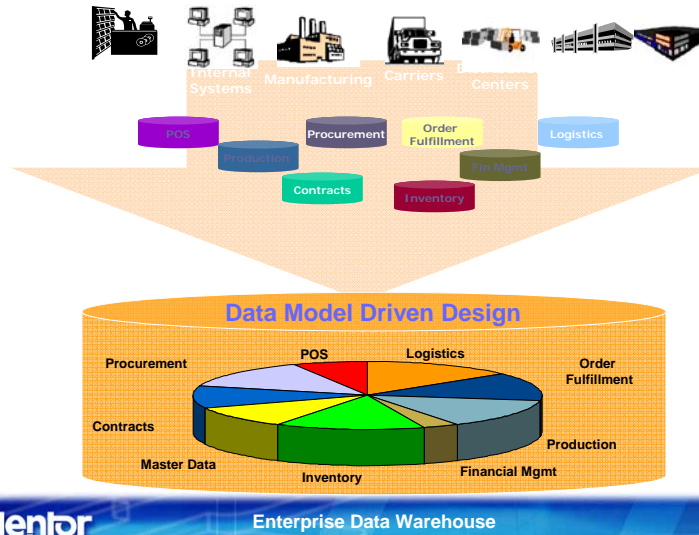
**Business Processes**





## MEASURES OF QUALITY

### Single Version of the Truth



### Tools that can help you improve your process . . .

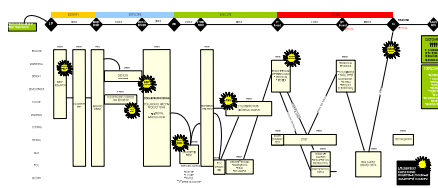


## Value Stream Mapping

Technique for capturing processes

"as is" (current state)

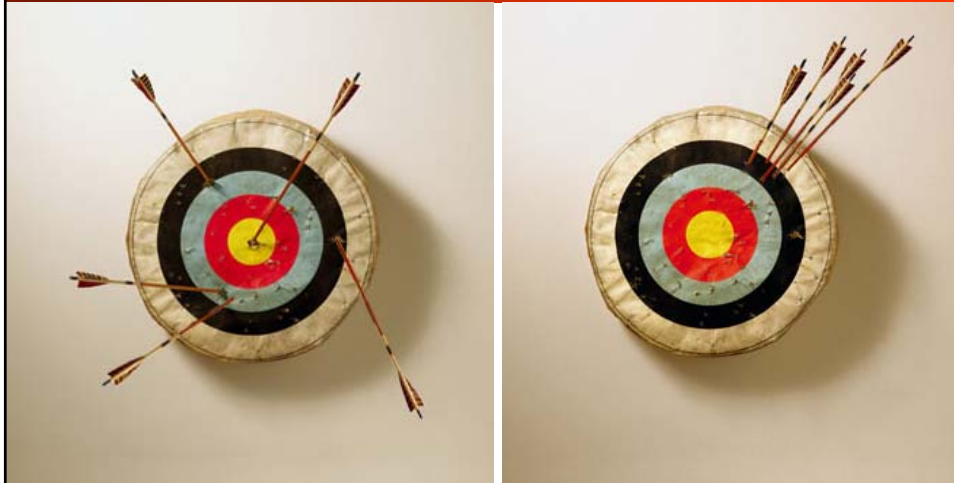
"to be" (future state)



Mentor  
Graphics



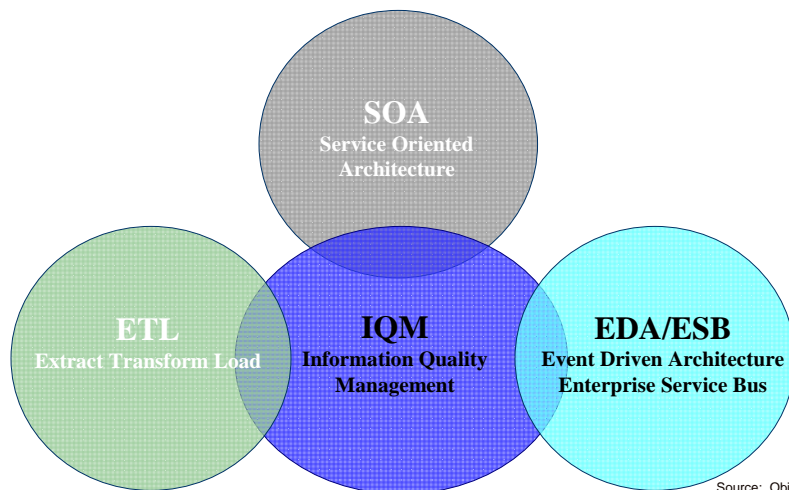
### *Stabilize before you Improve*



Which player did better in this round?  
Which is likely to do better after several rounds of play?

**Mentor  
Graphics**

### *Coupling Principles*

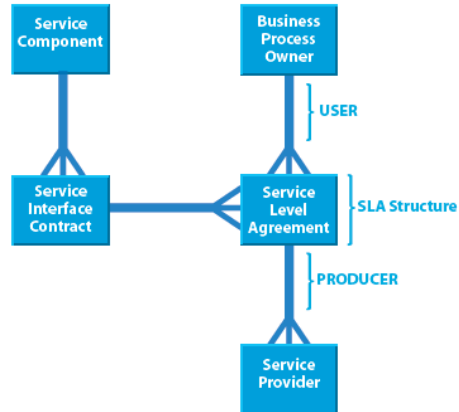


Source: ObjectRiver

**Mentor  
Graphics**

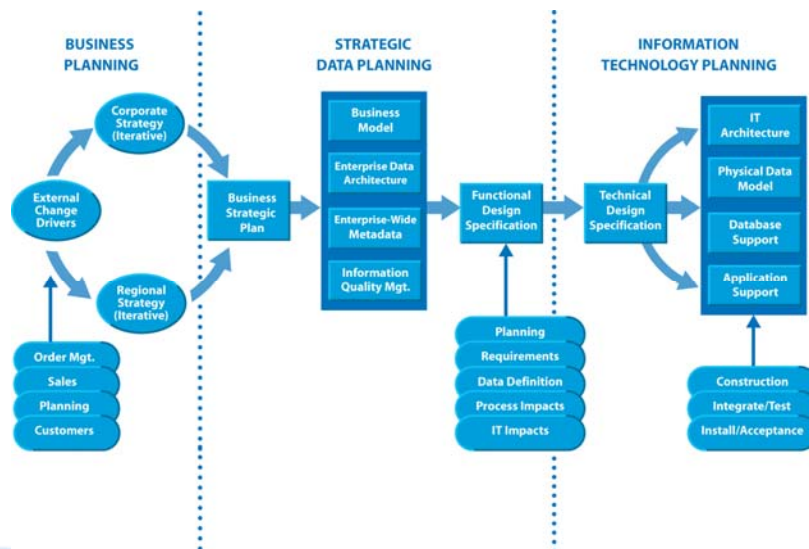
## Service Level Agreement

✓ A SLA is a formal negotiated agreement between two parties.

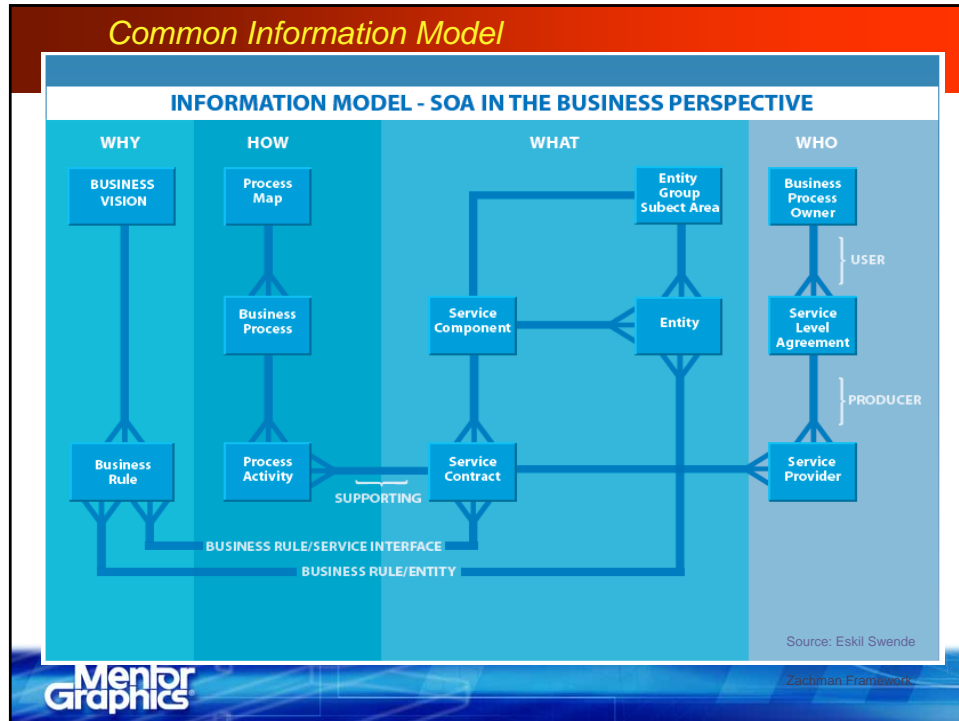


Mentor  
Graphics

## Information Quality Management Strategy drives SOA



Mentor  
Graphics



**Master the Art of Problem Solving**

- **Develop** a hypothesis and test it.
- **Check** results and adjust your plan.
- Be relentless in pausing to learn from “what just happened?”

Mentimeter

Mentimeter

## *What now . . . ?*

Taking it back to the  
job . . .



**Mentor  
Graphics**

## *Quality Information as a Service for SOA - Questions?*



**Mentor  
Graphics**



The MIT 2008 Information Quality Industry Symposium



## Unified Architecture for Integrating Intelligence Data

Suzanne Yoakum-Stover, Ph.D.

Potomac Institute for Policy Studies, Senior Research Fellow  
US Army CERDEC I2WD, Information Exploitation Futures Lab, Lead Scientist

Tatiana Malyuta, Ph.D.

New York City College of Technology, Associate Professor  
US Army CERDEC I2WD Information Exploitation Futures Lab, Knowledge Manager

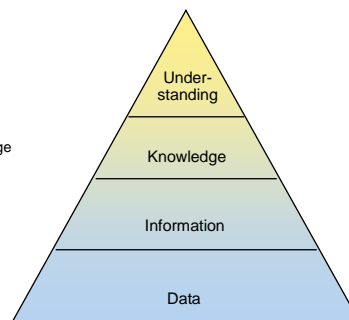


The MIT 2008 Information Quality Industry Symposium



## Problem Context and Statement

- Business of Intelligence
  - To develop and communicate understanding
- Intel Business Processes
  - Move Intel artifacts with respect to the cognitive hierarchy
    - **Into:** Data collection
    - **Up:** Semantic enhancement & fusion → Information & knowledge
    - **Out:** Communication and collaboration → understanding
- Data Integration Problem
  - Integrate all Intel into a coherent repository of knowledge
  - In an Ultra-Large-Scale systems environment<sup>1</sup>
    - Decentralized
    - Inherently conflicting, diverse, and unknowable requirements
    - Heterogeneous, changing, and inconsistent elements
    - Normal failures, continuous operation, evolution, and deployment
    - Immense scale along many dimensions
  - Without attempting to control
    - Data sources, types, data-models
    - Processing, usage, application



Cognitive hierarchy

- Data = symbols lacking explicit semantics
- Information = data + semantics
- Knowledge = information + logic
- Understanding = knowledge + human insight

1. Northrop, L., et al., *Ultra-Large-Scale Systems The Software Challenge of the Future*, Pittsburgh: Carnegie Mellon University, 2007.  
<http://www.sei.cmu.edu/publications/books/engineering/uls.html>



The MIT 2008 Information Quality Industry Symposium



## Current Practice Fails

Merging or harmonizing data models, either physically or virtually, fails to accommodate the demands of the fluid and rapidly growing intelligence enterprise

- Physical integration of disparate models into a single canonical data-model is untenable in the face of scale and complexity and cannot adapt as the system evolves.
- Virtual integration lacks authority over data sources and fails to support inter-source collaboration without introducing yet another database.

What begins as a neat solution for a handful of systems quickly becomes intractable with scale. This phenomenon is but one early symptom of our evolution toward Ultra-Large Scale (ULS) systems and as such, invites a completely different approach - one that remains viable in a freely evolving, interdependent collective of systems, people, policies, cultures, and economics, very little of which will ever be under our control.

3



The MIT 2008 Information Quality Industry Symposium



## New Approach

- Our approach to integrating Intelligence data in a ULS systems environment is data-centric (as opposed to data-model – centric) and proceeds in two stages
  - The first addresses the unified storage of the entire spectrum of intelligence artifacts regardless of modality or representation.
  - The second stage builds upon the foundation provided by the first to address the unified storage of structured data to enable semantic data integration.
- The result is a layered data architecture that can accommodate any kind of data without placing restrictions on vocabulary, structure, semantics, or constraints, in a way that addresses the needs of the Intelligence Community today while providing a seamless transition path toward a future of ULS systems imbued with semantic technologies.

4





The MIT 2008 Information Quality Industry Symposium



## Design Tenets

- Layer 1 of our data integration architecture supports an aspect of collection and rudimentary exploitation. Layer 2 supports the processing by which data is enhanced with semantics to produce information, and the processing by which information is enhanced with richer associations to produce knowledge.
- We embrace the diversity of domain-specific data-models employed throughout the Intelligence Community by taking a data-model agnostic approach wherein the integration model makes the least possible commitment to any particular data-model.
- The character and meaning of the source data-model, when existent, is preserved and made accessible by the data store.

5



The MIT 2008 Information Quality Industry Symposium



## Layer 1: Indigenous Artifacts

- In Layer 1 we seek to integrate the entire spectrum of indigenous artifacts by collecting them in one (possibly distributed) database using standard means for physical and or virtual data integration.
- Crucial principles
  - Avoid making any data or data-model transformations in the process of data ingestion
  - Make the least possible commitment to a data-model in the target storage schema

Consequently, the Layer 1 database schema is quite simple and flat, exposing a minimal set of essential meta-data fields whose main purpose is to support back-tracking to the original artifact and or source.

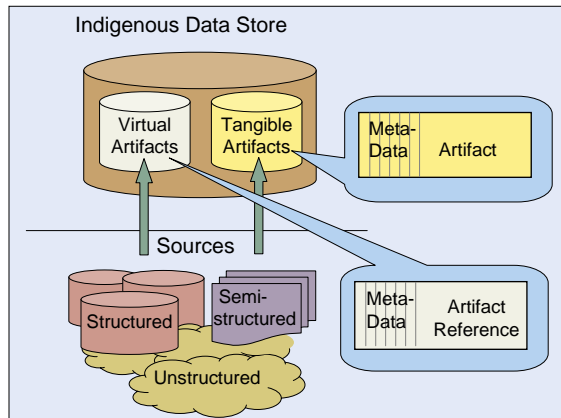
6



The MIT 2008 Information Quality Industry Symposium



## Layer 1: Universal Indigenous Store



### The benefits of this most trivial form of integration

- Provides a manageable yet powerful and standard interface to the source data
- Gives us the option to either “lazily” load and cache data as “virtual artifacts” for performance sake, or persist and control data as “tangible artifacts” for the long term
- Provides “one stop shopping” access to the indigenous data for analysts
- Establishes a foundation upon which deep data integration can be more effectively pursued

7



The MIT 2008 Information Quality Industry Symposium



## Layer 2: Universal Store for Structured Data

The challenge--a universal storage model for structured data

- To accommodate structured data in a way that *exposes* that structure for use, without *imposing* the structure on the data store itself
- Determine a method for storing and managing any kind of structured data, reflecting any data-model, so that it can be shared, efficiently exploited, and extended in unforeseen ways without requiring model-specific storage implementations

8





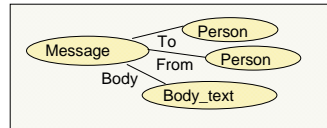
The MIT 2008 Information Quality Industry Symposium



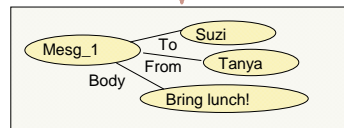
## The Problem with Structured Data

07/04/07  
To Suzi,  
Bring lunch!  
From Tanya

(a) Unstructured Data



(b) Data-model



(c) Structured Data

The data-model is imposed on  
the database

and

the data is frozen into it

Message	To	From	Body
Msg_1	Suzi	Tanya	Bring lunch!
...			

(d) Typical database structure

9



The MIT 2008 Information Quality Industry Symposium



## Layer 2: Data Model Abstraction

A domain-neutral storage model for structured data

- Decoupling that which varies, namely vocabularies and, more generally the data-models, from that which remains constant, namely the source artifact, and ideally the storage structure
- Considering structure, vocabulary, semantics, and constraints from a higher level of abstraction from which we then distill a minimal set of elements sufficient to capture any data-model

10



The MIT 2008 Information Quality Industry Symposium



## Layer 2: Elements

- **Mention:** A chunk of data, either physically located within a tangible artifact, or contained within an analyst's mind
- **Concept:** An abstract idea, defined explicitly or implicitly by a source data-model
- **Predicate:** An abstract idea used to express a relationship between "things"
- **Term:** A disambiguated *mention* abstracted from the source artifact or asserting analyst
- **Statement:** Encodes a binary relationship between a subject and an object mediated by a *predicate*

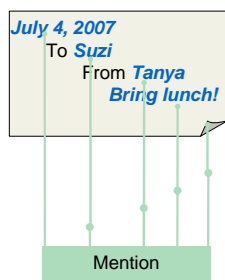
11



The MIT 2008 Information Quality Industry Symposium



## Layer 2: Data



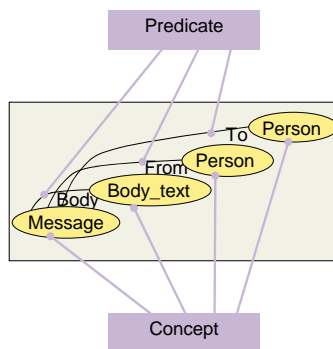
12



The MIT 2008 Information Quality Industry Symposium



## Layer 2: Data Model



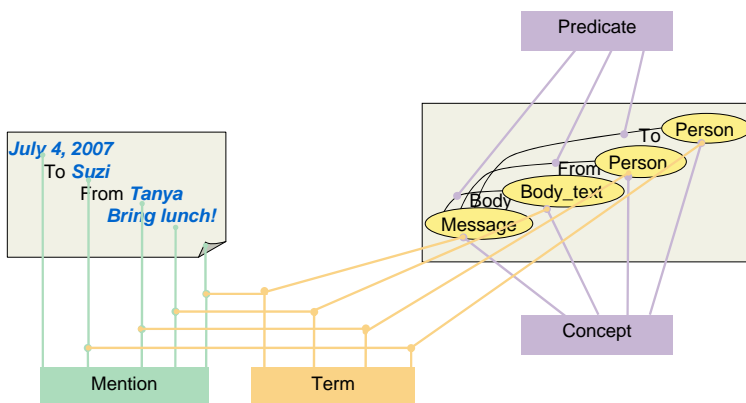
13



The MIT 2008 Information Quality Industry Symposium



## Layer 2: Semantics



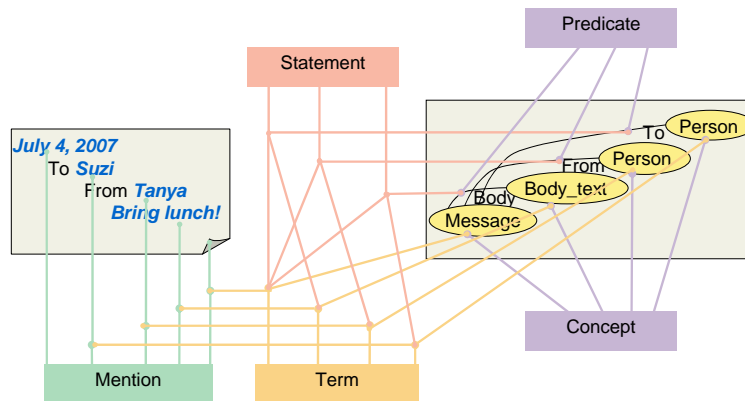
14



The MIT 2008 Information Quality Industry Symposium



## Layer 2: Semantic Associations



15



The MIT 2008 Information Quality Industry Symposium



## Data Description Framework (DDF)

The Layer 2 elementary constructs (concept, predicate, mention, term, and statement) provide the fixed-points of a data reference model that will ultimately serve as a practical data integration platform. We call this reference model the Data Description Framework (DDF).

Despite its simplicity, the DDF is a rich model that can be viewed from at least two different perspectives as a synergistic combination of two higher order models lying along different dimensions of abstraction

- Extrospective
  - Concept and predicate look outward toward domain knowledge.
  - Mention looks outward toward the data.
- Introspective
  - Term and statement form a semantic model and abstract data-model internals to expose structure in a uniform way.

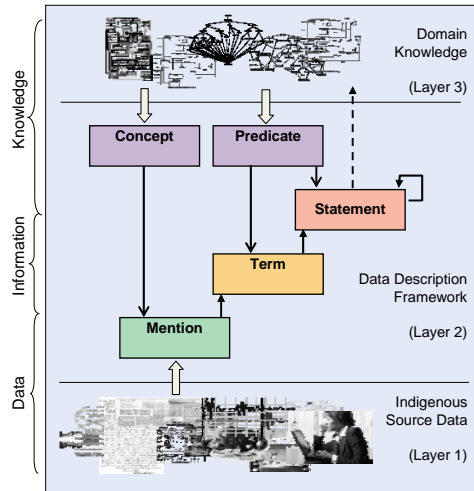
16



The MIT 2008 Information Quality Industry Symposium



## DDF: Vertical and Horizontal Integration



Together the introspective and extrospective models enable both horizontal and vertical data integration

- The extrospective abstraction bridges data and domain knowledge (vertical integration).
- The introspective abstraction bridges data structured by various disparate processes (horizontal integration) and binds the two outward looking faces of the extrospective model to provide a comprehensive data integration model.

17

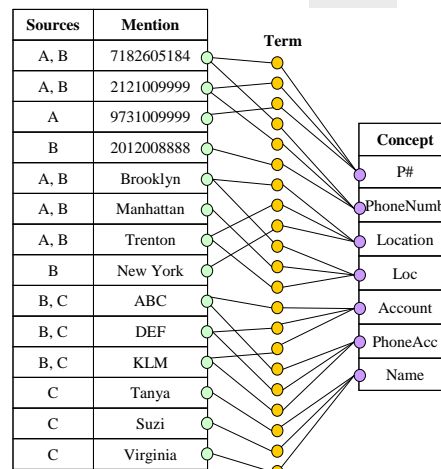


The MIT 2008 Information Quality Industry Symposium



## DDF: Simply Put

- Useful integration results just from putting data in the DDF
- Mostly automatic process
  - Data of interest selected from external data stores
  - Automatic load into DDF
  - No data-model harmonization
  - No information is lost
- Queries on Terms
  - What is 7182605184?
  - What sources mention 7182605184?
  - What of the Locations mentioned in DB-A are also mentioned in as Locs in DB-B?



DB-A	
P#	Loc
7182605184	Brooklyn
2121009999	Manhattan
9731009999	Trenton

DB-B		
PhoneNumh	Account	Location
7182605184	ABC	Brooklyn
2121009999	DEF	New York
2012008888	KLM	Trenton

DB-C	
PhoneAcc	Name
ABC	Tanya
DEF	Suzi
KLM	Virginia

18



The MIT 2008 Information Quality Industry Symposium



## DDF: Stating the Obvious

Term	Mention	Concept
T1	7182605184	P#
T2	2121009999	P#
T3	9731009999	P#
T4	7182605184	PhoneNumb
T5	2121009999	PhoneNumb
T7	2012008888	PhoneNumb
T8	Brooklyn	Loc
T9	Manhattan	Loc
T10	Trenton	Loc
T11	Brooklyn	Location
T12	New York	Location
T13	Trenton	Location
T14	ABC	Account
T15	DEF	Account
T16	KLM	Account
T17	ABC	PhoneAcc
T18	DEF	PhoneAcc
T19	KLM	PhoneAcc
T20	Tanya	Name
T21	Suzi	Name
T22	Virginia	Name

Statement

- Relations in source data automatically become statements
  - Only small sample illustrated
  - No data-model harmonization required
  - No information is lost
- Queries on Statements
  - Capability equivalent to that of the source system
  - Examples
    - What terms, concepts, or mentions are associated via the predicate hasName?
    - What phoneAccs hasName Tanya?

Predicate
hasLocation
hasAccount
hasName

19



The MIT 2008 Information Quality Industry Symposium



## DDF: Data Integration

Term	Mention	Concept
T1	7182605184	P#
T2	2121009999	P#
T3	9731009999	P#
T4	7182605184	PhoneNumb
T5	2121009999	PhoneNumb
T6	2012008888	PhoneNumb
T7	Brooklyn	Loc
T8	Manhattan	Loc
T9	Trenton	Loc
T10	Brooklyn	Location
T11	New York	Location
T12	Trenton	Location
T13	ABC	Account
T14	DEF	Account
T15	KLM	Account
T16	ABC	PhoneAcc
T17	DEF	PhoneAcc
T18	KLM	PhoneAcc
T19	Tanya	Name
T20	Suzi	Name
T21	Virginia	Name

Statement

- Nontrivial data integration by
  - Adding predicates
  - Creating statements that span across sources
- Enables
  - Correlation across data sources
  - Knowledge enhancement
  - More sophisticated queries
    - What are the PhoneAccs of those who work with Tanya?
    - What other labels does New York have?

Predicate
hasLocation
hasAccount
hasName
isSameAs
worksWith

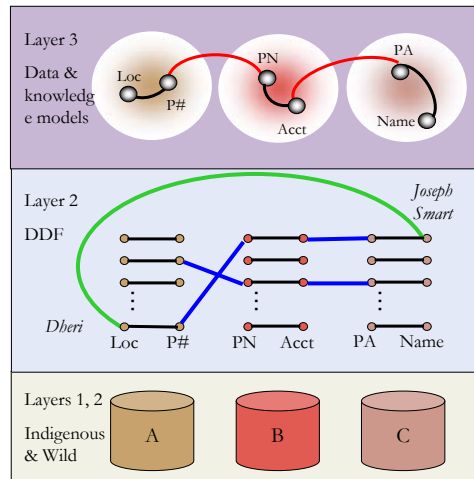
20



The MIT 2008 Information Quality Industry Symposium



## Above and Beyond (Layer 3)



### Connecting the Dots

- Halos represent distinct source systems.
- Associations
  1. Black: Automatic from ingestion into Layer 2
  2. Red: Added in Layer 3 to harmonize data-model elements
  3. Blue: Indicate data match, due to 2
  4. Green: Automatic result of 1-3
- Data in B used to generate new association between data in A and C (Green).

21



The MIT 2008 Information Quality Industry Symposium



## Conclusion

- We have presented the first two layers of a multi-layer data integration architecture that enables deep semantic data integration in a ULS systems environment.
- The underlying model, the DDF, supports both horizontal and vertical data integration (i.e. across disparate data-models and from data to knowledge) by embracing the diversity of data / knowledge models and processes by which data is structured.
- More importantly, the model admits a practical implementation ( "hard running code") that accommodates artifacts of any modality (e.g. text, audio, images, video, signals) in a single unified data store that enables true multi-intelligence data fusion and the continuous enrichment of data into knowledge.

22



The MIT 2008 Information Quality Industry Symposium



## Application of Practical Nominalism to Data Management

"Everyone is entitled to his own opinion, but not his own facts." Senator Daniel Patrick Moynihan

Fulton Wilcox  
Colts Neck Solutions LLC



The MIT 2008 Information Quality Industry Symposium



### Abstract

#### Application of Practical Nominalism to Data Management

Many information quality problems have as a root cause an over-reliance on the ontological notion that "entities" are "real" while events and transactions are merely transitory manifestations of "real" entities in action. The nominalist position is that an "entity" such as the "Massachusetts Institute of Technology" are not "real," but merely a name tagging a flow of transactions and events, and what the entity "is" by definition differs from day to day. Nominalism has been used to explain the "King Canute" impediments to creating taxonomies and ontologies : e.g., ..just as the taxonomy is defined, more events and transactions flood in to put it into disarray, but there is a more positive perspective.

Our capability to improve data quality will benefit if we exploit the growing power of our technology to run systems processes directly against transaction data and event data, and as a corollary, minimize reliance on "synthesized" data. Synthesized data looks "real" and may even look like an event, but it in fact has been synthesized by the application of rules and conventions to genuine transaction and event data. For example, a reported number of MIT employees is inherently "synthetic" data, because it fuses "realist" notions of what constitutes "MIT," what constitutes employment, what very detailed rules apply (hours per week) to transactions, and many others.

The evolution of technology favors this nominalist approach, because of increased processing capacity, the creation of SOA (service oriented architecture), rules engines and network services. The nominalist design approach also is liberating in that informational "gold" – our raw data – will not be held captive in synthetic outputs. It also greatly assists in supporting privacy, security and due process, because it becomes far easier to isolate data.





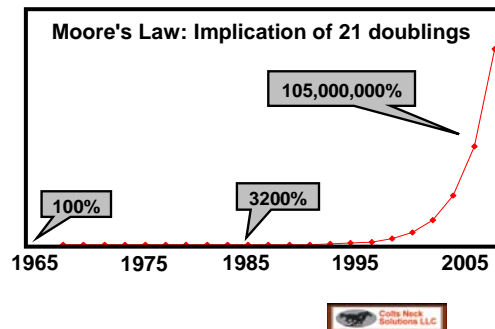


The MIT 2008 Information Quality Industry Symposium



## Technological Progress: The 900 Pound Gorilla

- Opens up new opportunities, e.g. Internet 2.0 and Internet 3.0
- Moore's Law is a proxy for brute force increases in capability across IT: processing, memory, storage, networking, virtualization ...
- Plotting it on a linear scale dramatizes effects
- The plot also prompts questions:
  - were IT users 150,000,000% better off in 2007 than in 1965?
  - Indeed, were they 800% better off in 2007 than in 2001? If not, why not?
  - With more doublings coming, how do we put the gorilla to work?
- What are the data management and data quality implications?



The MIT 2008 Information Quality Industry Symposium



## Must the data warehouse inmates take over the asylum?

- **ETL (extract, transform and load) processes rub people's noses into the shortcomings of today's data and data management processes**
  - What looks sensible and "good enough" within a given venue breaks down when used across venues
- **As described by Claudia Imhoff of Intelligent Solutions, those implementing data warehouses and ETL identified three needs:**
  - improve the quality of the data being integrated,
  - create sets of integrated master or reference data (MDM),
  - implement repositories of current data for management and operational purposes (ODS), and so on...
  - so the data warehouse team took a much broader, more diverse set of projects. "
- **Her reaction to this expansion of ETL/DW scope was "Back off!" ... None of these is a data warehouse project.**
  - From a scope creep perspective, she is right, but the problem needs to be addressed, and rescue is probably not coming from BPEL, etc.





The MIT 2008 Information Quality Industry Symposium



## Two Worldviews Regarding Data

### Conceptually-Based

- models and rules
- "realist" concepts
- abstractions
- mystical appliqués

Experience shows that divergence between the system/database conceptual world view and reality stimulates error and omission

### Event-oriented and observational



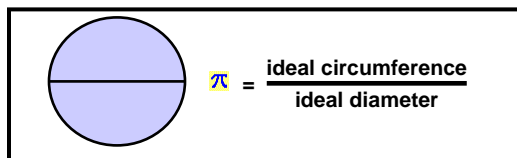
who, what, when, where and why, in empirical, unsmoothed detail



The MIT 2008 Information Quality Industry Symposium



## Abstractions or "universals" are disconnected from reality



↑ "Realist" perfection

↓ "Nominalist" particularity

### cosmological constraints

There is no such thing as a line, because  
... there are no real-world "points"  
... no such thing as a diameter line  
... No such thing as a "plane"  
... therefore, no such thing as a circle

### execution variances – e.g.

Engine components are not ideal cylinders  
Journal-To-Bearing Clearance  
0.001-0.003 Service Limit 0.006  
Journal Diameter = 1.7713-1.7720  
Out-of-Round Limit 0.0005 in. Max.

<http://www.merkurencyclopedia.com/Motor/enginespecs.html>





The MIT 2008 Information Quality Industry Symposium



The entity "MIT" exhibits "inbetweener" anomalies:  
e.g. it educates students for Harvard degrees

Harvard University



Massachusetts Institute of Technology

MIT-Whitacre College  
55% enrolled seek Harvard Degrees  
45% seek MIT degrees

We live in an age of organizational and role fuzziness: joint ventures, mergers stacked on mergers, systems consolidation, etc.



The MIT 2008 Information Quality Industry Symposium



## Synthetic Data: The Data "Bucket" Problem

- **Synthetic data is created by applying rules to source data to populate data "buckets."**
  - "Total revenue in June" may look like real data, but is a "bucket" of synthetic data
- **As synthetic data moves laterally and up the food chain, buckets are stacked on buckets –**
  - e.g., "revenue in June" is aligned with "cost of goods" in June" and summed up to corporate level.
  - You and I may be entitled to differing definitions of "June" (e.g., calendar versus fiscal), but those differing definitions will create collision between our synthetic "facts"
- **We often are unable to determine whether synthetic data fits purposes**
  - the rules are not expressed as rules, but as "data"





The MIT 2008 Information Quality Industry Symposium



## Rule Conflict: what is valid "there" may not be valid "here;" a rule that was appropriate yesterday may not fit today

- **Rules are essential and valuable, but one rule set does not fit all**
  - We need to accommodate multiple sets of rules, perhaps differing by place, perhaps over time, etc.
- **Also, we need to extricate rules from data and data from rules**
  - For example, move rules to a rules engine
  - Apply as needed the rules to source data to create synthetic data "fit for intended use"
- **Minimize use of synthetic data of unknown or inapplicable provenance**

Rules from a Model Pertaining to A Country other than the U.S.

"No employee may be older than 65 years."

"If a person is male he can't have a husband and if he has a wife it must be female. If a person is female, she can't have a wife and if she has a husband it must be male."



The MIT 2008 Information Quality Industry Symposium



## Summary: "Realist" Modeling Problems

- **Unboundedness:** "domain" is an arbitrary, subjective subset
- **Never-ending:** Who has time to map the grit of reality to abstract structures in the sky?
- **Unbridgeable disconnects:** Will my set of abstraction bridge map to your set of abstractions?
- **Rippling changes:** change creates unaddressed versioning issues
- **Inertia:** Having built it, users and data are force-fitted into a "solution," dampening feedback





The MIT 2008 Information Quality Industry Symposium



## Revisiting choices between "Realism" and "Nominalism"

- **Realism and its cousin, conceptualism, treat abstractions as "real"**
  - The "model" or the ideal of a given object are thought of as the highest truth, while real-world "instantiation" is viewed as annoyingly noisy
  - Real-world instances are idiosyncratic and, like snowflakes, no two are identical
- **Within narrowly bounded, disciplined abstractions are useful, while divergences from reality may be of minor importance**
- **However, as systems reach and data reuse overruns "stovepiped" domains**
  - Model design and detail cannot keep pace with expanding "footprints" of automation
  - Use of given models over extended time lead to accumulation of error
  - Tighter coupling of models to software development removes a buffering effect
- **To expand reach and increase the reuse of data we need to accommodate and exploit particularity rather than mask it**



The MIT 2008 Information Quality Industry Symposium



## Heraclitus: "We never step twice into the same river"

- To the nominalist, experience is flow
- We may name flowing water "river" or "stream" or "brook," but Heraclitus's point was that we are labeling a flow
  - Particular molecules flow by, to be replaced by other molecules
  - Words like "river" or "stream" are labels for fuzzy sets of instances
  - If it rains hard, the stream becomes a river; in drought the river becomes a riverbed
- An image of a river flowing over a waterfall is a good introduction to "transactions" and "all is flow"





The MIT 2008 Information Quality Industry Symposium



## "Practical" Nominalist Data Management Worldview

- "Truth" consists of the flow of transactions and events
- These center on action verbs, not "state" verbs (variants of "to be")
- Nouns do not represent static "entities," but dynamic balances of transactions and events
- "State" is merely how things were left by prior transactions
- Today's IT capabilities can capture and manipulate the "waterfall" of transactions
  - buy more processing, disk, network capacity and address space



Colin Neek Solutions LLC



The MIT 2008 Information Quality Industry Symposium



## Managing Transaction and Event Data as Transactions and Events

- **A transaction or granular event can be expressed as a declarative sentence**
  - The ancient, robust way of capturing events is in the declarative sentence
  - Of course, not all declarative sentences convey transactions
- **A transaction sentence can be repackaged as an XML document**
  - The transaction payload comprises the declarative sentence
  - The labels and tags provide the context-defining metadata
- **What is also critical is to emphasize the "action" verb**
  - All "properties" link back to transaction action verbs – bought, built, born, etc.
  - All "relationships" are a consequence of action verbs
  - Mere "properties" (e.g., "is an MIT grad student") may be synthetic overlays of "rules" and actions (e.g., paid, but never showed may = "is," showed but never paid may = "is not", failed to pay parking fines may = "never again")

Colin Neek Solutions LLC

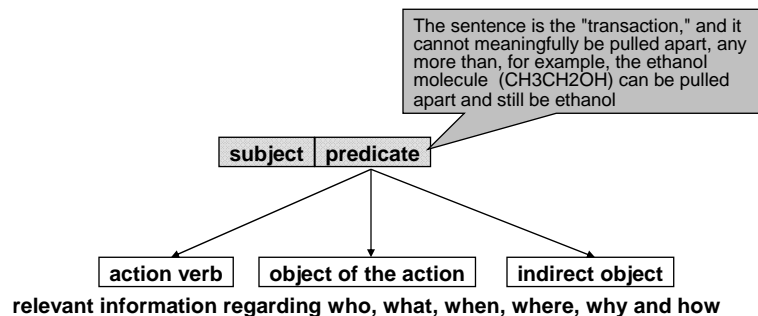




The MIT 2008 Information Quality Industry Symposium



## Transaction/event expressed as declarative sentence



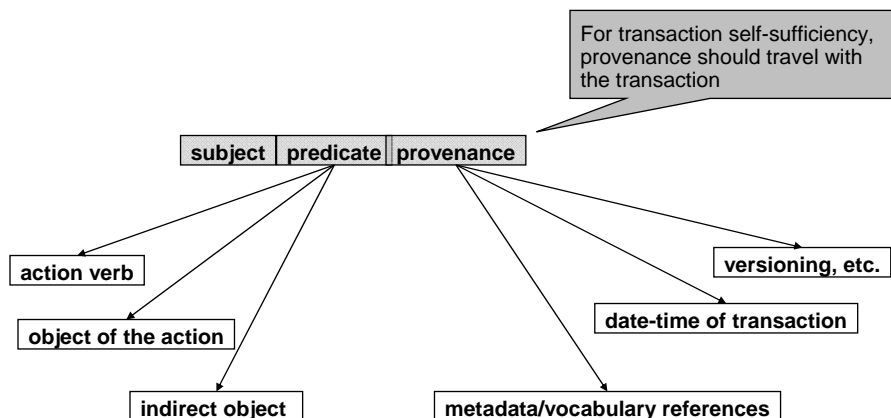
"... a phenomenon seen in almost all biomedical terminologies [is] the expression via single terms of information which should more properly be conveyed in the form of complete sentences." [http://ontology.buffalo.edu/medo/Onto\\_Epist.pdf](http://ontology.buffalo.edu/medo/Onto_Epist.pdf)



The MIT 2008 Information Quality Industry Symposium



## Transaction/event





The MIT 2008 Information Quality Industry Symposium



## RDF (Resource Description Framework) Triples

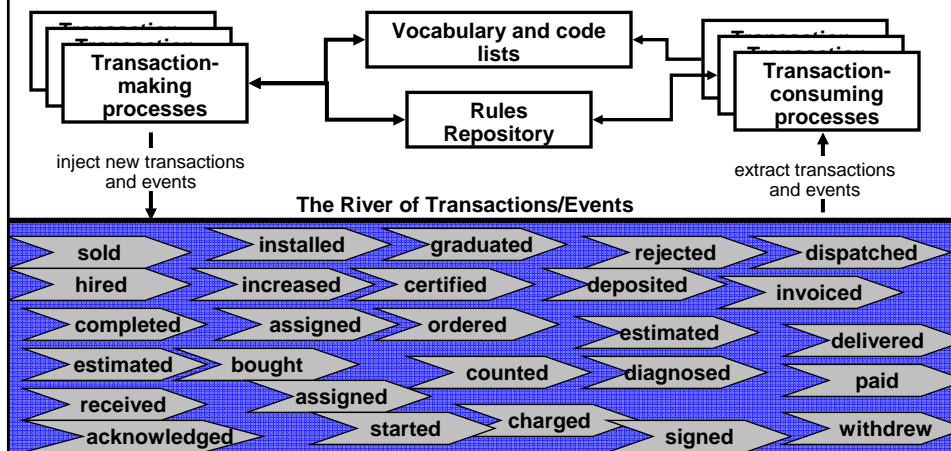
- **RDF triples also focus on "declarative sentence" subject-verb-predicate constructs**
  - As typically described, "triples" are not focused on preserving "events"
  - Triples instead are akin to "assignment statements," in which the predicate updates the subject's "properties"
- **Triples are often instantiated using state rather than action verbs**
  - For example, Susan "has" a PhD degree as opposed to "on June 14<sup>th</sup>, 2007" MIT awarded Susan a PhD degree
  - Therefore, the motivation behind a "triple" is more "realistic" rather than nominalist in nature
- **Given triples relationship to the declarative sentence, triples are open to nominalist application, given a nominalist mindset**



The MIT 2008 Information Quality Industry Symposium



## Experience: A "River" of Transactions and Events



Conventionally, transaction "fish" are shredded into "data bases"





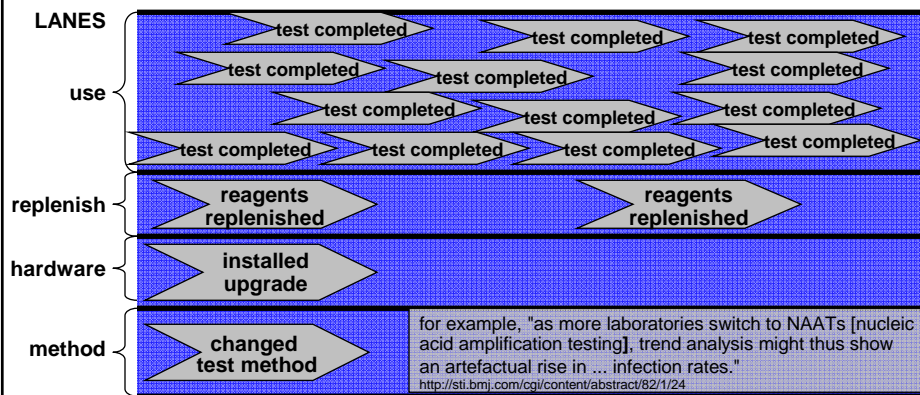


The MIT 2008 Information Quality Industry Symposium



## Flow of experience "swim lanes:" e.g. a healthcare example

For medical test result evaluation, it often is necessary to access maintenance, calibration, apparatus reconfiguration and change of method events



Slower-flowing "lanes" = "master data"



The MIT 2008 Information Quality Industry Symposium



## Three Nominalist Success "Cases":

A Few Minor Adjustments in Architecture





The MIT 2008 Information Quality Industry Symposium



### Case 1: Billing System as "Transaction/Event Generator"

- **Designed a "nominalist" billing system with a event level database**
  - Each billable "event" transformed into an informationally self-sufficient, priced out, discounted, tax-rated, fully tagged invoice transaction
  - Business process involved a complex, fast-evolving mix of services and products
  - Customers could be multi-level or for pricing purposes related by affinity group
- **Generated invoices as a data warehouse "query"**
  - Merely selected transactions to be billed via a query
  - Few billing-time lookups or pricing calculations needed, because the billed transactions were already priced, discounted, etc.
- **Credits and other reversal events were symmetrically aligned with the original billable event**
- **Resulting data was highly "portable" and auditable**
  - Users could access or copy relevant transaction detail
  - Little need to access rules and reference tables (e.g., contract pricing) because the relevant data was embedded in the transaction



The MIT 2008 Information Quality Industry Symposium



### Case 2: Accounting and Project Accounting System

- **Transactions were maintained as detailed, informationally self sufficient row at the lowest level of informational granularity**
  - e.g., if a transaction spanned multiple projects, general ledger accounts/sub accounts, or organizations, the detailed splits were created
- **Design was of great help in coping with rapid changes in project, G/L and organizational structure**
  - e.g., because of an internal organizational restructuring, there was a need to run simultaneously under two, radically different general ledgers
  - with GL codes being merely "tags" in each transaction, reporting in two different accounting contexts was both easy and highly auditable
  - no synthetic data "buckets" existed because transactions were held at the lowest feasible level
- **In many respects, Cases 1 & 2 were architected as data warehouses even though they were the actual billing and accounting systems**
  - In effect, the "data warehouse" paradigm "took over" the vertical application





The MIT 2008 Information Quality Industry Symposium



### Case 3: eBusiness transaction processes

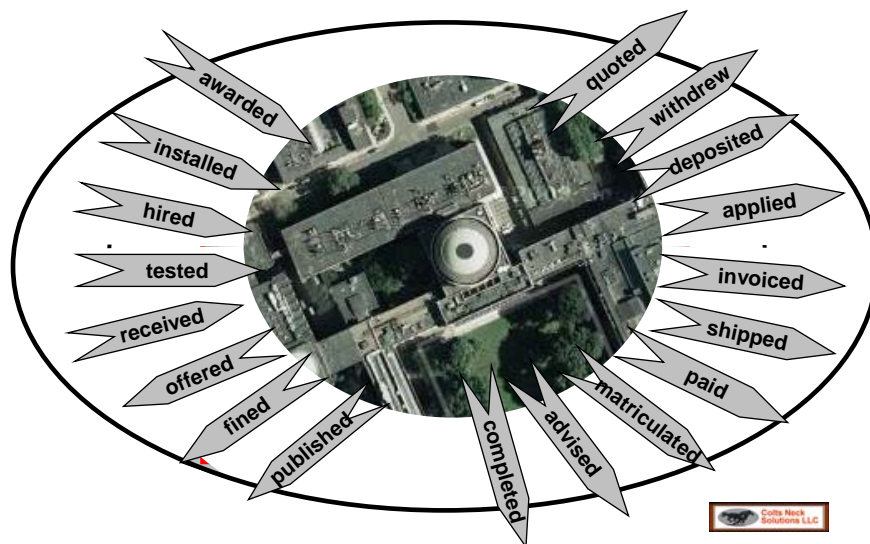
- **B2B eBusiness transaction flow is inherently "nominalistic"**
  - eBusiness transactions typically are not "native" to any of the parties
  - Each transaction has to "speak for itself" because recipients have little or no access to the sender's systems and data
- **B2B transactions comprise an event-driven "flow"**
  - The data passed is aligned with the eBusiness transaction's action "verb," (such as "ordered", "shipped", "invoiced") etc.
  - Although transactions are standardized, the flow often is not because not all the transactions are used or are used in non-standard sequences
- **An organization's eBusiness transaction "corona" becomes a potentially important source of "nominalistic" transaction data**
  - "You are what you eat:" an entity is defined by its inputs and outputs
  - However, today most eBusiness transactions are treated as perishable, initiated for a specific need and unceremoniously unpackaged on arrival at the "far end"



The MIT 2008 Information Quality Industry Symposium



### MIT's nominalist electronic transaction "corona"





The MIT 2008 Information Quality Industry Symposium



## eBusiness transactions constitute "declarative sentences"

On January 31, Customer party X ordered products a and b from supplier party Y

```

<xsd:element ref="cbc:IssueDate" minOccurs="1" maxOccurs="1">
  <xsd:annotation>
    <xsd:documentation>
      <ccts:Component>
        <ccts:ComponentType>BBIE</ccts:ComponentType>
        <ccts:DictionaryEntryName>Order. Issue Date. Date</ccts:DictionaryEntryName>
        <ccts:Definition>The date assigned by the Buyer on which the Order was
        issued.</ccts:Definition>
        <ccts:Cardinality>1</ccts:Cardinality>
        <ccts:ObjectClass>Order</ccts:ObjectClass>
        <ccts:PropertyTerm>Issue Date</ccts:PropertyTerm>
        <ccts:RepresentationTerm>Date</ccts:RepresentationTerm>
        <ccts:DataType>Date. Type</ccts:DataType>
        <ccts:AlternativeBusinessTerms>OrderDate</ccts:AlternativeBusinessTerms>
      </ccts:Component>
    </xsd:documentation>
  </xsd:annotation>
</xsd:element>
  
```

**OASIS UBL 2.0 order fragment**

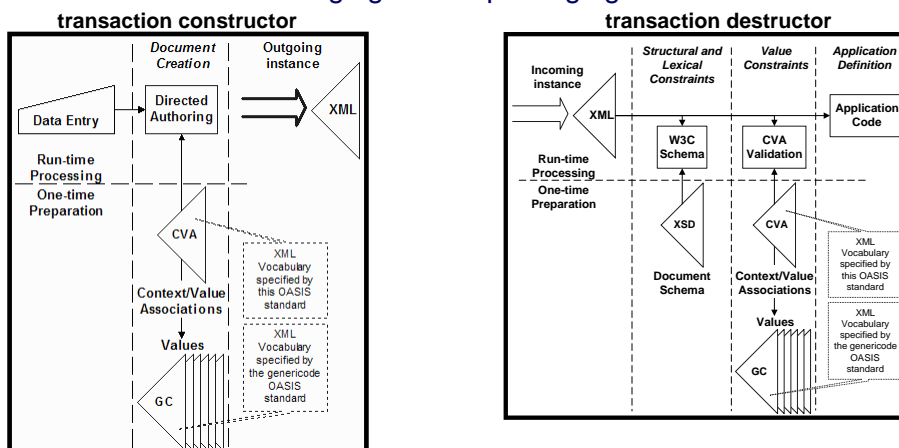
- The transaction/event verb is expressed as the transaction type, and the subject and rest of the predicate is contained in the transaction payload.



The MIT 2008 Information Quality Industry Symposium



## B2B Transaction Packaging and Unpackaging – OASIS UML 2.0



- In OASIS UBL 2.0, metadata/tags are built up from xBIEs (Business Information Entity) and Components





The MIT 2008 Information Quality Industry Symposium



## Conclusions

- **The notion that abstractions such as data models are "real" generates conflict and risk as we expand system "footprints"**
  - The modeler inevitably falls behind events and, in any case, cannot agree with other modelers
- **A "nominalist" approach maintains data as transactions/events**
  - Expresses and stores transactions as "declarative sentences" built around action verbs (as opposed to "state" verbs) and structured vocabulary and code lists
  - Applications that can construct suitable "sentences" can inject new transactions into the "flow," even if not integrated with other transaction "constructors"
- **Today's principal "existence proof" is found in b2b eBusiness flows**
  - B2B transactions are necessarily self-standing and "portable"
  - An entity's eBusiness "corona" (flow of transactions in and out) increasingly defines that entity's "reality" better than a "model"
- **Many Internet 2.0 and 3.0 prospective solutions are facilitated by the nominalist data management approach because**
  - it optimizes the portability and reuse of source data
  - It opens the way for "fit for purpose" rules and conventions





The MIT 2008 Information Quality Industry Symposium



## Data Governance With a Focus on Information Quality

By Gwen Thomas,  
President, The Data Governance Institute



The MIT 2008 Information Quality Industry Symposium



### Objectives of this presentation

- Identify interdependencies between Information Quality (IQ) programs and many “flavors” of Data Governance.
- Describe “flavors” of Data Governance, their stakeholders, and their focus areas.
- Identify opportunities for IQ to piggyback on Data Governance budgets and executive mindshare.





The MIT 2008 Information Quality Industry Symposium



## Three Case Studies

- An “information factory” with a thriving IQ function that became better supported because of Data Governance.
- A large financial institution that wanted a formal IQ function and got it – after funding foundational efforts using budgets from Enterprise Data Management and an executive-sponsored, cross-functional “special project” administered by Data Governance.
- A smaller financial institution that wants formal IQ, and is using Compliance-based and Data Governance-driven requirements, mindshare, and budget to pave the way.

**Three organizations. Three “flavors” of Data Governance.  
Three sets of happy IQ sponsors and evangelists.**

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
[www.DataGovernance.com](http://www.DataGovernance.com)

3



The MIT 2008 Information Quality Industry Symposium



## A Definition for Data Governance



### **Data Governance**

*Data governance* is the organization and implementation of policies, procedures, structure, roles, and responsibilities which outline and enforce rules of engagement, decision rights, and accountabilities for the effective management of information assets.

– John Ladley, Danette McGilvray, Anne-Marie Smith, Gwen Thomas

From *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information™*  
by Danette McGilvray

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
[www.DataGovernance.com](http://www.DataGovernance.com)

4



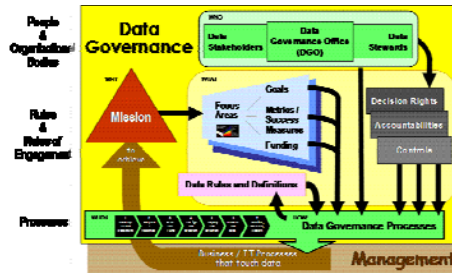
The MIT 2008 Information Quality Industry Symposium



## How Data Governance Can Address Data Quality

- All Data Governance frameworks address data-related rules:  
Making the rules (which can include Data Quality policy, standards, guidelines, and rules), enforcing them, resolving issues, etc.
- How? through
  - People and organizational bodies
  - The “Rules of Engagement” for people, process, and technology
  - Data Governance processes.

The DGI Data Governance Framework  
from The Data Governance Institute



**Any Data Governance framework should be able to address IQ rules and processes.**

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
www.DataGovernance.com

5



The MIT 2008 Information Quality Industry Symposium



## Data Governance “Rules of Engagement”

For projects, programs, or ongoing data-related processes:

- All **data stakeholders** are identified, and their perspectives, needs, and constraints have been considered as the effort's **goals** are clarified.
- The right data stakeholders have been granted appropriate **Decision Rights** to make **rules** and resolve issues.
- **Accountabilities** are established and accepted.
- Efforts are scoped to include human-based and technology-based **controls**.

**Sound familiar? Data Governance programs can establish a firm foundation for IQ efforts.**

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
www.DataGovernance.com

6



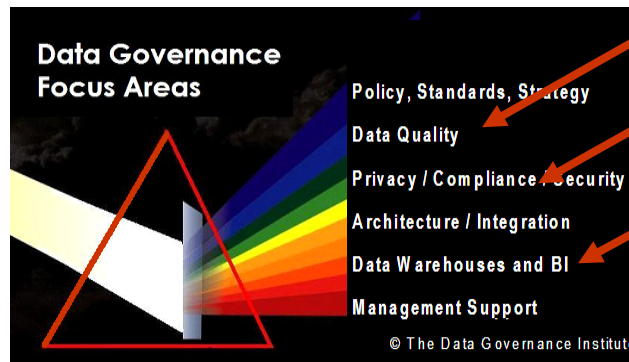


The MIT 2008 Information Quality Industry Symposium



Of course, not all Data Governance efforts focus their attention on the same goals...

### Six Common “Flavors” of Data Governance



...but at least three out of six common “flavors” of Data Governance are concerned about improving the quality of data.

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
www.DataGovernance.com

7



The MIT 2008 Information Quality Industry Symposium



### Data Governance With a Focus on Policy, Standards, Strategy

**IQ is typically not the major focus for this “flavor”**



- **What problem is this addressing?**
  - Some group needs support from a cross-functional leadership body.
- **Who might originate the program?**
  - Data Architecture, Data Management, BPR, or a cross-functional team that needs to align policies, standards, requirements.
- **What is the scope?**
  - The scope of the team needing support.
- **What might Data Governance do (besides work with rules, resolve issues, and provide stakeholder CARE)?**
  - Review, approve, monitor policy; Align sets of policies and standards.
  - Collect, choose, review, approve, monitor standards.
  - Contribute to Business Rules.
  - Identify stakeholders and establish decision rights.

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
www.DataGovernance.com

8



The MIT 2008 Information Quality Industry Symposium



## Data Governance With a Focus on **Data Quality**



- **What problem is this addressing?**
  - – Quality, integrity, usability, of data.
- **Who might originate the program?**
  - – Data Quality group or a business team that needs better quality data. Often starts with a focus on Master Data.
- **What is the scope?**
  - Could be enterprise, local to a department, or local to a project.
- **What might Data Governance do (besides work with rules, resolve issues, and provide stakeholder CARE)?**
  - Set direction for Data Quality. ←
  - Monitor Data Quality. ←
  - Ensure consistent Data Definitions.
  - Identify stakeholders, establish decision rights, clarify accountabilities.

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
www.DataGovernance.com

9



The MIT 2008 Information Quality Industry Symposium



## Data Governance With a Focus on **Privacy / Compliance / Security**

**Typically, Access and Quality are concerns.**



- **What problem is this addressing?**
  - – Data Privacy, Access Management, Information Security controls, **Information Quality**, regulatory compliance.
- **Who might originate the program?**
  - Business or IT. Often comes from a senior management mandate.
- **What is the scope?**
  - Generally enterprise, but often limited to specific types of data.
- **What might Data Governance do (besides work with rules, resolve issues, and provide stakeholder CARE)?**
  - Help protect sensitive data through support for Access Management and Security Requirements.
  - Help define risk, controls, and rules about information quality. ←
  - Help enforce regulatory, contractual, architectural compliance requirements.
  - Identify stakeholders, establish decision rights, clarify accountabilities.

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
www.DataGovernance.com

10

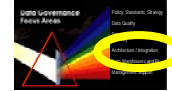


The MIT 2008 Information Quality Industry Symposium



## Data Governance With a Focus on Architecture / Integration

**IQ is typically not  
the major focus  
for this "flavor"**



- **What problem is this addressing?**
  - Challenges moving from a silo environment to integrated or enterprise systems.
- **Who might originate the program?**
  - Data Architecture group or a project addressing a data integration challenge.
- **What is the scope?**
  - Could be enterprise, local to a department, or local to a project.
- **What might Data Governance do  
(besides work with rules, resolve issues, and provide stakeholder CARE)?**
  - Ensure consistent data definitions.
  - Support Architectural Policies and Standards.
  - Support Metadata Programs, SOA, Master Data Management.
  - Bring cross-functional attention to integration challenges.
  - Identify stakeholders, establish decision rights, clarify accountabilities.

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
www.DataGovernance.com

11



The MIT 2008 Information Quality Industry Symposium



## Data Governance With a Focus on Data Warehouses and BI

**Typically,  
Quality is a  
concern.**



- **What problem is this addressing?**
  - Enforcement of rules that affect the format of or the **quality** of data in Data Warehouses, Data Marts, or Business Intelligence systems.
- **Who might originate the program?**
  - Data Management teams or the Business Groups who sponsor/use these systems.
- **What is the scope?**
  - Generally limited to roles and responsibilities for the warehouse. Sometimes this prototype grows to an enterprise effort.
- **What might Data Governance do  
(besides work with rules, resolve issues, and provide stakeholder CARE)?**
  - Establish rules for data usage, data quality, and data definitions.
  - Identify stakeholders, establish decision rights, clarify accountabilities.
  - Clarify the value of data assets and data-related projects.

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
www.DataGovernance.com

12



The MIT 2008 Information Quality Industry Symposium



## Data Governance With a Focus on Management Support

**IQ is typically not the major focus for this “flavor”**



- **What problem is this addressing?**
  - Managers need to make collaborative decisions but either don't know all the stakeholders to involve or have an obstacle to assembling them.
  - The value/impact of data and data-related efforts needs to be assessed.
- **Who might originate the program?**
  - Leadership.
- **What is the scope?**
  - Could be enterprise, local to a department, or local to a project.
- **What might Data Governance do (besides Issue Resolution and Stakeholder CARE)?**
  - Measure the value of data and data-related efforts.
  - Align frameworks and initiatives.
  - Identify stakeholders, establish decision rights, clarify accountabilities.
  - Identify SDLC embedded governance steps and loop-outs for projects.

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
www.DataGovernance.com

13



The MIT 2008 Information Quality Industry Symposium



## What Six “Flavors” of Data Governance Means for IQ Evangelists...

You have the opportunity to

- Piggyback your IQ message onto the Data Governance message, to reach new audiences.
- Inject your requirements into “hot” programs.
- Have these programs lay the foundation for your efforts.
- Have these programs embed IQ rule-making, rules enforcement, and other efforts into your organization's Project Management Lifecycles (PMLCs) and/or System Development Lifecycles (SDLCs).
- Take advantage of diverse funding buckets.



July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
www.DataGovernance.com

14



The MIT 2008 Information Quality Industry Symposium



## Who Can Help You Get on the Data Governance Agenda?

- Most Data Governance programs are designed with multiple layers of decision-making
  - High-level Council that makes strategic decisions, sets direction and prioritizes efforts, provides top-down support, and resolves issues with an enterprise impact. (Cross-functional representation)
  - A committee or team of data stakeholders working at a tactical level to set rules (policies, standards, guidelines, requirements, definitions), and deal with exceptions/infractions. (Cross-functional representation)
  - “In-the-trenches” Data Stewards and/or Data Custodians who work with data as part of their daily jobs. (Federated roles)
  - Plus Data Governance support personnel, typically from a Data Governance Office (DGO) or a Data Management team. (Centralized)

July 16-17, 2008

Gwen Thomas, The Data Governance Institute  
[www.DataGovernance.com](http://www.DataGovernance.com)

15



The MIT 2008 Information Quality Industry Symposium



## Discussion / Questions?

Gwen.Thomas@DataGovernance.com  
+1.321.438.0774



## The MIT 2008 Information Quality Industry Symposium



### About the Data Governance Institute

- The DGI provides consulting, executive mentoring, program development, and information services, including the web's largest data governance resource, [www.DataGovernance.com](http://www.DataGovernance.com).
- The Institute provides a wealth of resources: the free DGI Data Governance Framework, information on data laws, regulations, and standards, whitepapers, case studies, best practices, data humor, and non-technical briefings on data-related issues and disciplines.
- The Data Governance Institute also publishes [www.DataGovernanceSoftware.com](http://www.DataGovernanceSoftware.com), the DGI Data Governance Vendor Showcase, and [www.SOX-online.com](http://www.SOX-online.com), the web's largest source of vendor-neutral Sarbanes-Oxley information.



July 16-17, 2008

### About Gwen Thomas

- President, The Data Governance Institute
- Principal author, The DGI Data Governance Framework
- Author, *Alpha Males and Data Disasters: The Case for Data Governance*
- Personally designed Data Governance programs or helped existing programs become more mature at companies such as Washington Mutual Bank (WaMu), BankUnited, Sallie Mae, NDCHealth/Wolters Kluwer, Wachovia Bank, Disney, and Coors.
- Background in Systems Integration.



Gwen Thomas, The Data Governance Institute  
[www.DataGovernance.com](http://www.DataGovernance.com)

17



The MIT 2008 Information Quality Industry Symposium



## ISO 8000 the International Standard for Data Quality

Peter Benson  
ISO 8000 Project Leader  
Executive director and chief technical officer  
Electronic Commerce Code Management Association (ECCMA)



The MIT 2008 Information Quality Industry Symposium



### ISO 8000 - Data Quality

- ISO 8000 addresses data quality. ISO 8000 is concerned with:
- the principles of data quality;
- the characteristics of data that determine its quality;
- the processes to ensure data quality.



The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – ISO Definitions

### **information**

knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning [ISO/IEC 2382-1:1993]

meaningful data [ISO 9000:2005]

### **data**

re-interpretable representation of information in a formalized manner suitable for communication, interpretation, or processing [ISO/IEC 2382-1:1993]

### **quality**

degree to which a set of inherent characteristics fulfils requirements [ISO 9000:2005]

### **characteristic**

distinguishing feature [ISO 9000:2005]

### **requirement**

need or expectation that is stated, generally implied or obligatory [ISO 9000:2005]



The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – Parts under development

Part 1: Overview, principles and general requirements

Part 2: Terminology

Part 100: Master data: Overview

**Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification**

Part 120: Master data: Provenance

Part 130: Master data: Accuracy

Part 140: Master data: Completeness

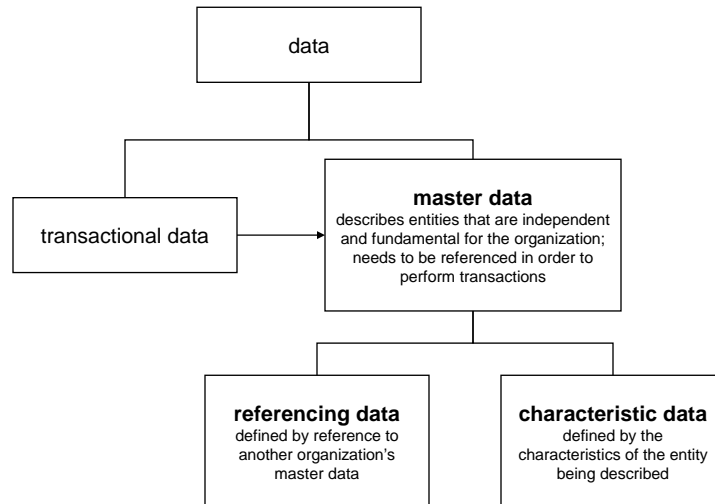




The MIT 2008 Information Quality Industry Symposium



## ISO 8000-100: – Master Data Quality



The MIT 2008 Information Quality Industry Symposium



## ISO 8000-100: – Master Data Quality

### master data

data held by an organization that describes the entities that are both independent and fundamental for an enterprise, that it needs to reference in order to perform its transactions

A master data record is a collection of data element values. Data element values are the fundamental building blocks of electronically stored information, the quality of the data element values is a major determinant of the quality of the information and consequently the accuracy and reliability of the knowledge that can be derived from the information.

The following are considered to be intrinsic components of establishing data quality:

1. **Syntax:** The syntax or arrangement of data element values determines the ease with which data can be integrated within and across organizations.
2. **Semantics (metadata):** The ability to retrieve the definition of metadata (data label) and the quality of the definition in terms of clarity of meaning to all stakeholders determines the portability of the data within and across organizations.
3. **Source of data (provenance):** The ability to track the organization that owns the process that created, validated or transferred the data and the time when the process was performed determines the traceability of the data.
4. **Fitness:** The ability to assess if data meets the requirements of a specific function determines its fitness for the purpose.
5. **Accuracy:** The method through which accuracy is asserted determines the ability to validate accuracy.
6. **Completeness:** The method through which completeness is asserted determines the ability to validate completeness.



The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – Part 110

This part of ISO 8000 specifies requirements that can be checked by computer for the exchange, between organizations and systems, of master data that consists of characteristic data.

**The following are within the scope of this part of ISO 8000:**

- requirements regarding conformance to a formal syntax for master data messages;
- semantic encoding requirements for master data messages;
- requirements regarding conformance to data specifications for master data messages;

**The following are outside the scope of this part of ISO 8000:**

- requirements regarding data not in messages;
- requirements regarding exchange of data that are not master data;
- requirements regarding master data that are not characteristic data;
- Records of the history of the origination, modification, and transfer of custody or ownership of data are commonly referred to as the data provenance ( these are in part 120)
- requirements regarding recording the history of master data; (these are in part 120)
- requirements regarding accuracy of master data; (these are in part 130)
- requirements regarding the management of master data internally within an organization;
- *Data within an organization's enterprise resource planning (ERP) or product data management (PDM) system is out of scope.*



The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – Part 110

### Syntax

Each data set shall contain a reference to the syntax to which the data set complies....The reference shall be resolvable to the specification of the syntax through a mechanism that is publicly available.

### Semantic encoding

Each data element value shall reference all concepts necessary to unambiguously define its meaning. Each reference shall be to a concept dictionary entry contained in a concept dictionary that supports an interface for resolution of a concept identifier.

**Syntax and semantic resolution shall be available at no charge unless the data carries a “fee based encoding” warning label.**

### Conformance to requirements

Each data set shall contain a reference to the data requirements statement to which the data set complies. The reference shall be a globally unambiguous identifier that was used to encode the data set. The reference shall be resolvable to the data requirements statement. The data requirements statement shall be publicly available.



The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – Part 120

### Provenance

This part of ISO 8000 describes requirements for representation and exchange of information about provenance of property value pairs and data sets.

**The following are within scope of this part of ISO 8000:**

- scenarios for data provenance;
- data provenance roles;
- requirements for capture and exchange of data provenance information;

**The following are outside the scope of this part of ISO 8000:**

- exchange format for data provenance information;
- scheme for registering and resolving organization identifiers and person identifiers;
- provenance of data that are not property value pairs or data sets.



The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – Part 130

### Accuracy

This part of ISO 8000 describes requirements for representation and exchange of information about accuracy of property value pairs, records, and data sets.

**The following are within scope of this part of ISO 8000:**

- scenarios for data accuracy;
- data accuracy roles;
- requirements for capture and exchange of data accuracy information;

**The following are outside the scope of this part of ISO 8000:**

- exchange format for data accuracy information;
- accuracy of data that are not master data;
- accuracy of data that are not property value pairs, records, or data sets.



The MIT 2008 Information Quality Industry Symposium



## ISO 8000 - Data Quality – Part 140

### Completeness

This part of ISO 8000 describes requirements for completeness of master data.

#### The following are within scope of this part of ISO 8000:

- scenarios for completeness;
- completeness roles;
- requirements for completeness of master data;
- requirements for exchange of information about master data completeness.

#### The following are outside the scope of this part of ISO 8000:

- exchange format for completeness information;
- completeness of data that are not master data;
- completeness of data that are not property value pairs, records, or data sets.



The MIT 2008 Information Quality Industry Symposium



## ISO 8000-110:2008 Certification

- **Certified Software applications and data cleaning services**
  - Know how to use an open technical dictionary for metadata encoding and mapping
  - Know how to read and write master data requirement specifications
  - Know how to generate requests for master data
  - Know how to read and write encoded master data messages
- **Certified Master Data Quality Managers (data requestors)**
  - Know how to specify their master data requirements
  - Know how to ask for they data they need to validate or complete their master data
- **Certified Quality Master Data Providers**
  - Know how to respond to a request for master data



## **Data Integration and Data Quality: Pharmaceutical Industry Case.**

Sergiy Sirichenko, Vadim Tantsyura, Olive Yuan, Ph.D.  
(Regeneron Pharmaceuticals, Inc., Tarrytown, NY)

Max Kanevsky (Pinnacle21, Plymouth Meeting, PA )

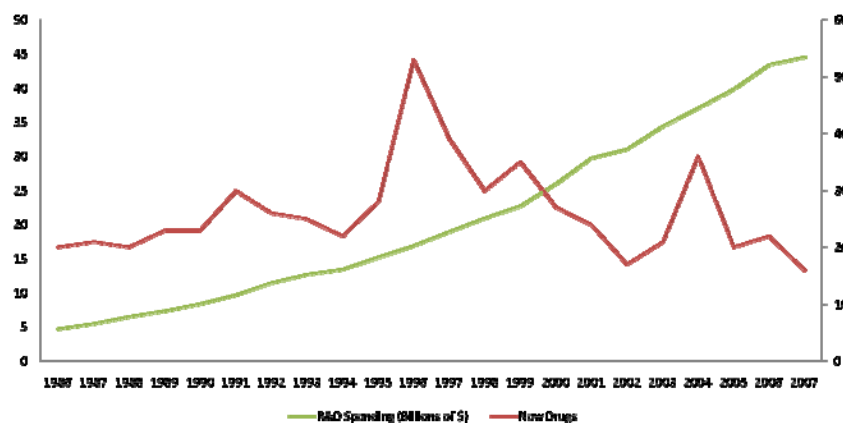
## **Agenda**

- Current process
- Current definition of DQ in pharma
- Data integration issue and examples
- Recommendations

## Introduction

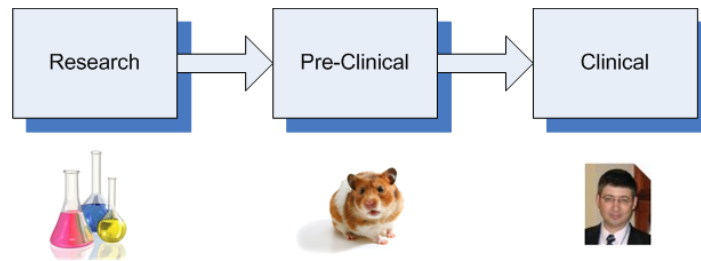
- “The review and approval of new pharmaceuticals by federal regulatory agencies is contingent upon a trust that the clinical trials data presented are of sufficient integrity to ensure confidence in the results and conclusions presented by the sponsor company.” (Society for CDM, Charter of the Committee for Standards for GCDMP, 1998.)

U.S. Drug Industry Spending on Research and Development vs. New Drug Approvals by FDA (1986 – 2007)

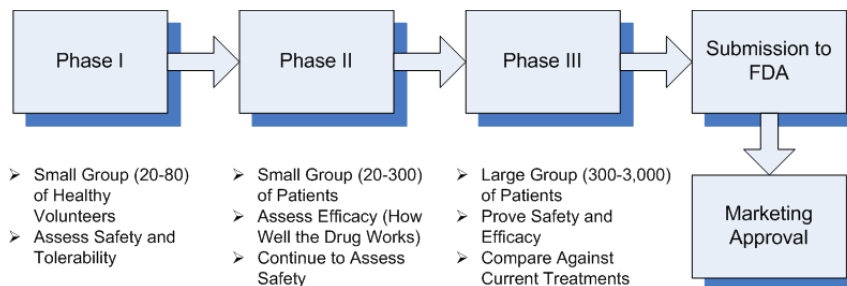


Source: Pharmaceutical Research and Manufacturers of America, *Pharmaceutical Industry Profile* 2006, [http://www.phrma.org/files/2008 Profile.pdf](http://www.phrma.org/files/2008%20Profile.pdf) and FDA Center for Drug Evaluation and Research, *CDER Drug and Biologic Approval Reports*, <http://www.fda.gov/cder/rdmt/>

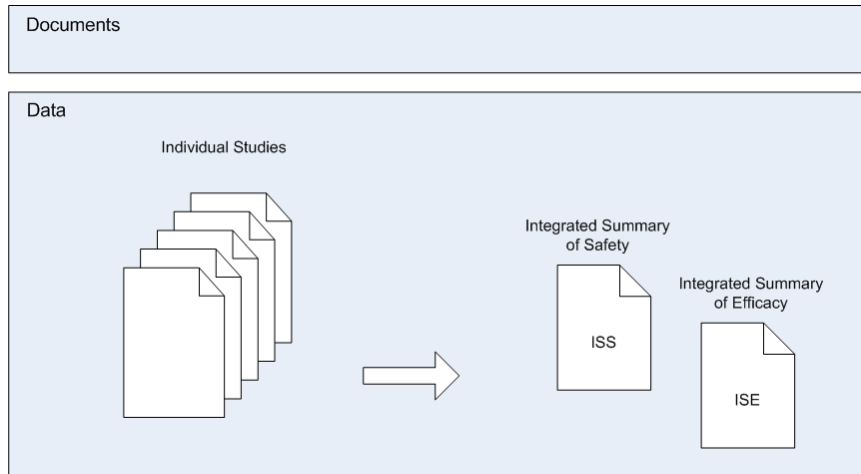
## Drug Development Process



## Clinical Trials



## Submission to FDA



## Definition of Quality

- SCDM adapted the IOM definition:
  - (sufficient) “quality data is data that support conclusions and interpretations equivalent to those derived from error-free data” (Institute of Medicine, Roundtable Report, 1999)
- **Risk-based approach:** Quote from Janet Woodcock (Science Board: FDA’s New Bioresearch Monitoring Initiative, 04 November, 2005):
  - “High-quality Clinical Trial Data:
    - Support integrity of clinical research enterprise
    - Support confidence of public/patients in human studies
    - Provide evidentiary base for product approvals and medical practices”



## FDA Guidance for pharmaceutical industry:

Computerized Systems Used in Clinical Trials (April 1999).

Data quality: attributes

- attributable
- original
- accurate
- contemporaneous
- legible

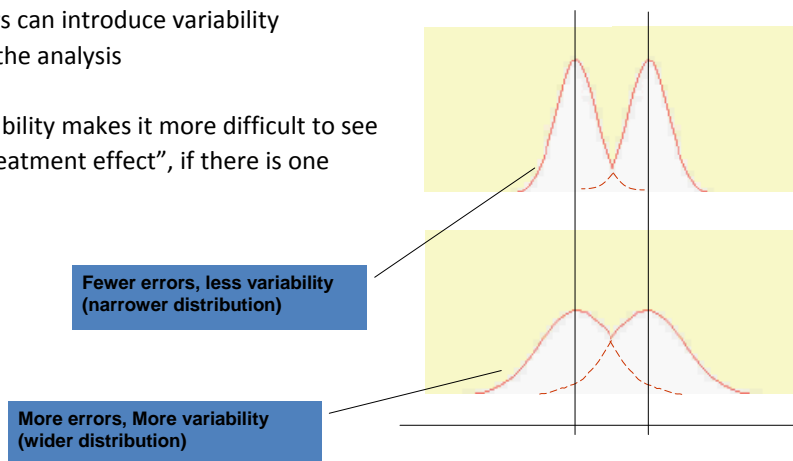
## Common definitions of DATA error in CDM

- **Very often “conveniently” defined as “mismatch between database and the source”.** (This definition is misleading.)
- **GCDMP definition** (v4, p. 77)
  - “A data error occurs when a data point inaccurately represents **a true value**... This definition is intentionally broad and includes errors with root causes of misunderstanding, mistakes, mismanagement, negligence and fraud.”
- **What is true value?**
  - The foundation of traditional data manager is the belief that “source document” represents true value (in 99+% of cases source represents “true value”)
  - Let’s keep in mind this is not always the case (example: a value inconsistent with life)
- The following are NOT examples of DATA errors (according to the definition above):
  - Lack of compliance of procedures to regulations
  - Lack of compliance of practices to written documentation

## Why are Errors Undesirable?

Errors can introduce variability into the analysis

Variability makes it more difficult to see a “treatment effect”, if there is one



Meredith Nahm, MS, CCDM

Director, Clinical Data Integration, Duke Clinical Research Institute

Author, GCDMP Measuring Data Quality and Assuring Data Quality Sections

## Assuring Clinical Trial Data Validity: The Current Process

- **The complexity of the design and the amount of data collected have important influences on data quality**
  - Design of protocol
  - CRFs
  - Data collection systems
- **Training is critical to ensuring that the protocol is followed correctly and the CRFs are properly completed**
  - Clinical investigator
  - Study personnel
- **Clinical site monitoring (can consume 15 to 30 percent of overall trial costs)**
- **Industry data QA procedures**
  - Assembly of all the data from trial
  - Entry of the information into databases
  - Evaluation of the data for quality
  - Audits of clinical sites

## Assuring Clinical Trial Data Validity: The Current Process (cont'd)

- **FDA data analysis (includes clinical and statistical review)**
  - Checking and verification of data from important analyses submitted by the sponsor
  - Performance of exploratory analyses to answer questions that emerge from the review
- **FDA data QA evaluation**
  - Auditing of CRFs to verify the accuracy of tabulated data
  - Evaluation of follow-ups on reported AEs
  - Verification of primary outcome measure at the CRF level
  - An overall assessment of data quality is developed. If serious questions regarding overall data integrity are not resolved, FDA will not approve the application
- **FDA clinical study audit program**
  - A thorough on-site review of these sites is conducted by trained FDA inspectors. Record keeping, adherence to the protocol, informed-consent procedures, and other aspects of the study are assessed. If objectionable conditions are found, a report (FDA Form 483) is provided to the PI at the conclusion of the audit.
- **FDA enforcement activities**
  - If an investigator found to have serious or repeated problems in performing clinical studies, FDA will take steps to debar the individual from performing trials for regulatory purposes. In cases of fraud, criminal prosecution may be pursued.

## Main Thesis:

- Individual study data can be perfect (e.g. 100% clean), but when you integrate them the quality suffers

## Data Integration Issue

- FDA collects all data from all studies, but their actual usage is limited to individual studies
- To perform meta analysis data integration is needed. This is a very time and resource consuming process that makes the collected data much less useful.
  - Example: FDA blood pressure drugs analysis
  - Analogy: Cryptography

## Why clinical data are different?

- No required data standards in Pharma
- Usually each study is designed separately and physician rather than statistician is its owner
- Different data collection systems

## Data

- Structure
  - Terminology
  - Content
- } Standards: CDISC, HL7, ISO, etc.

## Structure

- Different structure is
  - a potential source of errors during integration and
  - obstacle for integration itself
- In addition, lack of standards leads to errors during data collection

Examples:

- Gender: M=1, F=2 vs. F=1, M=2
- Date formats: 07-08-12, 08-12-07, 12-07-08

## Structure: (cont'd)

- Adverse Events:
  - Headache volunteered
  - Headache elicited from checklist (recall bias)
- Smoking
  - Dichotomy
    - Smoker
    - Non-smoker
  - Quantification or Qualification:
    - Smokes less than 1 pack per year
    - Quit smoking within past year
    - Smokes less than 1 pack per week, 2-4 pack, more...
    - Smokes cigarettes, cigars, pipes;

## Terminology

- Definition- the same term is understood in the same way by different people in different places at different times

### Examples:

- Myocardial infarction
- Age in China
- Age in raw data and SAE reporting
- Same name for different lab analytes: Lymphocyte count and percentages
- Same name for bilirubin in hematology and in urinalysis

## Contents: Subject Race example

Study A	Study B	Integrated
<i>White</i>	<i>White</i>	<i>White</i>
<i>Black</i>	<i>Black</i>	<i>Black</i>
<i>Other</i>	<i>Asian</i>	<i>Other</i>
	<i>Other</i>	

Any coding leads to loosing of information

## Contents: Subject Ethnicity example

Old way	Current way	
Race:	Race:	Ethnicity:
<i>White</i>	<i>White</i>	<i>Hispanic</i>
<i>Black</i>	<i>Black</i>	<i>Non-hispanic</i>
<i>Hispanic</i>	.....	

Any coding leads to loosing of information

## Content: AE Causality example

Study A

Study B

*Related*

*Definitely*

*Not Related*

*Definitely Not*

*Possibly*

*?*

*Probably*

*?*

*Unlikely*

*?*

Depends on your choice the analysis results can be different

## Content: Study Day example

Study A: *Day 1*

*Day 28*

Study B: *Day 1*

*Day 21 Day 35*

Integration:

*D21+D28 or D35+D28 ?*

Depends on your choice the analysis results can be different



## Do not forget during analysis they are still apples and oranges!

Data can be perfect, but source populations are different

Phase 1: Healthy volunteers

Phase 2: Ideal patients and different diseases

Phase 3: Real patients with targeted disease

10 apples and 90 oranges → They are fruits → More than 90% of fruits are orange in color

Sometimes less data means better quality

## Conclusion

- Issue:
  - Data integration has bearing on data quality that is under-researched in pharmaceutical industry now.
- Recommendations:
  - Use existing standards. If none, develop your own ones
  - Design individual studies keeping integrating data base as final goal in mind



## *MDM Enterprise Analyzer: a framework to support centralized and local master data quality analysis*

---

MIT IQ Industry Symposium,  
Cambridge, Massachusetts, USA

by Kai-Uwe Baryga

SYDECON  
Systems Design & Construction GmbH  
Hans-Urmiller-Ring 46  
D-82515 Wolfratshausen  
Germany  
Email: [kai-uwe.baryga@sydecon.de](mailto:kai-uwe.baryga@sydecon.de)

[www.sydecon.de](http://www.sydecon.de)



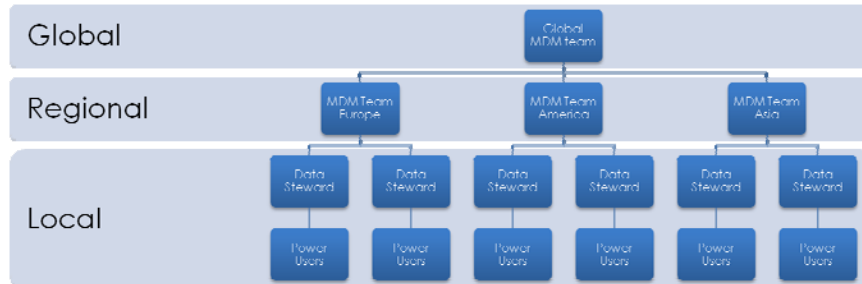
## **Master Data Key Success Factors** Strategy and Governance

---

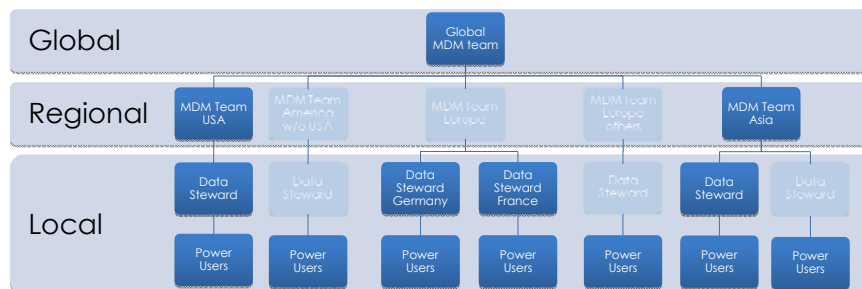
- Global MDM Management
  - Master data strategy and organization
  - Global master data quality and service controlling
- Global MDM Teams
  - Definition of global standards for master data objects and processes
  - Definition of global maintenance rules
  - Definition of global quality and service standards
  - Collaboration with regional MDM teams and global business drivers
- Regional MDM Teams
  - Definition of regional standards for master data objects and processes
  - Definition of regional rules in addition to global rules
  - Regional quality and service controlling
- Data Stewards
  - Collaboration with local business and MDM power users
- MDM Power Users
  - Maintenance of certain MDM objects or object parts



## Hierarchical MDM Organizations



## Typical MDM Organizations



- Incomplete, unbalanced hierarchies
- Inhomogeneous MDM experience and skills



## Why improve MD Quality?

---

- Transactional processes base on MD objects, i.e. insufficient MD quality leads to ...
  - ... delays in the supply chain
  - ... expensive additional manual work
  - ... increased risks (credit limit, dangerous goods)
  - ... issues with reporting and analysis systems
- Demands due to
  - ... local laws and legal restraints
  - ... customer or vendor requirements
  - ... internal company policies
  - ... restraints required due to stock exchange (e.g. NYSE/SEC)



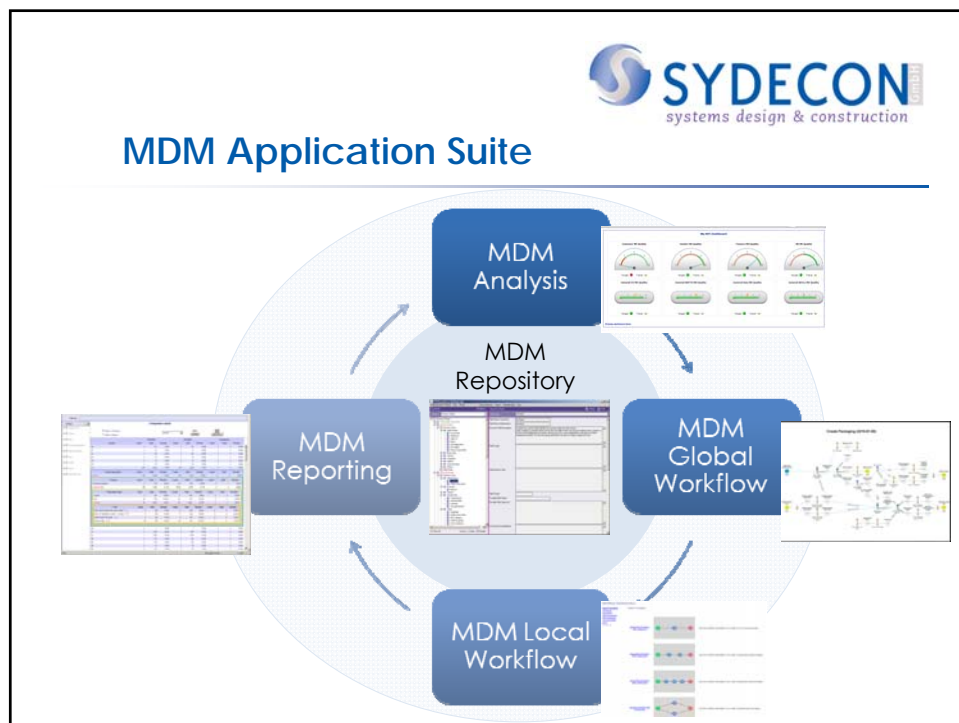
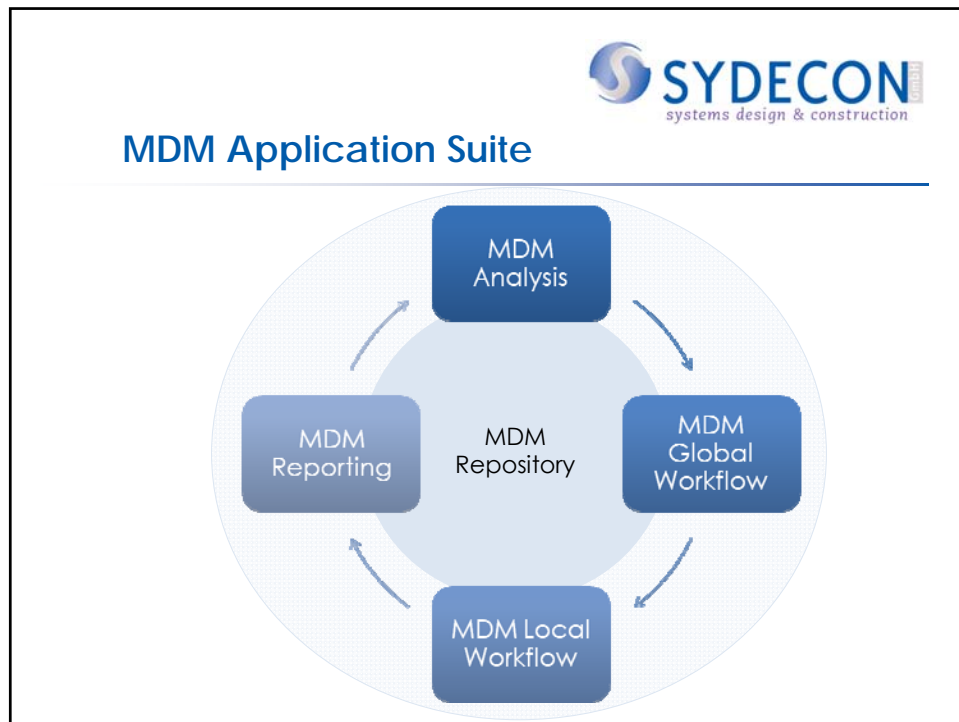
## Conclusion

---

Increased MD quality ...

- ... reduces costs
- ... reduces risks
- ... increases transparency (reporting / analysis)
- ... speeds up the supply chain
- ... observes the laws and restraints

Better MD quality makes your CEO sleep better!



## MDM Analyzer Technology



### MDM Analyzer ...

- ... is an analytical system for master data
- ... allows rules-based analysis
- ... supports local and global rules
- ... supports local and global responsibilities
- ... allows views restricted by access rights
- ... bases on data warehouse technology
- ... is a centrally installed and maintained system

## MDM Analyzer Business



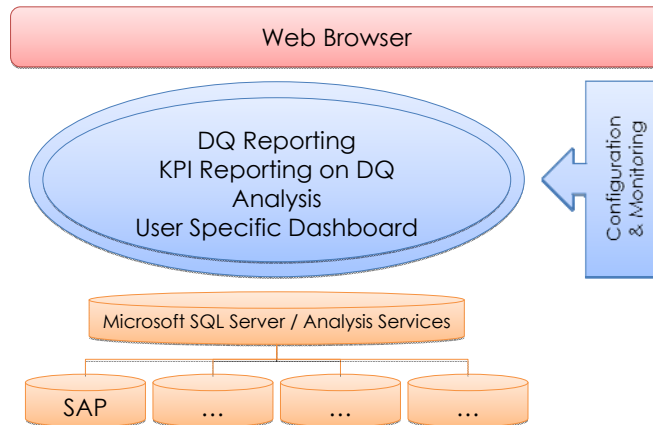
### MDM Analyzer ...

- ... shows MD quality indicators on high and detailed level
- ... shows the increase and decrease of MD quality over time
- ... finds invalid master data and helps to initiate correction processes in MDM Enterprise Workflow

MDM Analyzer supports MD governance  
and controlling tasks in the complete MD organization



## Architecture



## Basic Principles of MDM Analysis

- Administration Areas  
An administration area allows the delegation of responsibilities to area administrators and the mapping between the MDM organization and the MDM Analysis system.
- Sources and Rules  
Every rule is based on a source. A source defines the data set that is analyzed by a rule:  
$$\text{Rule Data Set} \subseteq \text{Source Data Set} \subseteq \text{Database}$$



## Rules based Analysis

---

- There are other ways to detect incorrect data in IT systems, e.g. statistical methods, neuronal networks, etc.
- The rules-based approach is used because the requirements on MD quality are based on rules.
  - MD is used in systems where transactions require well specified information
  - Legal and financial restraints are well specified and define rules
  - Customer and vendor requirements are rules based



## Rules and Localization

---

- Rules can be global
  - Required by a global IT system
  - Based on global company policies
  - ...
- Rules can be local
  - Local laws and restraints
  - Local customer or vendors requirements (e.g. by local logistics providers)
  - ...





## What can be analyzed?

Sample rules:

- Customer names should start with a capital letter
- European customers need to have a VAT number
- For each vendor at least one contact needs to be assigned
- Ordered products should have status 'in process'
- For each material in SAP marked as dangerous good, security instructions must be available in the fire department system
- ...



## Source Management

Manage sources

Source: 200 Supplier table with Purchase info

Name: 200 Supplier table with Purchase info

Description: Supplier table with Purchase info (LFA1, EINA, EINE)

Source table name(s): LFA1, EINA, EINE

Data specification: LFA1.LIFNR = EINA.LIFNR  
AND EINA.INFNR = EINE.INFNR

Source type: Global scope and iterable

Distinct object: Distinct object Table name ID column name

KDS assignments:

KDS type	KDS table name	KDS ID column name	KDS value
Country	LFA1	LAND1	
=select item			

Show source history


Clear form Validate SQL syntax Save Save as new Delete

Original database tables

Joins

Distinct counting

Access restrictions



**SYDECON**  
systems design & construction

## Rules Management

Manage rules

Process: 06 Customer Management [Edit Process Assignment](#)

Rule: 300 Customer ICC with wrong Rec.account [Search rules](#)

---

Name: 300 Customer ICC with wrong Rec.account [Show source meta data](#)

Description: 300 Customer ICC with wrong Rec.account

Detail display columns:

SQL Name	Display Name	Sort Order	
KNAL.KUNNR	KNAL.KUNNR	1	<a href="#">Edit</a> <a href="#">Delete</a>
KNBL.AKONT	KNBL.AKONT	5	<a href="#">Edit</a> <a href="#">Delete</a>

[Add](#)

Source: 300 Customer Master Comp Code and Sales Data

Local KDS identifier:

KDS type	KDS value
Sales Organization	RUS0 <a href="#">Browse</a>

[Show source definition](#)

Rule specification: KNAL.TOKD='ICC' and KNBL.BUKR='0182' and KNBL.AKONT<>'0000300000'

Area: Area East Europe (Russia/Ukraine)

Result type: Critical Error

KPI rule: ☐ [Add KPI Calculation](#)

[Assign workflows](#)

[Show value lists](#)


[Show rule history](#)

Results

Source

Filter

Error types



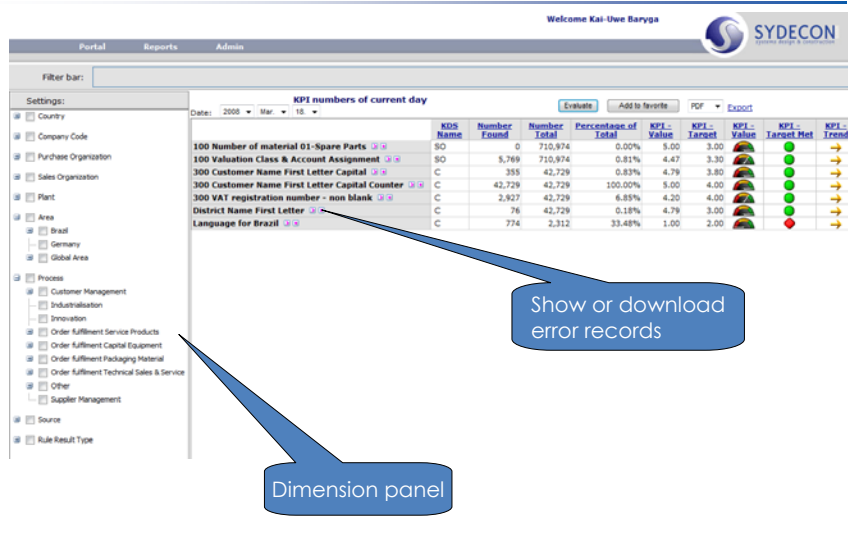
**SYDECON**  
systems design & construction

## MDM Analysis Results

- Dimension Panel in order to navigate in MDM dimensions:
  - Processes, Sources, Rules
  - Areas
  - Key data structures
- Different types of result tables and graphs
  - Current day
  - Compare with history
  - Drill down
  - Drill through to source data
- Integration
  - Download of error records
  - Integration of MDM Enterprise Workflow

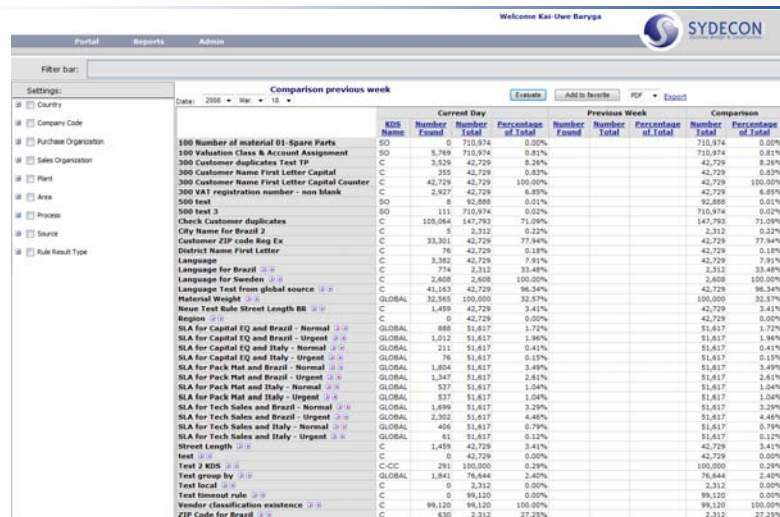
## MD Quality Analysis

### List of KPI with target and trend



## MD Quality Analysis

### Compare with last week



## MD Quality Analysis

### Drill Down



Filter bar

Settings

☒ Company Code

☒ Country

☒ Plant

☒ Purchasing Organization

☒ Sales Organization

☒ Area

☒ Division with Restriction

☒ Process

☒ Source

☒ Rule Result Type

Comparison report

Country

Period

Start To End

10/10/2007

10/10/2007

Comparison

	Count	Total	Percent	Count	Total	Percent	Count	Total	Percent
AE	0	511	0.00%	0	511	0.00%	0	0	0.00%
AF	0	21	0.00%	0	21	0.00%	0	0	0.00%
AL	1	161	0.62%	1	161	0.62%	0	0	0.00%
AO	0	63	0.00%	0	63	0.00%	0	0	0.00%
AIH	0	91	0.00%	0	91	0.00%	0	0	0.00%
AS	0	133	0.00%	0	133	0.00%	0	0	0.00%
AA	104	9,322	3.52%	104	9,322	3.52%	0	0	0.00%
SA Sales Organization									
Count	Total	Percent	Count	Total	Percent	Count	Total	Percent	
SA/USA	104	9,322	3.52%	104	9,322	3.52%	0	0	0.00%
Source									
Count	Total	Percent	Count	Total	Percent	Count	Total	Percent	
Customer Accounts	0	740	0.00%	0	740	0.00%	0	0	0.00%
Customer type	104	4,476	4.17%	104	4,476	4.17%	0	0	0.00%
Rule Result Type									
Count	Total	Percent	Count	Total	Percent	Count	Total	Percent	
Counter	0	740	0.00%	0	740	0.00%	0	0	0.00%
Error	92	740	12.33%	92	740	12.33%	0	0	0.00%
Warning	92	2,364	3.09%	92	2,364	3.09%	0	0	0.00%
Rule									
Count	Total	Percent	Count	Total	Percent	Count	Total	Percent	
300 Customer Name First Letter Capital	0	740	0.00%	0	740	0.00%	0	0	0.00%
303 VAT registration number - non tags	0	740	0.00%	0	740	0.00%	0	0	0.00%
302 VAT registration number - non tags	0	740	0.00%	0	740	0.00%	0	0	0.00%
Street Length	92	740	12.33%	92	740	12.33%	0	0	0.00%
AS	0	7	0.00%	0	7	0.00%	0	0	0.00%
A7	0	1,585	0.00%	0	1,585	0.00%	0	0	0.00%
A7	7	7,063	0.10%	7	7,063	0.10%	0	0	0.00%
AZ	0	161	0.00%	0	161	0.00%	0	0	0.00%
AA	2	357	0.56%	2	357	0.56%	0	0	0.00%
BB	0	39	0.00%	0	39	0.00%	0	0	0.00%
AB	0	170	0.00%	0	170	0.00%	0	0	0.00%

Print

7

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/2007

10/10/20

27. Mai 2008

## MD Quality Analysis

### Initiate MD Change Request



Welcome Cornelius Wolf



Error records: Customer name first letter as capital						
Column display mode: <input type="text" value="UserFriendlyName"/>						
Customer Number	Address Number	Name 1	Country	Language	MDM Requests	
0000010268	1120114504	Krey to be used	NO	DE		
0000017111	1120047889	Reinhold Kallisch	DE	DE		
0000090117	1120024408	Sturm-Haupt, Zweigstelle-Hessing	DE	DE		
0000090269	1120124214	Speiser, Adm.	DE	DE		
0000000236	1120166494	Musical Elements GmbH	DE	DE		
0000000402	1120161165	Wien-Ad	DE	DE		
0000000466	1120119998	WOLFF Leasing GmbH	DE	DE		
0000099771	1120024611	Musik-Haus GmbH, Vertriebsgesellschaft	DE	DE		
0000090830	1120166434	Wien-Ad	DE	DE		
0000094552	7000005333	Deutsche Decker AG/DA	DE	DE		
0000094981	7000009840	Monomats	DE	DE		
0000090861	7000009018	Als Wirtschaftsprüfungsbüro für	DE	DE		
0000095010	7000009101	WERTZ Fachhandel und	DE	DE		
0000099025	7000009283	W & B MC	DE	DE		
0000110543	1120285237	Speiser-Haus, Zweigstelle-Hessing	PL	DE		
0000143138	1120281536	Deutsche Postbank AG	NL	DE		
0000149028	1120297803	Kyushu	NL	DE		
0000130992	7000002339	refracto & V	GB	DE		
0000111157	7000007247	Star bay products Ltd	GB	DE		
0000176057	7000007365	Star bay products Ltd	GB	DE		
0000176069	7000007381	musical elements ltd	GB	DE		
0000176070	7000007376	Musical Elements Ltd	GB	DE		
0000176076	7000007354	Musical Elements Ltd	GB	DE		
0000176078	7000007387	Als Buchverlag	GB	DE		
0000176080	7000007396	musical elements ltd	GB	DE		
0000176107	7000007431	exterior fluids ltd	GB	DE		
0000176116	7000007442	exterior fluids ltd	GB	DE		
0000176120	7000007425	golden fluids products ltd	GB	DE		
0000176131	7000007467	by hanna & son ltd	GB	DE		
0000176134	7000007451	Went Imaging fluids ltd	GB	DE		

Link to MDM  
Enterprise  
Workflow Process



## MDM Enterprise Workflow Initiated by MDM Analyzer

Save Request   Validate Form   Submit Request

**Change Vendor Master Data**

**General Request Data**

**Standard Data**

Requester Name: Kai-Owe Baryga

Title for Inbox: Change vendor due to missing VAT number

Request Urgency: Normal

Comments:

**General Data**

**0. Initial Screen**

Vendor: 0005026039

Company Code: 0120 TP International s.A.

Purchase Organization: BL32 TP Eq.Land Pur.Org

Account Group: LHM Locally Managed Vendors

**1. Address**

Name: Sydecon Systems Design & Constructi

Name 2: GmbH

Name 3:

Name 4:

Street: Hans-Ullrich-Ring

Street 4:

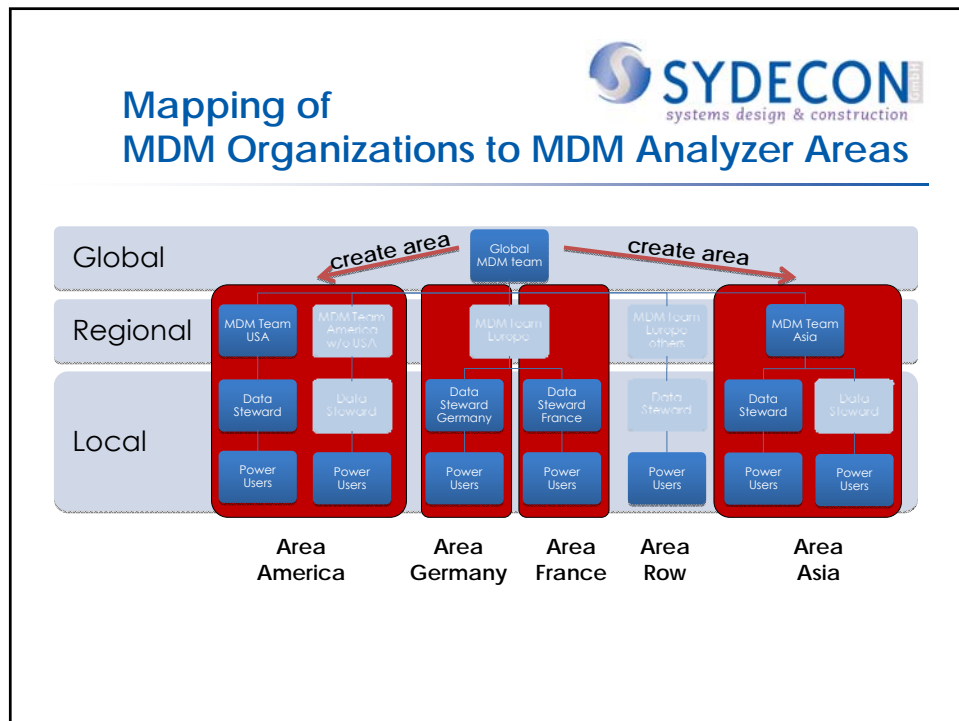
Street 5:


Postal code: 82515

City: Wolfratshausen

Country: DE GERMANY

Data  
preloaded  
from ERP  
System (SAP)





## Area Management

Areas:

Name:


Description:

Area administrator assignment

Available users

Assigned area admins

Baryga, KaiUwe



Area

Area Administrators




## Use Best Practice

Area administrators ...

- ... often have excellent business experience
- ... have to 'feel' the results of insufficient master data quality
- ... do their best to get perfect master data quality in order to reduce effort for corrections

This leads to best practice solutions that can be reused and delivered centrally



**SYDECON**  
systems design & construction

## Best Practice

### Setting in rules

Manage rules

Process: 06 Customer Management [Edit Process Assignment](#)

Rule: 300 Customer ICC with wrong Rec account [Search rules](#)

[Show source meta data](#)

Name: 300 Customer ICC with wrong Rec account

Description: 300 Customer ICC with wrong Rec account

Detail display columns:

SQL Name	Display Name	Sort Order	
KNAL_KUNNR	KNAL_KUNNR	1	<a href="#">Edit</a>
KNBLAKONT	KNBLAKONT	5	<a href="#">Edit</a>

Source: Local KDS identifier

Area: Brazil

Result type: Error

KPI rule: ☒ [Edit KPI Calculation](#) [Remove KPI Calculation](#)

[Assign Workflows](#)

[Show value lists](#)

[Show rule history](#)

Area: [Area](#)

Result type: Critical Error

KPI rule: ☐ [Add KPI Calculation](#)

[Assign Workflows](#)

[Show value lists](#)

[Show rule history](#)

Global Administrator can make rules "global"



**SYDECON**  
systems design & construction

## Access Rights

- Access rights base on the MDM areas and KDS
- Key Data Structures (KDS) are business object related organizational structures, e.g. sales or purchasing organizations, countries and plants
- A user can see data of global rules and rules of his area, if he has the KDS assigned that is required for the data
- E.g. in order to see Brazilian customers with an incorrect address, the user needs to belong to the area South America and have access to the sales organization BR00

## Access Rights Setting in Rules



**Manage rules**

Process: 06 Customer Management [Edit Process Assignment](#)  
 Rule: 300 Customer ICC with wrong Rec. account [Search rules](#)

Name: 300 Customer ICC with wrong Rec. account [Show source meta data](#)  
 Description: 300 Customer ICC with wrong Rec. account

Detail display columns:

SQL Name	Display Name	Sort Order	
KNAL.KUNNR	KNAL.KUNNR	1	<a href="#">Edit</a> <a href="#">Delete</a>
KNBL.AKONT	KNBL.AKONT	5	<a href="#">Edit</a> <a href="#">Delete</a>

Source: 300 Customer Master Comp Code and Sales Data  
 Local KDS identifier: **KDS type** Sales Organization **KDS value** RU60 [Browse](#)

[Show source definition](#)  
 Rule specification: KNAL.KTOID='3CC' and KNBL.BUKRS='0182' and KNBL.AKONT<>'0000300000'

Area: Area East Europe (Russia/Ukraine) [Add KPI Calculation](#)  
 Result type: Critical Error  
 KPI rule: [Assign workflows](#)  
[Show value lists](#)  
[Show rule history](#)

KDS assignment

Area assignment

## Access Rights Setting for Users



**Manage users**

[Show search path](#)  
 Users: Beryga, KaiUwe

Last name: Beryga First name: KaiUwe  
 Login name: TP1GEBARYGAK E-Mail: Kai-Uwe.Beryga@sydecon.de  
 Column display mode: Surname

**User Roles**

Available user roles: KDSQualityAdministrator  
 Assigned user roles: MCMAreaAdmin, MCMReportingAdmin, MCMReportingUser

**KDS Assignment**

Available countries: AD, AE, AF, AG, AI  
 Assigned countries: AD, AE, AF, AG, AI

Available company codes: 0001 SAP Brazil, 0102 TP Packaging Sol. S.p.A., 0103 Tetra Pak Europe S.A., 0107 Tetra Pak (China) S.A., 0108 TP Technical Service Asia, 0110 TP Technical Service M.E.  
 Assigned company codes: 0001 SAP Brazil, 0102 TP Packaging Sol. S.p.A., 0103 Tetra Pak Europe S.A., 0107 Tetra Pak (China) S.A., 0108 TP Technical Service Asia, 0110 TP Technical Service M.E.

Available purchase orgs: 0001 Einkauf.org, 0001 AEST RDC OME Pur. Org, JAR31 TP Argentina Pur. Org  
 Assigned purchase orgs: 0001 Einkauf.org, 0001 AEST RDC OME Pur. Org, JAR31 TP Argentina Pur. Org

Application role

KDS assignments

Area assignment





## KPI Reporting

### General

---

- Users can specify KPIs and KPI groups based on existing rules
- Multiple KPIs per rule can be defined
- KPIs are shown in main analysis page
- KPIs can be shown in the KPI dashboard of MDM Analyzer and in the 'Welcome Page' of the MDM portal.
- A user can specify which KPI he wants to use in his dashboard on an individual base



## KPI Reporting

### Rules Result Normalization

---

- The results of rules can be normalized based on the absolute number of errors found or the error percentage
- Normalized Rule result =  $f(\text{rule result}) \in [0, 1]$
- The function  $f$  can be
  - linear, based on the percentage
  - A step function based on values or percentages

## KPI Reporting

### Result Weighting



- The normalized values can be weighted to a grade between 1 and 5
- In addition a target can be assigned to a rule. A target is also a value between 1 and 5.


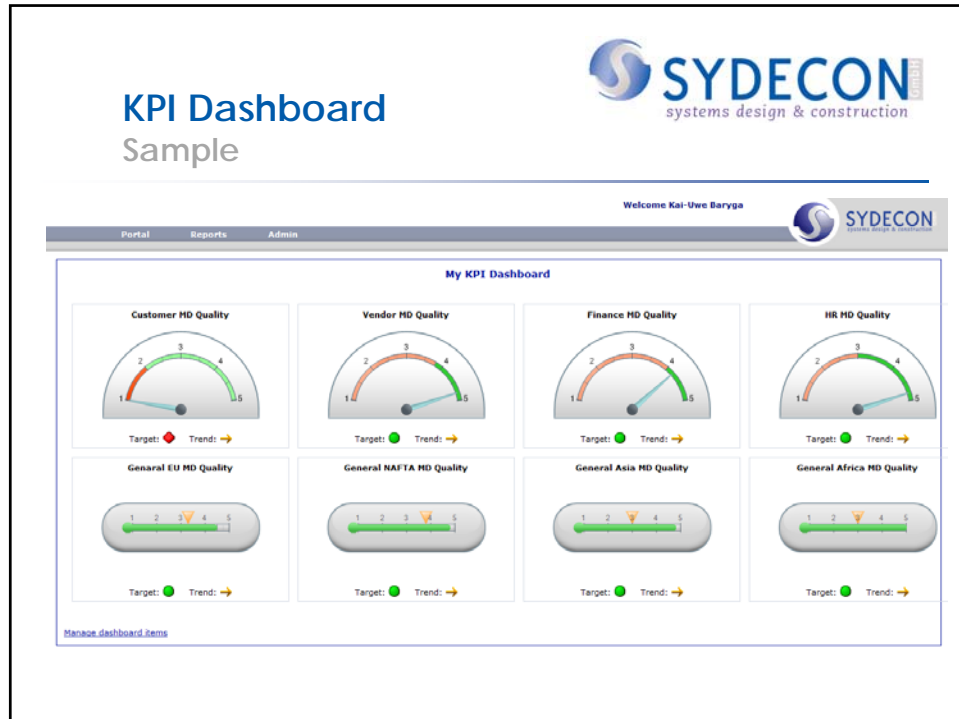


## KPI Reporting

### Grouped KPI



- KPIs can be grouped
  - A KPI group consist of one or more KPIs  
E.g. Customer KPI group can consist of a Customer Address KPI, a Customer Finance KPI, etc.
- Each KPI in a KPI group can be weighted with percentages.  
A more important KPI will have a higher percentage than a less important
- The user can drill down from KPI groups to the individual KPIs in his dashboard



## Performance Aspects

Critical performance areas

- Data loads from ERP systems to the staging area database (less critical than in transactional data warehouses but still a point)
- The Source should not join too many very large tables (we currently use tables with > 250 million records)
- Multiple KDS assignments should be handled with care (no problem in case of real hierarchies, unfortunately this is not always found in SAP environments)
- Drill through with many records

After processing  
the analysis does not cause high loads!



## Technology

---

- Microsoft SQL Server 2005
  - Database Service
  - Analysis Services
  - SQL Server Integration Services (SAP connector)
- Microsoft Internet Information Server
  - Microsoft .NET
  - Developed in C#
- SYDECON is a 



## Summary

---

- The support of central and local analysis is essential for MD Quality control
- Rule based analysis allows direct implementation of business and governance requirements
- Central and local management of rules and access rights make the system more efficient and effective
- Central systems management allows global control of system resources
- Integration in MD maintenance systems adds additional benefits



## Ongoing Development

---

- Closer integration of repository and analysis
  - Specify rules with data specification in repository
  - Specify relationship in repository that can be used in analysis
- Additional automated tasks based on analysis results
  - Mass maintenance
  - Automated data correction



---

## Discussion & Demo