The MIT 2008 Information Quality Industry Symposium

Rapid Corporate Growth and Information

Steve Sarsfield, Trillium Software

The MIT 2008 Information Quality Industry Symposium

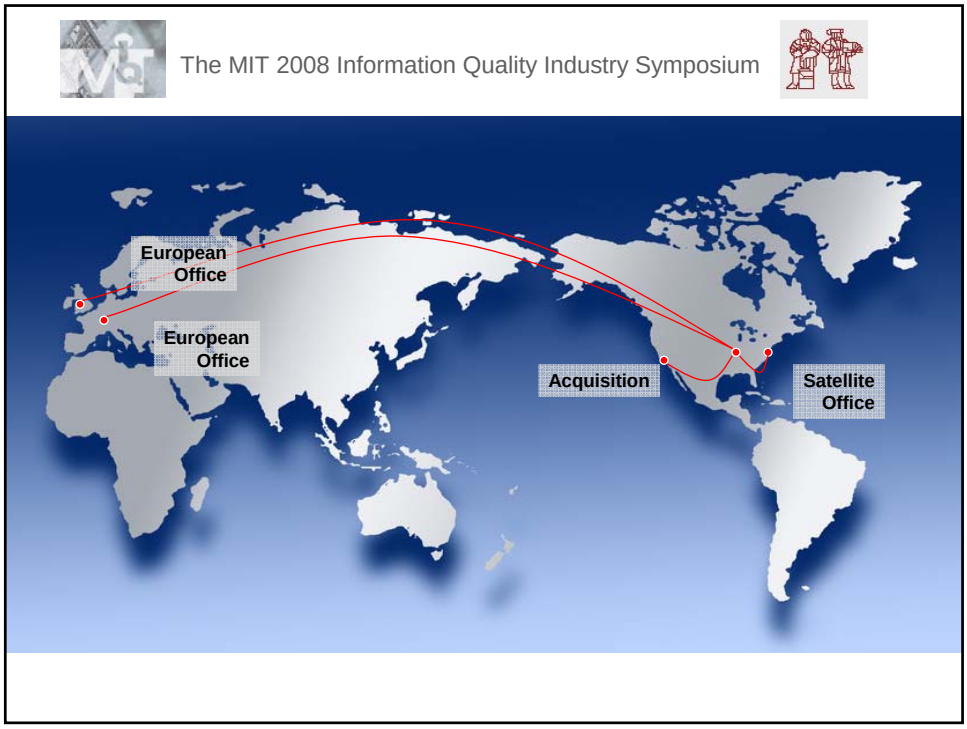
Agenda

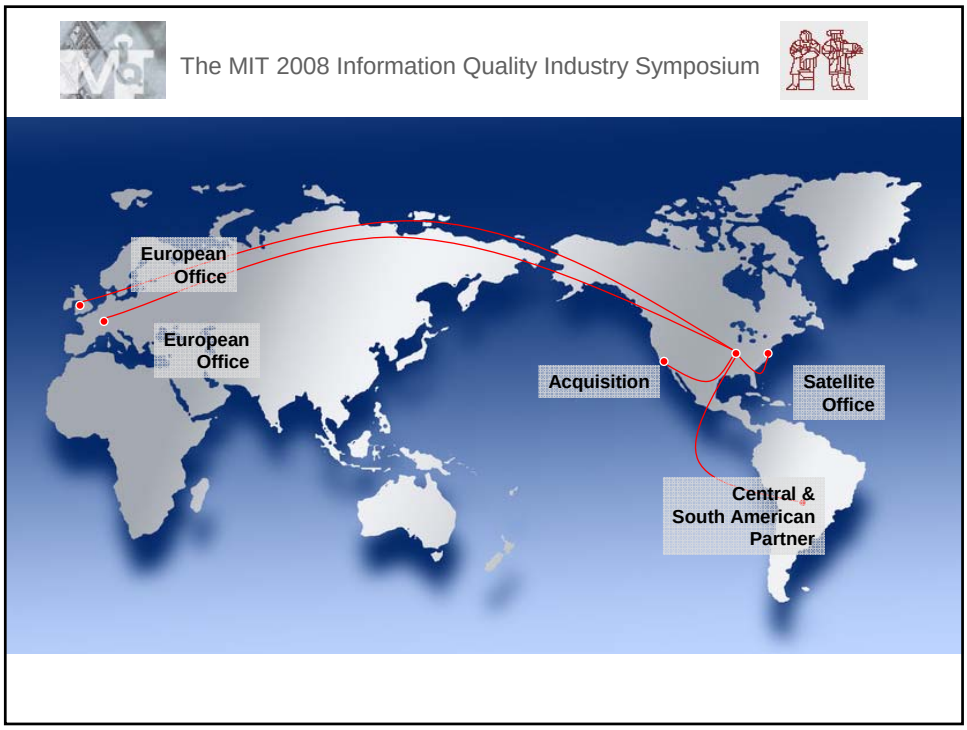
- How Companies Grow
- The Data Components of Company Value
- Effects of Rogue Data Quality Processes
- Sorting Out a Large Company's Problems
  - Cleansing and Matching
  - Domain Specific Knowledge
  - Platform Unification

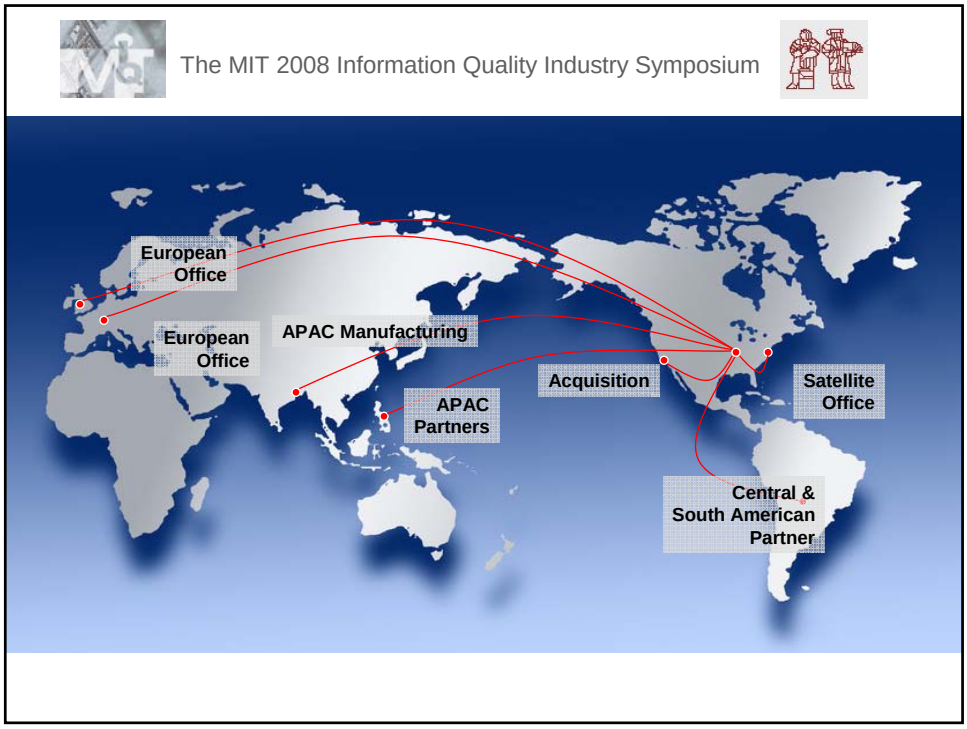
The MIT 2008 Information Quality Industry Symposium

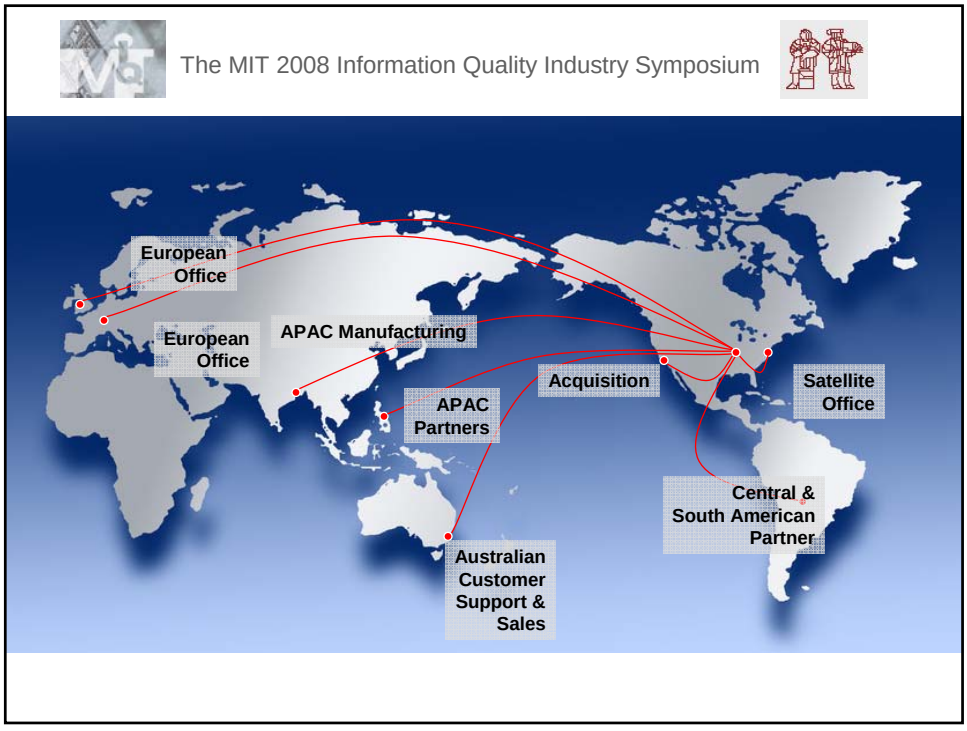
How Companies Grow

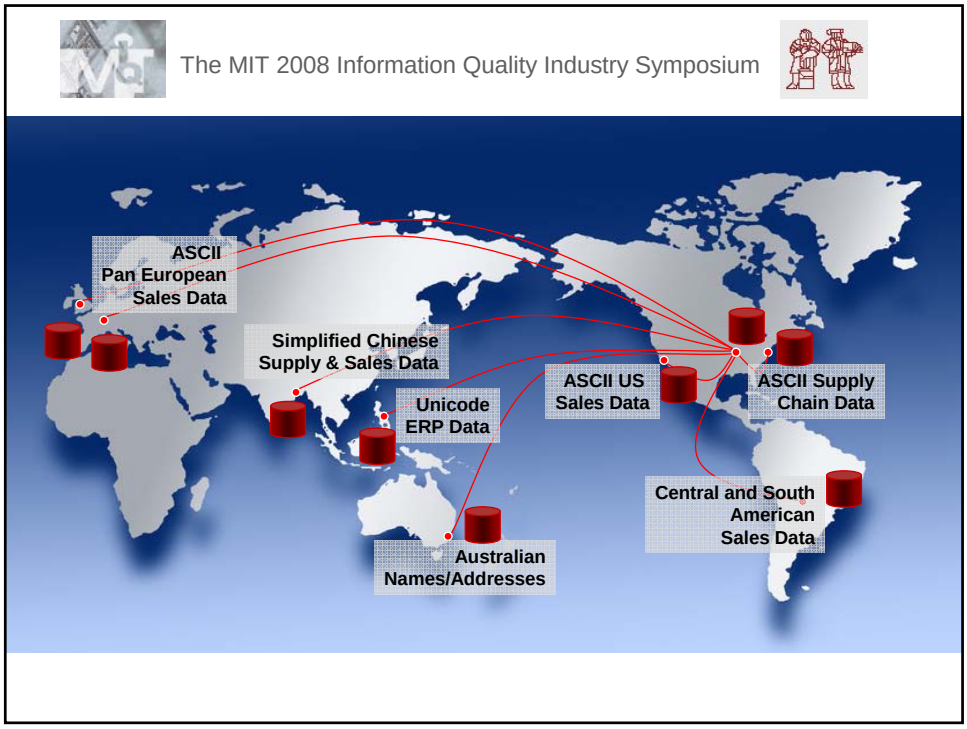
The MIT 2008 Information Quality Industry Symposium

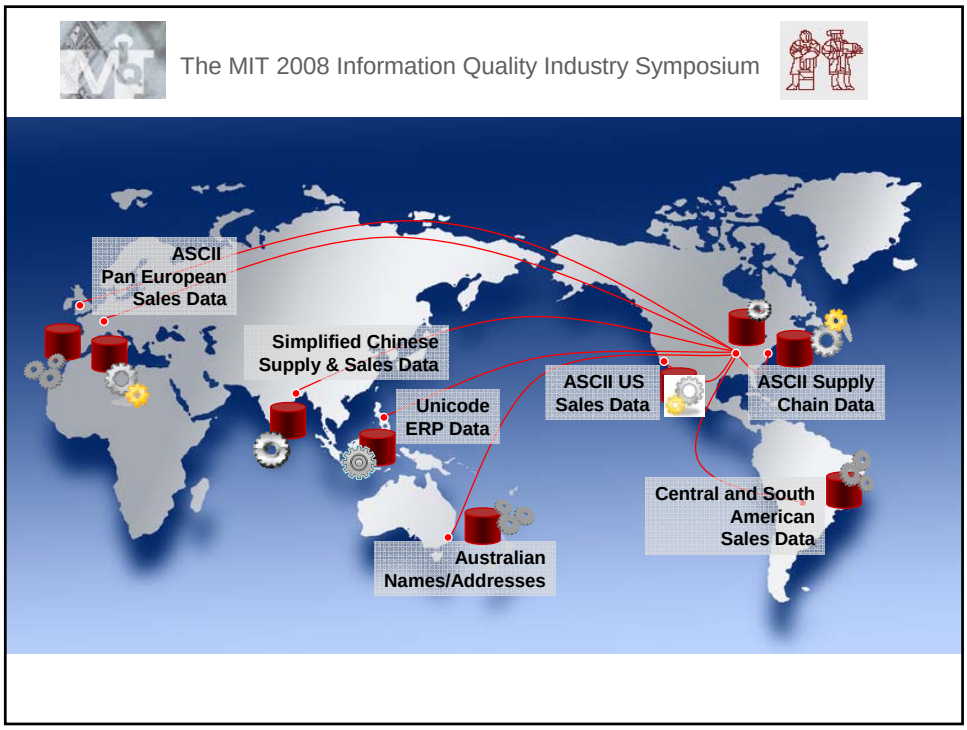

The MIT 2008 Information Quality Industry Symposium
European Office
European Office
Acquisition
Satellite Office

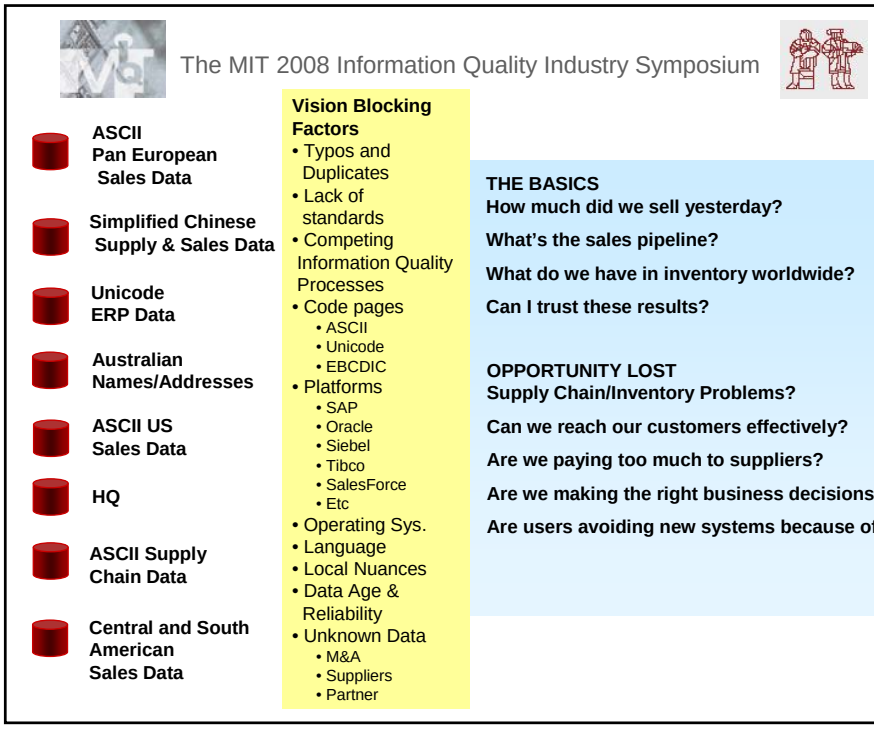

The MIT 2008 Information Quality Industry Symposium
European Office
European Office
Acquisition
Satellite Office
Central & South American Partner


The MIT 2008 Information Quality Industry Symposium
European Office
European Office
APAC Manufacturing
APAC Partners
Acquisition
Satellite Office
Central & South American Partner


The MIT 2008 Information Quality Industry Symposium
European Office
European Office
APAC Manufacturing
APAC Partners
Acquisition
Satellite Office
Central & South American Partner
Australian Customer Support & Sales


The MIT 2008 Information Quality Industry Symposium
ASCII
Pan European
Sales Data
Simplified Chinese
Supply & Sales Data
Unicode
ERP Data
Australian
Names/Addresses
ASCII US
Sales Data
ASCII Supply
Chain Data
Central and South
American
Sales Data


The MIT 2008 Information Quality Industry Symposium
ASCII
Pan European
Sales Data
Simplified Chinese
Supply & Sales Data
Unicode
ERP Data
Australian
Names/Addresses
ASCII US
Sales Data
ASCII Supply
Chain Data
Central and South
American
Sales Data

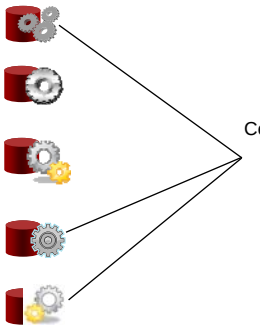

The MIT 2008 Information Quality Industry Symposium
ASCII
Pan European
Sales Data
Simplified Chinese
Supply & Sales Data
Unicode
ERP Data
Australian
Names/Addresses
ASCII US
Sales Data
HQ
ASCII Supply
Chain Data
Central and South
American
Sales Data
Vision Blocking
Factors
• Typos and
Duplicates
• Lack of
standards
• Competing
Information Quality
Processes
• Code pages
• ASCII
• Unicode
• EBCDIC
• Platforms
• SAP
• Oracle
• Siebel
• Tibco
• SalesForce
• Etc
• Operating Sys.
• Language
• Local Nuances
• Data Age &
Reliability
• Unknown Data
• M&A
• Suppliers
• Partner
THE BASICS
How much did we sell yesterday?
What's the sales pipeline?
What do we have in inventory worldwide?
Can I trust these results?
OPPORTUNITY LOST
Supply Chain/Inventory Problems?
Can we reach our customers effectively?
Are we paying too much to suppliers?
Are we making the right business decisions?
Are users avoiding new systems because of data?


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



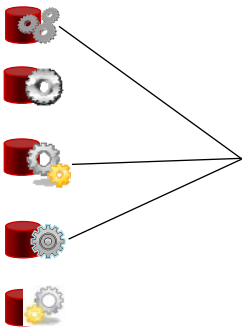
Certain Solutions can't understand this street address.

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

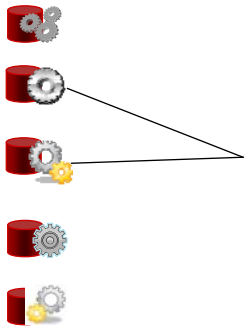


Certain solutions can't handle delivery info intermingled with name and address.

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

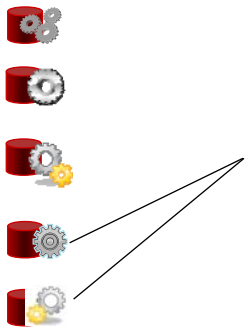


Certain Solutions
Can't handle the C/O

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

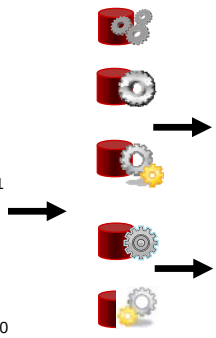
Differences in Data Quality Processes

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

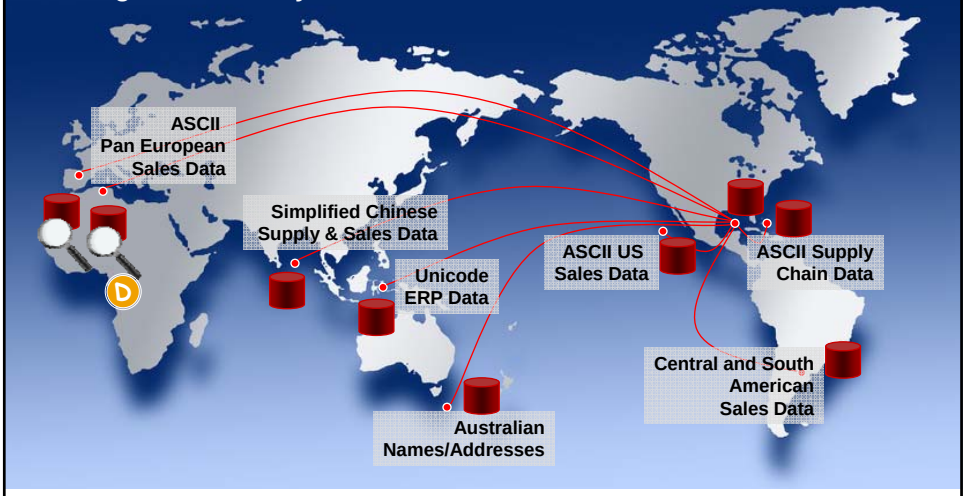
- Real-time cleansing of incoming data

Clean the Lake

- Batch cleansing of existing data

The MIT 2008 Information Quality Industry Symposium

Starting Point: Analysis



Pick High Value Targets and Identify Business Value of Improved Information Quality

The MIT 2008 Information Quality Industry Symposium

Cleansing Is Key to Matching

Original Record 1

Name:	Peggy Smith
Address:	345 6th Ave
City:	NY
State:	NY
Zip:	01012
Country:	

Original Record 2

Name:	Margaret Smith
Address:	345 Avenue of the Americas
City	Manhattan
State:	NY
Zip:	1012
Country:	USA

⇩ ⇩

Standardized Record 1

Root First Name:	Margaret
Last Name:	Smith
Address:	345 Ave of the Americas
City:	New York
State:	NY
Post Code:	10012
Country:	USA

Standardized Record 2

Root First Name:	Margaret
Last Name:	Smith
Address:	345 Ave of the Americas
City:	New York
State:	NY
Post Code:	10012
Country:	USA

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 오포읍양벌리**94-5**대주파크빌**2**차**207**동 **101**호

Level 1	**Block Number**
Level 2	**Sub-Block Num**
Level 3	**Apt. Number**
Level 4	**House Number**
	Postal Code

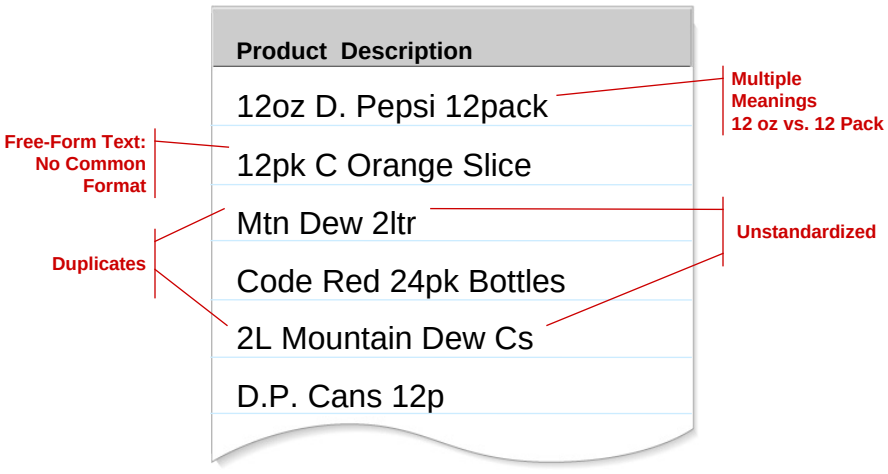
The MIT 2008 Information Quality Industry Symposium

Understanding Global Data (Korean)

Level 1	경기도	Block Number	94
Level 2	광주시	Sub-Block Num	5
Level 3	오포읍	Apt. Number	207
Level 4	양벌리 대주파크빌2차아파트	House Number	101
		Postal Code	464-764

The MIT 2008 Information Quality Industry Symposium

Understanding Product Data



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	**CANS**	12 PACK
12pk C Orange Slice	ORANGE SLICE	**12 OZ**	CANS	12 PACK
Mtn Dew 2ltr	MOUNTAIN DEW	2 L	**BOTTLES**	**8 CASE**
Code Red 24pk Bottles	CODE RED	**20 OZ**	BOTTLES	24 PACK
2L Mountain Dew Cs	MOUNTAIN DEW	2 L	BOTTLES	**8 CASE**
D.P. Cans 12p	DIET PEPSI	**12 OZ**	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 (978) 436-8768