Data Quality and Master Data Management

David Loshin
Knowledge Integrity, Inc.
loshin@knowledge-integrity.com
(301) 754-6350

1

Agenda

1) Master Data Management and Data Quality
2) Data Profiling and Data Standards for Master Data Analysis
3) Data Quality Tools and Technology
4) Data Governance

Part 1:

Master Data Management

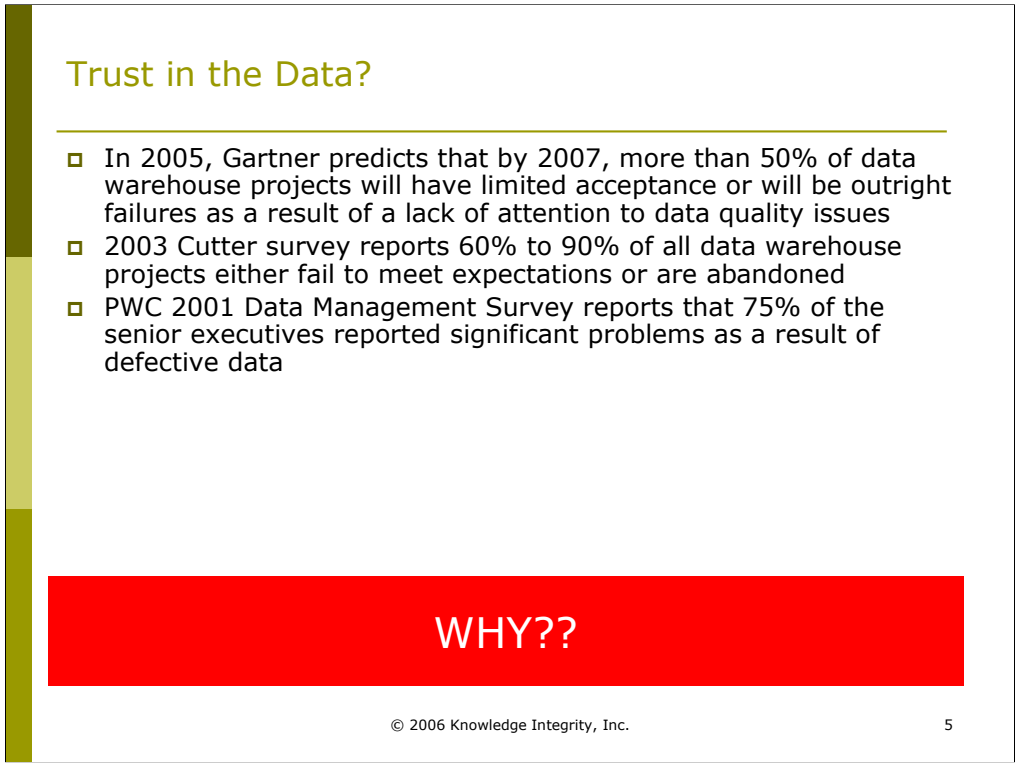
and Data Quality

3

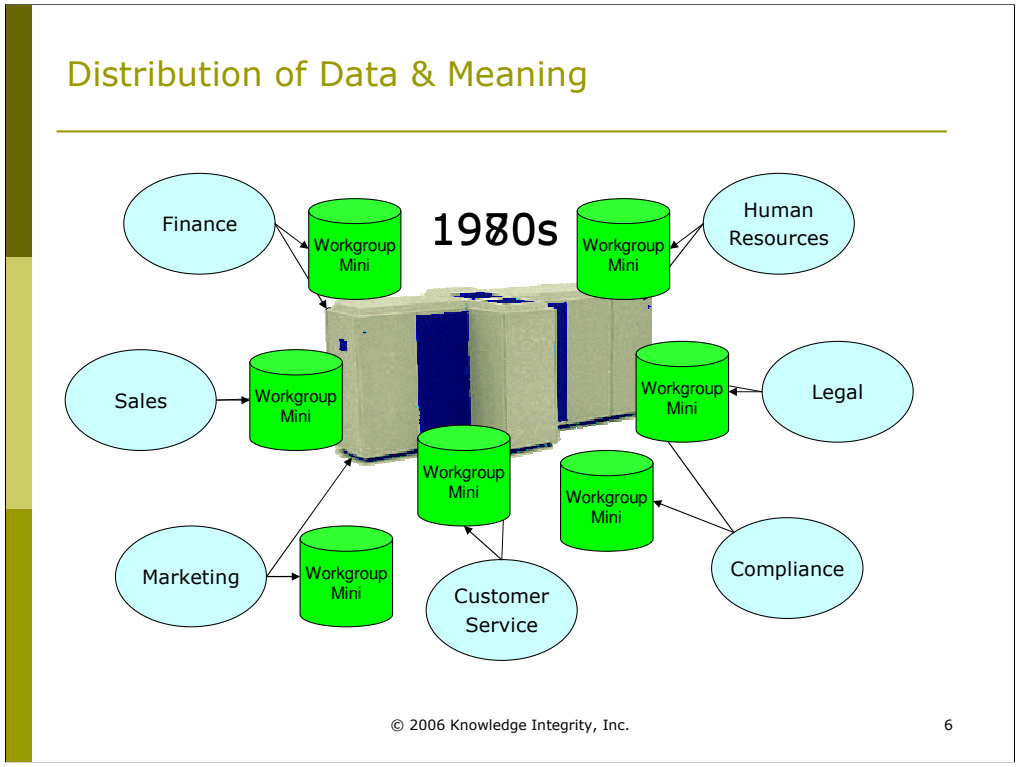
CRM: Holy Grail or Xeno's Paradox?

- CRM - Customer Relationship Management
  - Help a company manage relationship with customer
  - Help a company understand the roles and relationships of parties within their business environment
  - Help a company understand how customers relate to each other
- Is the 360° view a myth?
- Critical Questions about Missed Expectations:
  - What are the business objectives of CRM?
  - Why has traditional CRM not lived up to its promise?

Customer Relationship Management is often seen as not having lived up to its promise. Although the intentions were appropriate, the failure lies in the focus on the application infrastructure and the information model, and not on the information content. The problem is that when disparate systems are expected to source the population of a CRM system, that CRM system itself becomes yet another database that no longer stays synchronized with the other data sets across the organization.



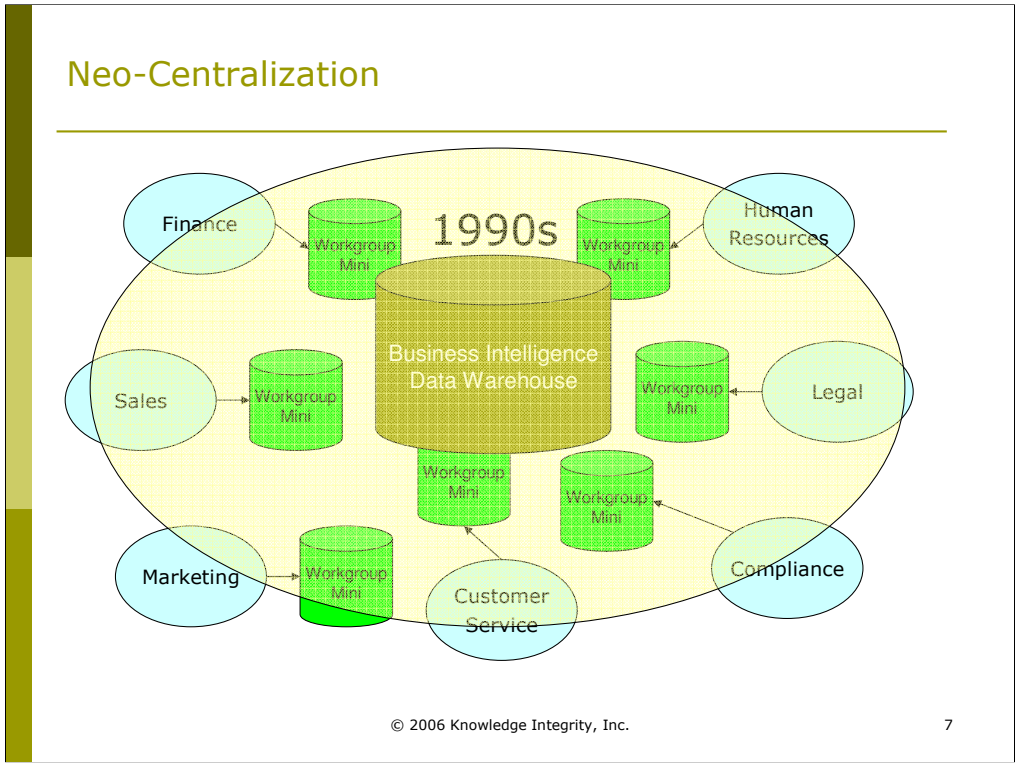
What is curious about these quotes and predictions is not the focus on failure of analytical (or even operational) systems due to poor data quality. The strange thing is that year after year, the analysts continue to beat the drum about poor data quality, but the message has trouble reaching the intended audience. If organizations focused on improving data quality, then one might consider that the prediction would be that fewer data warehouse projects will fail!



What is the origin of master data?

In the 1970s, all data systems were based on a single system that was time-shared across different applications. With limits on memory and disk space, collaborative efforts to define the exact sizes of data values for data files provided some level of oversight for data management.

With the introduction of workgroup computing, many different subgroups within an organization acquired their own computers along with their own application and data management teams. As these teams developed new relational models, their approaches, while similar, may have had slight variations in format, sizes, and data types.



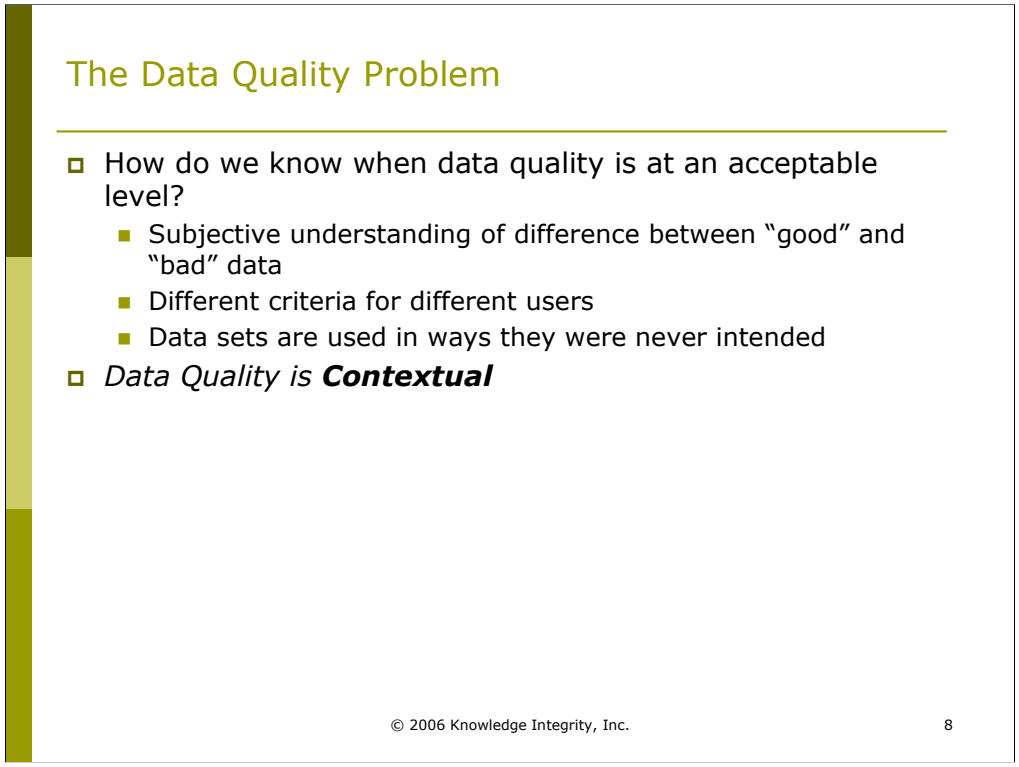
However, by the 1990s, the desire to aggregate enterprise data into decision support systems, and ultimately into data warehouses that service multiple business intelligence, reporting, and analytic functions, created the need to recentralize data.

The challenges of data integration overwhelmed the data warehouse process, with estimates of 70-80% of the time spent in implementing a data warehouse dedicated to addressing data extraction, transformation, and loading. This effort underlines the key issues that drive the need for a Master Data management strategy:

-The need to govern the way that information is shared between operational applications and the analytical functions

-The absence of governance and accountability for the quality of shared data

-The ability to communicate opportunities for information value improvement back upstream



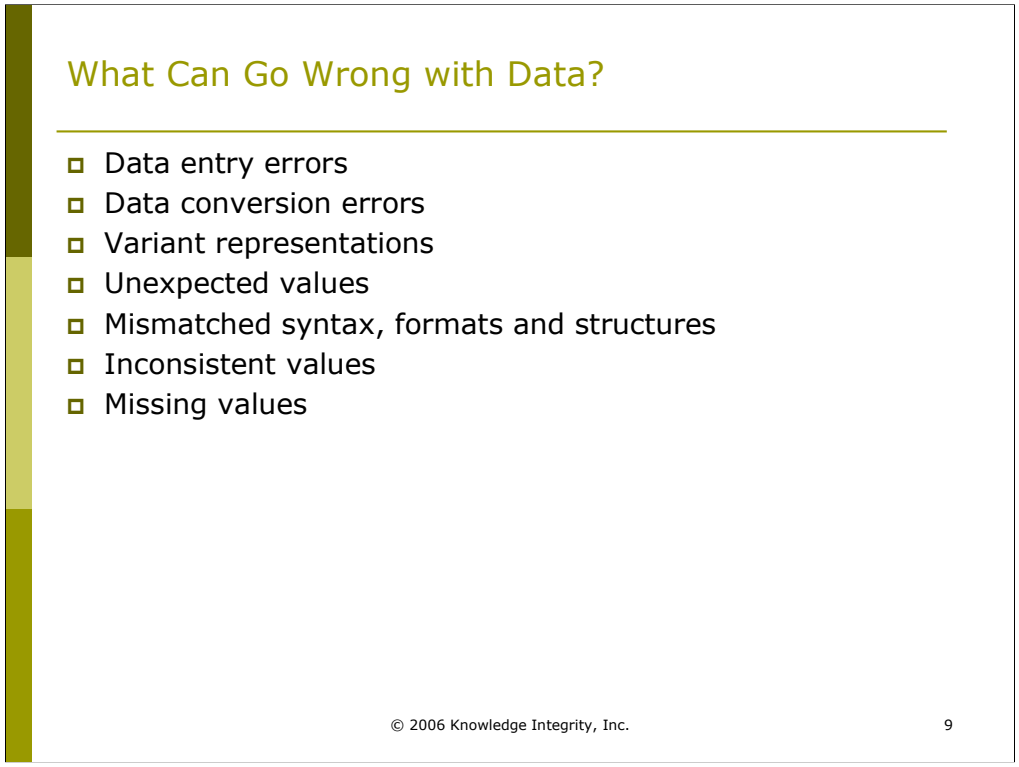
A lot of the data quality literature discusses the merits of measuring levels of data quality, but how do we do this? One user's view of a "good" piece of data may be completely irrelevant in the eyes of another user.

One of the biggest hurdles is the lack of consensus of what constitutes a good or bad value. In addition, it is possible that the same data items, used in different execution contexts, may have completely different meanings! Especially in matrixed organizations that are attempting to execute Business Intelligence projects, the cross-pollination of information across the organization leads to more complex data usage constraints.

Data sets in an enterprise evolve in a distributed manner – the accounts system, the order management system, the products system, etc. We can apply a "Peter Principle" to data in that is rises to its necessary level of competence, but even though different data systems are merged or linked, there is no internal motivation on behalf of the data owners to improve the level of quality higher than their own needs.

Note that "Data Quality" is not the same thing as "Data Cleansing." Data cleansing is a static operation applied to specific data items, potentially on a repetitive basis. For example, data sets extracted and delivered to a staging area for periodic data warehouse refreshes are then subjected to a cleansing operation, but if the same data values are being extracted for each refresh, the same cleansing operations will be applied each time the data is extracted.

Data Quality is a process and methodology for identifying the source of a continual problem and then fixing the problem at the source to prevent further occurrences. But in order to measure quality in an objective way we must first decide what kinds of metrics to use.



What kinds of problems crop up in data sets? If we are dealing with pristine data that was created and entered into a relational system in 3rd-normal form, we might assume we are in good shape. Unfortunately, very little data appears in this "pure" form. Most data sets emanate from actual people typing data into disparate and distributed entry systems, which have little or no validation checking at entry. In this slide we look at typical causes for the introduction of noncompliant data.

In addition, there is a world of information that lives in systems running before the advent of relational database systems, in which case there is no guarantee of null constraints or referential integrity. In addition, a lot of data exists in alternate forms existing outside the realm of database systems, such as flat file data, web data, Excel spreadsheets, etc. Because of this disparity, the potential for problems is multiplied. Any of the problems enumerated here can appear in almost any collection of data sets.

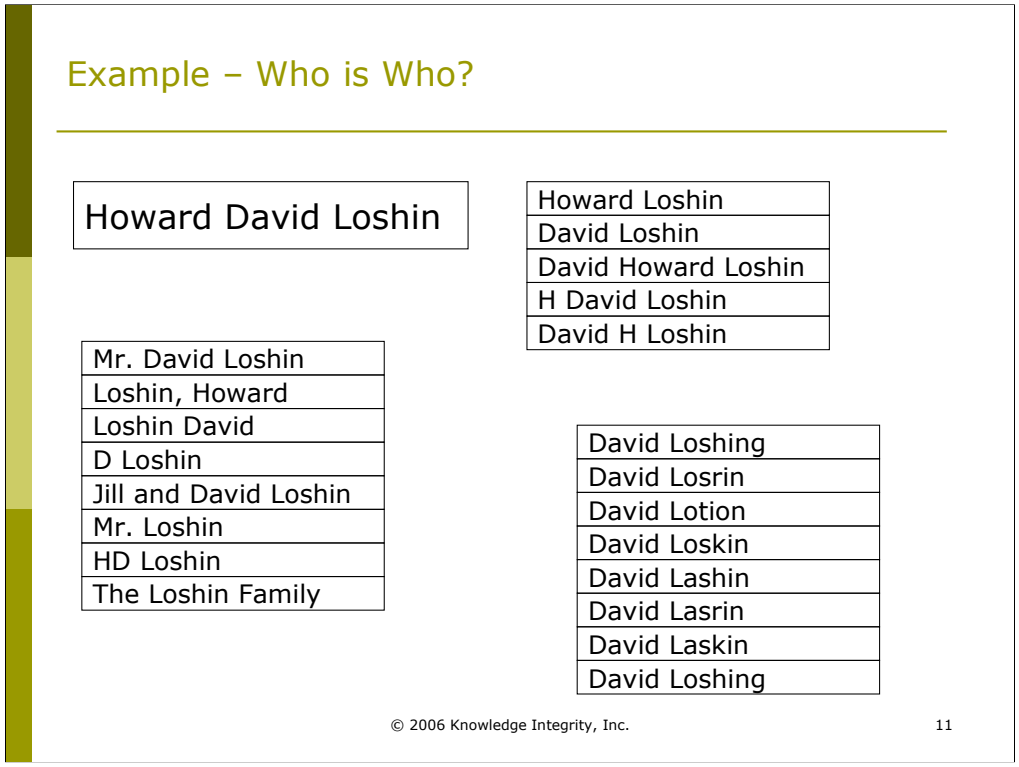
Data entry errors are very typical, either from lack of focus by those tasked with entering data, or perhaps due to improper construction of the input system. For example, one of our clients found that although a certain table's column was no longer used, the input screen still held a place for entering the data. **Conversion errors** often are introduced when migrating from a legacy system to a new one, when extracting data during an ETL process, or when loading data into a target system. Depending on the kinds of constraints or controls one can exercise over the information flow process, there are different likelihoods that unexpected values are introduced. **Mismatched formats** occur with data that we expect to see conforming to a set of patterns, like telephone numbers or addresses.

Inconsistencies are sometimes deliberately introduced (for example, in order to force transactions to be performed, such as entering a "dummy" telephone number at the point of sale), or to insert **incomplete** (or to allow for **missing**) data (as an example, in one financial services environment, one could enter a security trade for a customer even if the customer had not yet been entered into the system).

Example: Common Name Variations

Variation Type	Examples
Nicknames	William, Bill, Billy, Will
Name Variation	Chris, Kris, Christie, Krissy, Christy, Christine, Tina
Abbreviations/Spelling	Mohammed, Mohd, Mohamad, Mhd, Muhammad
Foreign Versions	Peter, Pete, Pietro, Piere, Pierre
Spelling Variation	Johnson, Johnsen, Johnsson, Johnston, Johnstone, Jonson
Suffix Variation	Smith II, Smith 11, Smith Jr, Smith jnr
Initials/Order	Frank Lee Adam; A. Frank Lee; Lee Frank
Anglcization	De La Grande, Delagrande, D L Grande
Out of Order	Henry Tun Lye Aun; Mr Aun Tun Lye (Henry)
Titles	Dr. Henry Lee, Henry Lee, M.D., Mr. Henry Lee

Although not all data errors occur in names, the desire to deploy customer-oriented applications is a common theme in the MDM world. Therefore, reviewing the kinds of errors and variations that occur in names is a good exercise in evaluating some of the data quality issues that will probably need to be addressed, especially when attempting to consolidate multiple records into unique identities.



This slide shows numerous variations of name strings that are intended to represent one individual. Exact matching will lead to missed matches (also referred to as false negatives).

Note the following aspects of the different kinds of variance:

-Name component ordering (first name, middle name, last name subject to float across the field)

-Optional use of punctuation

-Multiple entities within a single field (married couples, “family”)

-Misspellings

-Phonetic transformation (and reverse as well – “Loshin” spelled as “Lotion”)

-Hand-transcription errors (switching “r” for “h,” or “k” for “h”)

-Hearing errors (changing “-in” to “-ing”)

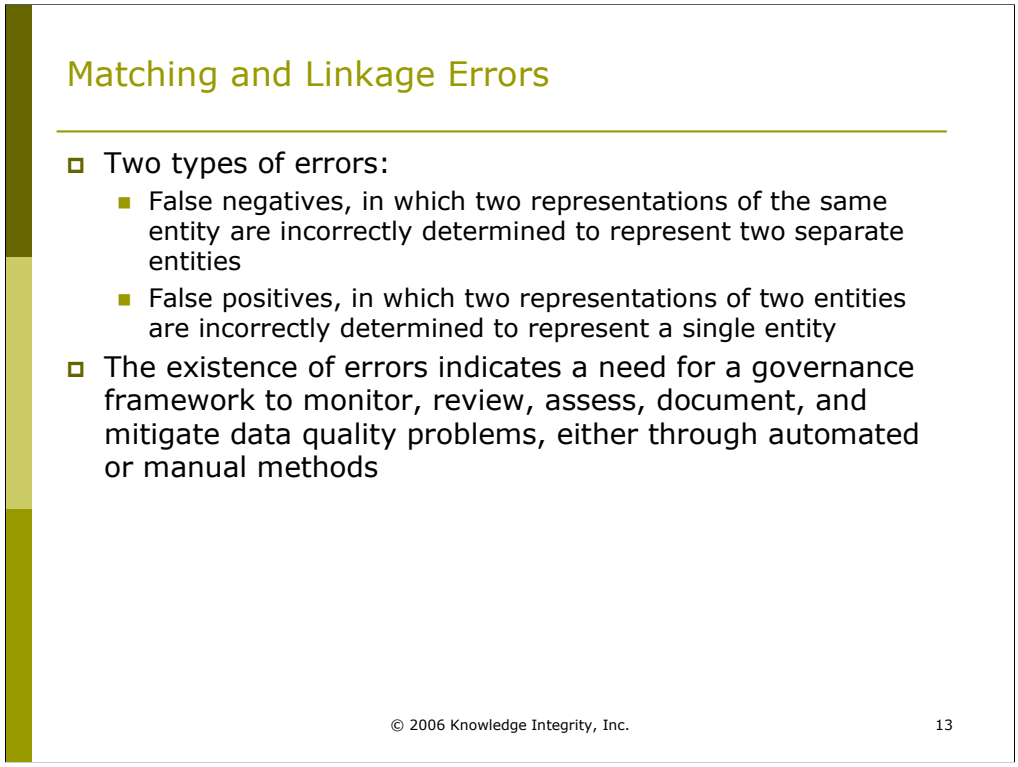
No, Really – *Who is Who?*

Enterprise Knowledge Management : The Data Quality Approach (The Morgan Kaufmann Series in Data Management Systems) (**Paperback**) by David Loshin — Me

Business Intelligence: The Savvy Manager's Guide (The Savvy Manager's Guides) (**Paperback** - June 2003) by David Loshin — Me

The Geometrical Optics Workbook(**Paperback** - June 1991) by David S Loshin — Not Me

Here is a good example of a false positive consolidation. Not only do these two authors share the same given and relatively uncommon last name, they also share demographic attribution of being authors. This demonstrates the fact that even factoring in statistical probability associated with the name frequencies, assigning weights based on different attribute values can still lead to falsely unifying two distinct entities.

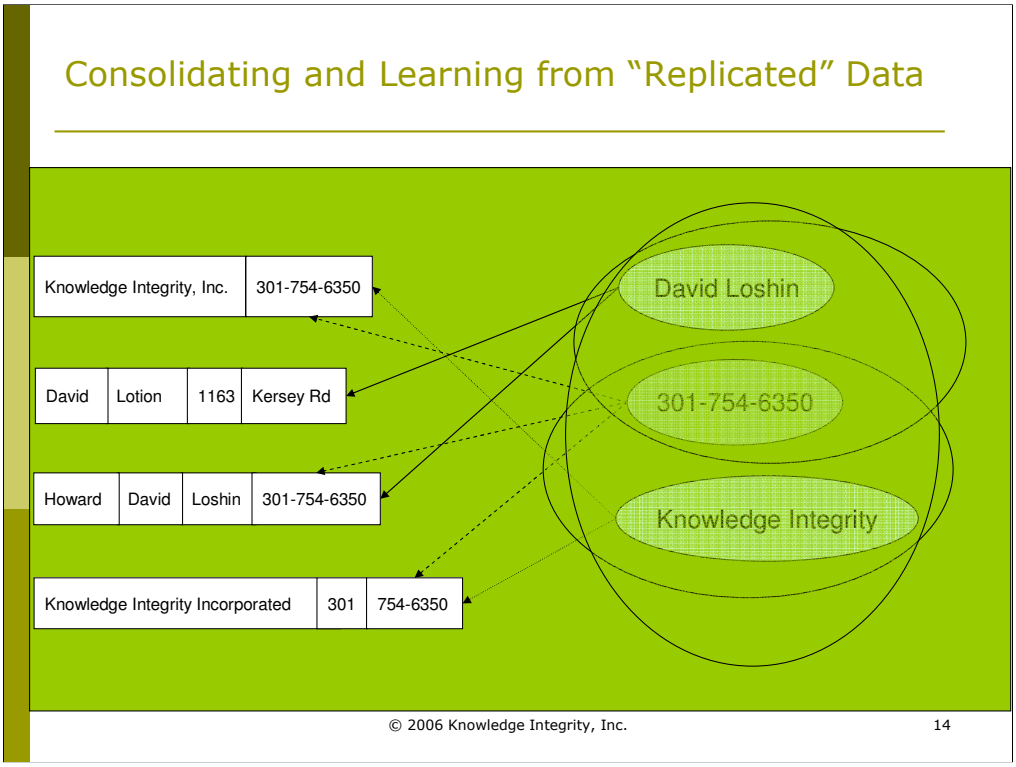


The potential for both kinds of errors introduces conceptual questions of data quality governance. The issue here is that while automated tools can be used to cluster data instances for the purpose of unique identity resolution or consolidation, it is nearly impossible to automate the determination that false positives or negatives exist.

Yet we may establish data quality expectations regarding proper data aggregation, which is measured based on the number of these kinds of errors.

This suggests that data quality management and data governance requires manual processes for review and assessment. These processes must have a reasonable foundation for believability. For example, a manual review might take into account a random selection of records, with the sample size determined based on the size of the original data set and the tolerance in the level of confidence of the result (i.e., margin of error in evaluation).

It also suggests that protocols be defined to mitigate the risks introduced by these kinds of errors. How are incorrect linkages broken, or missed ones accounted for, how is that information is communicated to participants and contributors, etc.



Another valuable by-product of the data integration process is the ability to learn new information about the relationships between entities that exist within your data sets. In this example, there are no records that contain a specific link between the individual "David Loshin" and the company "Knowledge Integrity." The ability to cluster records based on content, though, establishes a transitive connection:

David Loshin is associated with the telephone number 301-754-6350

The telephone number 301-754-6350 is associated with Knowledge Integrity

Therefore: David Loshin is associated with the company Knowledge Integrity

Not only that, establishing the connectivity between this set of records provides an address for Knowledge Integrity. By applying a data integration process, we are also deploying a knowledge discovery process as well!

The Problem Gets Worse

- Exploitation of "shortcuts" for implementation and operations
- Numerous customer touch points with their own front end interfaces feed the same back end
- Limited system functionality driving "creativity"
- *Changes in use and perception of data*

When systems are initially designed, the requirements are carefully documented. However, as time progresses, changed are made to the system even though the documentation becomes woefully out of date. In some instances, implementation shortcuts are routinely taken, such as using the same

In our experience, we have also seen problems introduced when **multiple front end interfaces** feed the same back end system. In particular, the same information is inserted by different processes, and there are no checks to resolve potential duplication.

The End Result

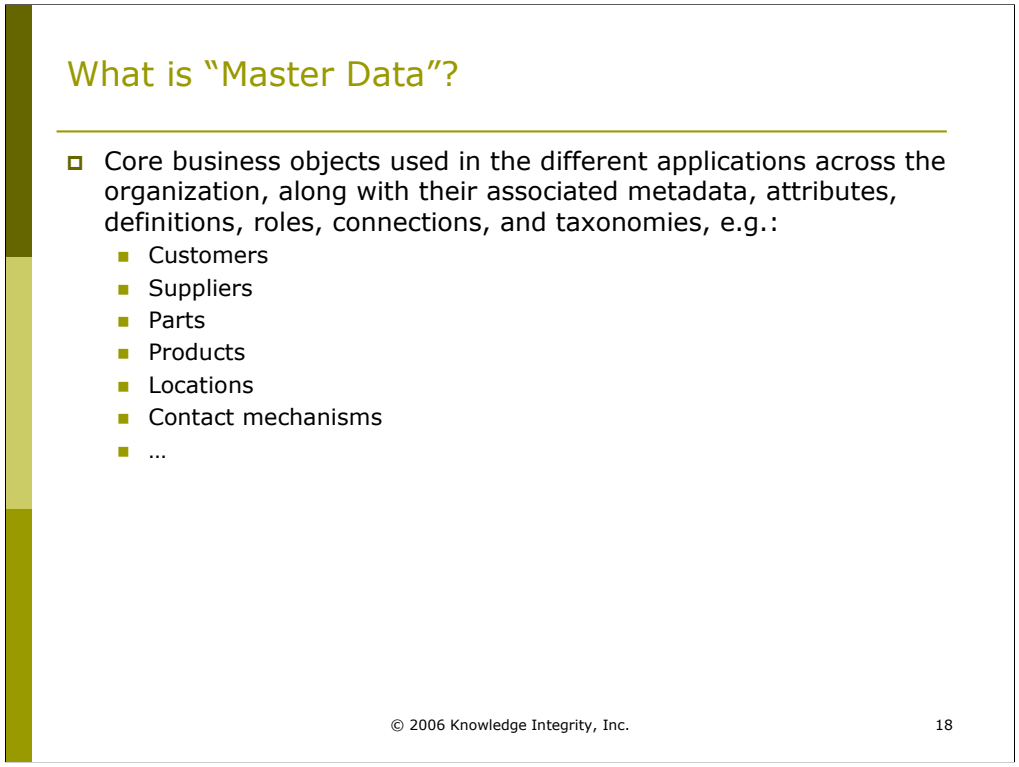
- Inability to effectively identify unique "entities" within the distributed environment
- Inability to accurately quantify or qualify the roles that individuals play in the business context
- Inability to establish relationships across customer space

No wonder CRM has not lived up to its promise

Master Data Management

- New idea:
  - Resurrect technical approaches to data integration
  - Make the information sharing framework b-directional
  - Institute governance to manage technology, automation, and behavioral gaps

➡ Master Data Management (MDM)



Master data objects are those "things" that we care about – the things that are logged in our transaction systems, measured and reported on in our reporting systems, and analyzed in our analytical systems. Common examples of master data include:

Customers

Suppliers

Parts

Products

Locations

Contact mechanisms

Policies

Transactions Use Master Data

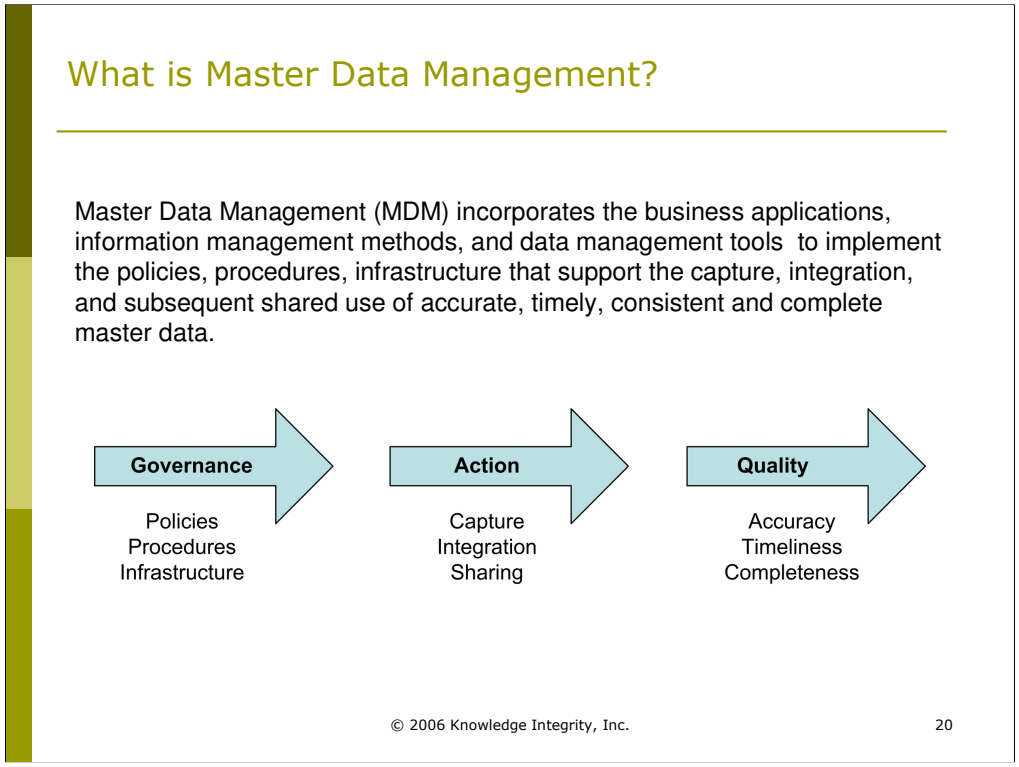
"David Loshin purchased seat 15B on US Airways flight 238 from Baltimore (BWI) to San Francisco (SFO) on July 20, 2006."

Master Data Object	Value
Customer	David Loshin
Product	Seat 15B
Flight	238
Location	BWI
Location	SFO

The transaction demonstrates the use by reference to the master data objects. The fact that the same objects may appear in other applications, whether they be functional or analytical, lends credence to their status as master objects.

In this example, we see four kinds of master data objects, which can be characterized based on the business application: customer, location, flight, and product.

The fact that we have identified "flight" as a master data category should trigger some thought – what is the difference between a "flight" and a "product"? Does one category fit into the other, and what are the relationships between them? This exposes one major issue associated with MDM data governance: standardizing the definitions and semantics associated with master data objects.



An MDM program is intended to:

- Assess the use of core information objects, data value domains and business rules
- Identify core information objects relevant to business success
- Instantiate a shared standardized object model
- Manage collected and discovered metadata as an accessible, browsable resource
- Collect and harmonize replicated data instances
- Integrate the harmonized view via a service-oriented approach
- Institute the proper data governance policies and procedures at the corporate or organizational level

CDI – One kind of Master Data Management

- According to Gartner:

"Customer data integration (CDI) is the combination of the technology, processes and services needed to create and maintain an accurate, timely and complete view of the customer across multiple channels, business lines and, enterprises, where there are multiple sources of customer data in multiple application systems and databases"

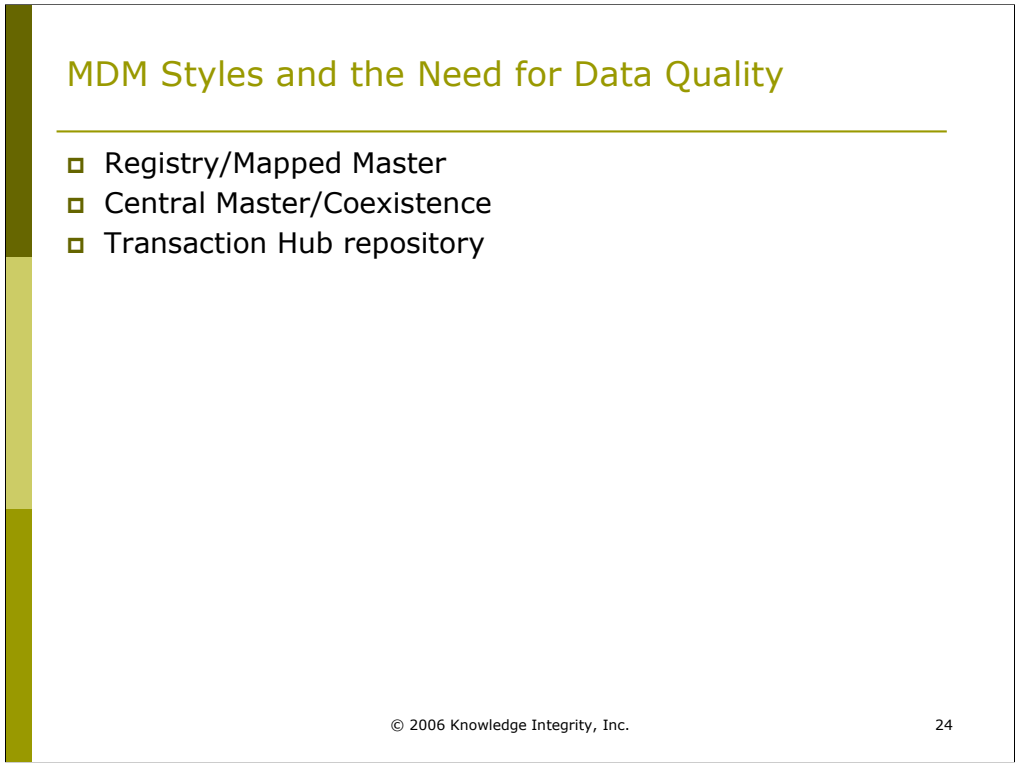
MDM Goals

- Organizational
  - Provide a unified view of each customer derived from information available across the enterprise
  - Integrate the unified view back into the applications
- Business
  - Improve cross/up-sells
  - Improve marketing program efficiency
  - Enhance customer experience
  - Reduce risk

Business Objectives for MDM

This Business Objective	*Really Means…*
"accurate, timely and complete view of the customer"	Enforceable and measurable data quality
"across multiple channels, business lines, and enterprises,"	Matrixed collaboration between business managers across vertical lines of business
"multiple sources of customer data in multiple application systems and databases"	Complete and efficient data integration and aggregation from heterogeneous data sources in a coordinated manner

The business objectives effectively translate into requirements associated with the underlying quality of the data to be integrated, the technical processes employed, and the need for oversight and governance.

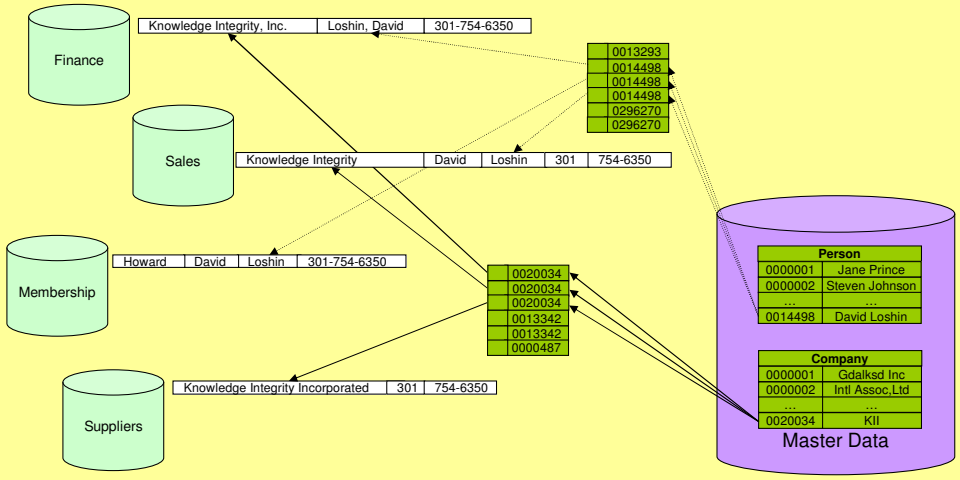


"System of Record" acts as an external reference master registry.

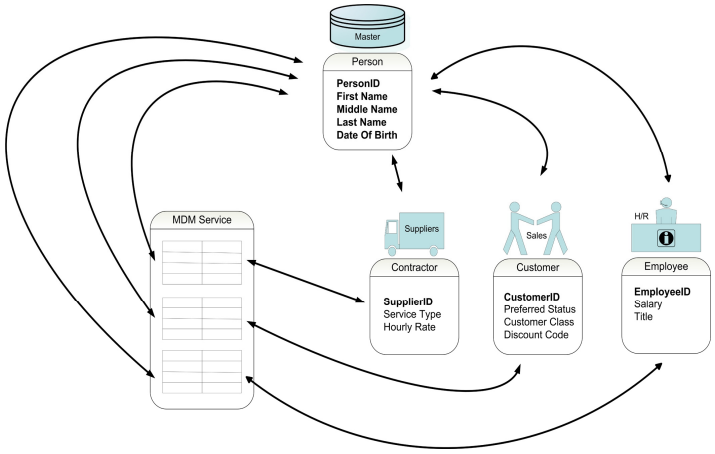
Alternate options:

- •Registry (maintain a registry of identities with links to systems that hold instances)
- •Coexistence (source of truth used to update and harmonize existing system data)
- •Integrated Transaction hub (replaces existing system use of customer data altogether)

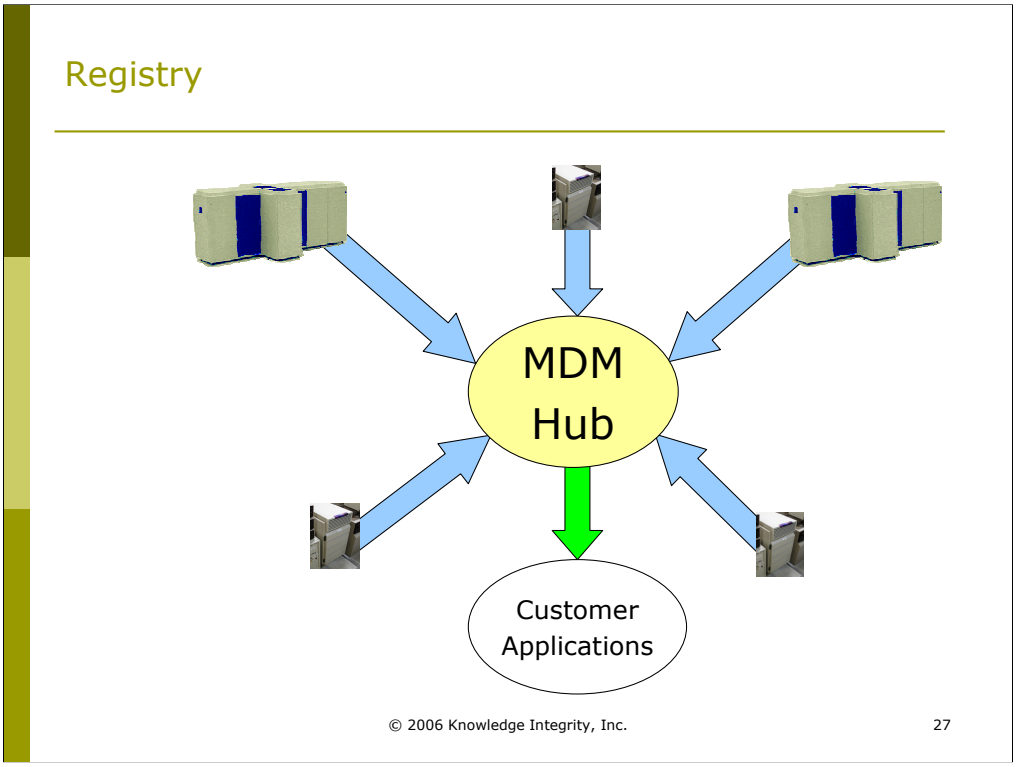
Master Customer Registry



Registry



In a **Mapped Master System**, an existing application system is selected to be the master, and other systems become dependent on that system as the main repository. New data instances are created in the master, which are then propagated to other systems. In this approach, different data objects are not necessarily linked by a global primary key, so it may be necessary to define mappings to link objects between systems. When new objects are created, they are distributed out to other applications, and similar to the central master approach, the other applications may not modify core attribute values but may introduce and modify their own application-specific attributes.

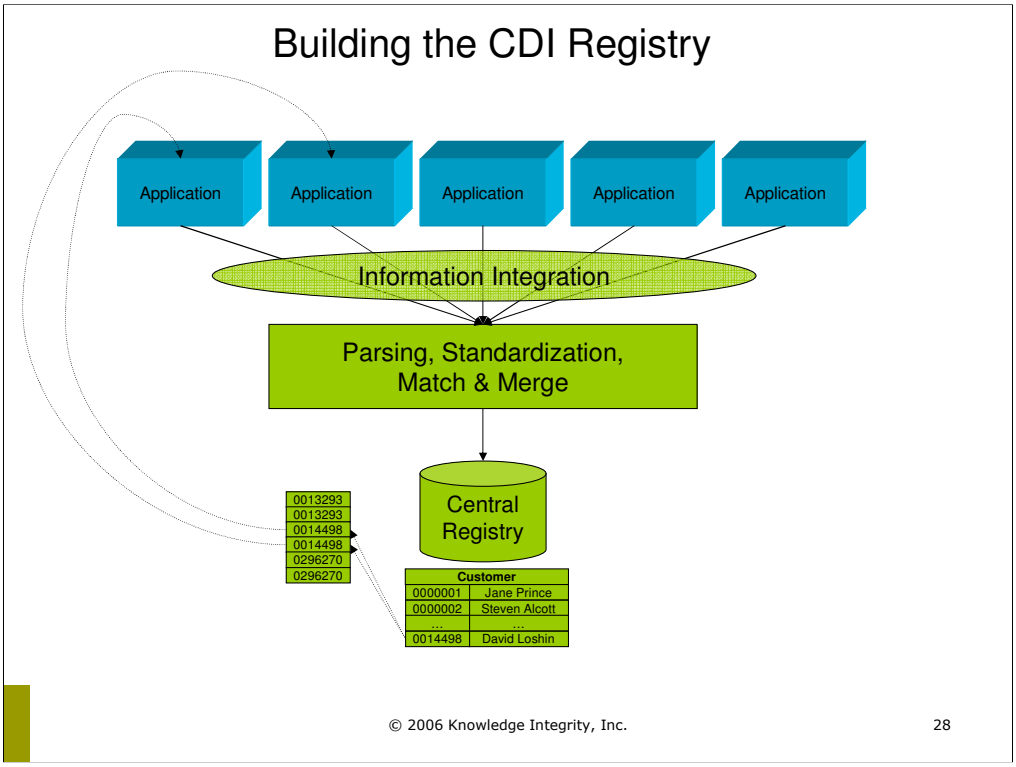


•A master registry maintains a unique reference for each identified entity

• Rules for transformation are managed within the CDI hub

• For customer-related applications, the master registry is consulted at run time and records are materialized in real time

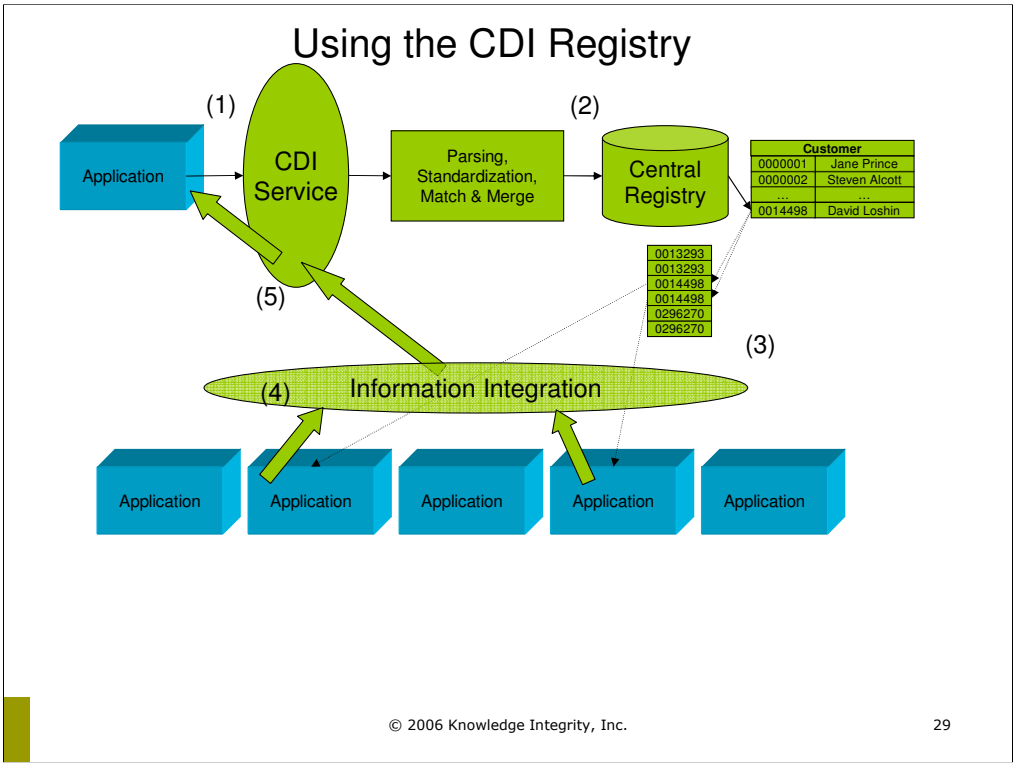
• Legacy systems are not affected



•The CDI Registry is initially populated with data extracted from source systems.

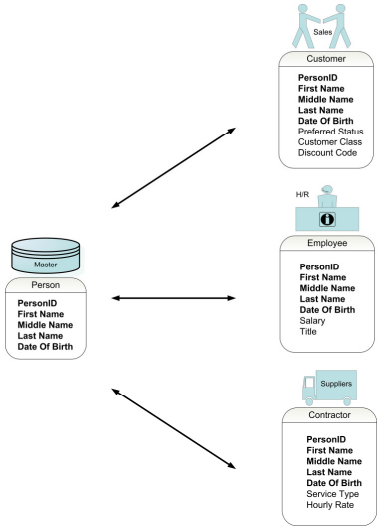
•The data is subjected to parsing, standardization, cleansing, and then to an Identity resolution/match & merge process.

•The result of the identity resolution process is a set of cross-references back to data sets in original source systems, both representing the party and its associated relationships.



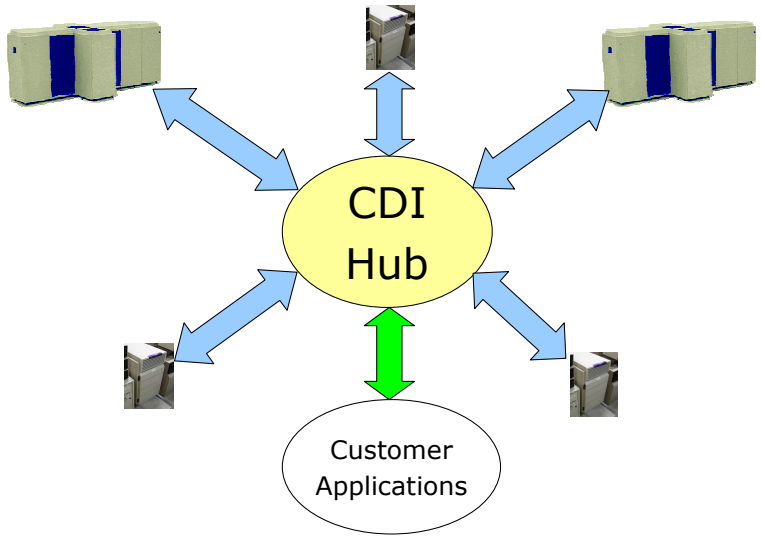
1. An application wants to create a new customer record. It passes the new data to the CDI service, which passes it through to the Central Registry
2. The customer demographic data is searched in the cross-reference, and two references are found in other applications.
3. The Information Integration service requests an extract of the customer data from the applications, and
4. When the data is accessed, it is transformed into a packet for return to the CDI service requestor
5. The packaged customer view is returned to the original requesting applications

Central Master/Coexistence

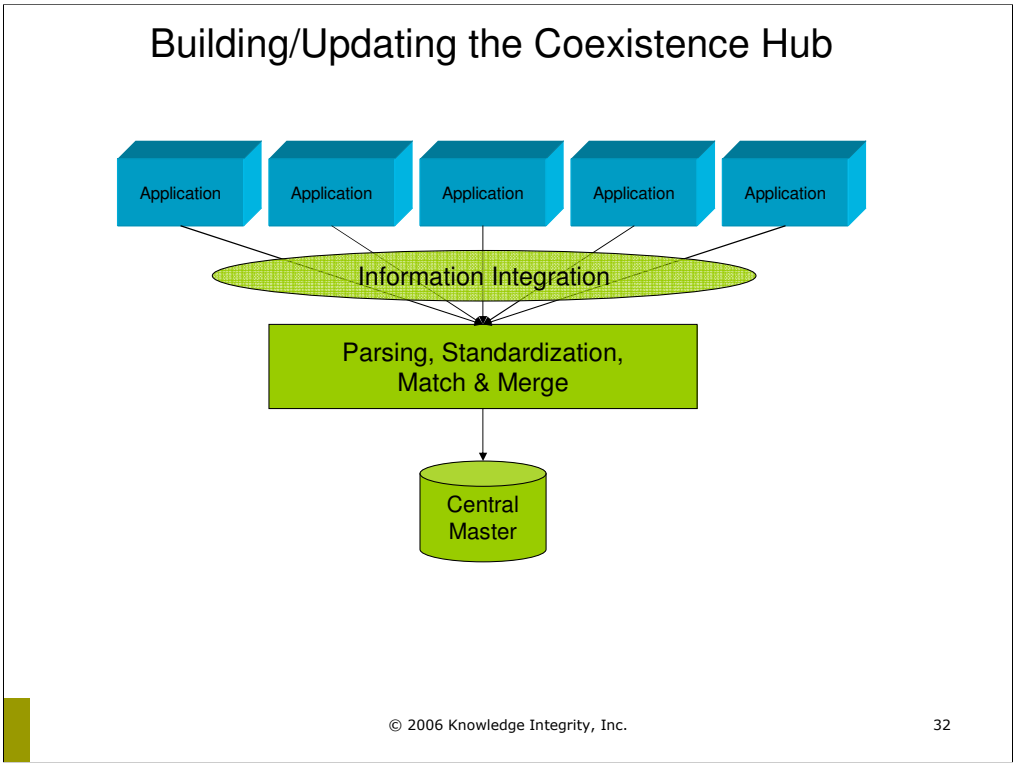


In a **Central Master Data System**, for each data "domain," a set of core attributes associated with each master data model is defined and managed within a single master system. The master repository is the source for managing these core master data objects, which are subsequently published out to the application systems. Within each dependent system, application-specific attributes are managed locally, but are linked back to the master instance via a shared global primary key. In this approach, new data instances may be created in each application, but those newly-created records must be synchronized with central system.

Coexistence



- A master repository maintains a unique record for each identified entity
- Rules for transformation are managed within the CDI hub
- For customer-related applications, the master repository is consulted at run time
- When new customer data is introduced, the master repository is consulted and the information is “harmonized” back into the production systems

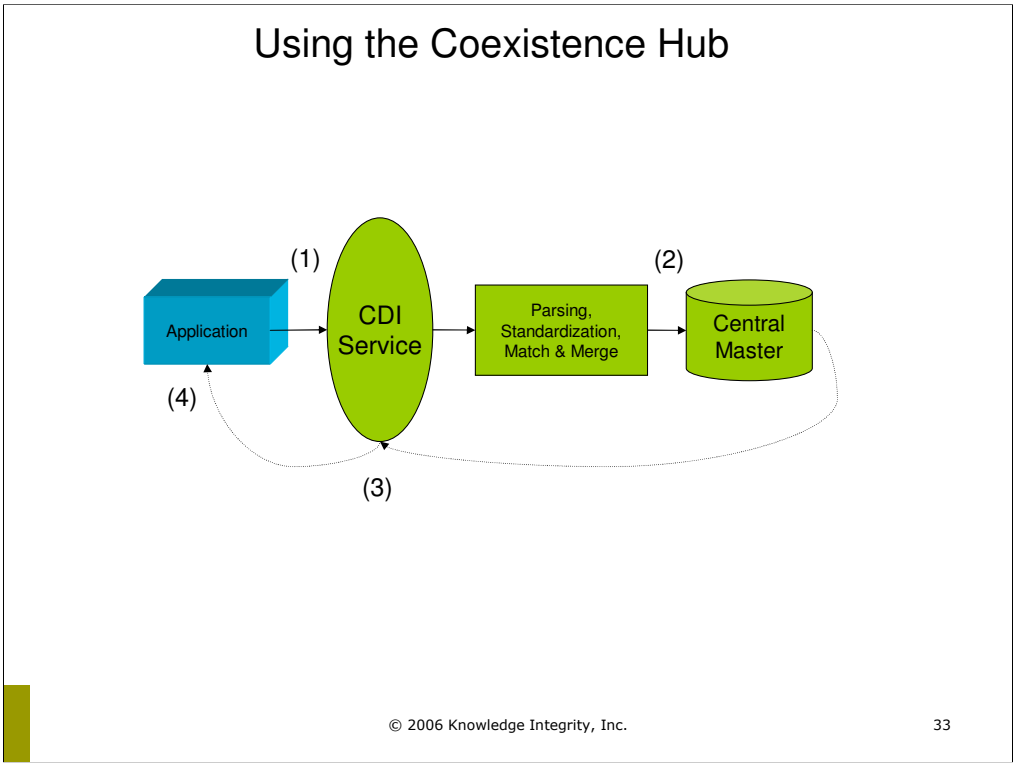


•The Central Master is initially populated with data extracted from source systems.

•The data is subjected to cleansing, and then to an Identity resolution/match & merge process.

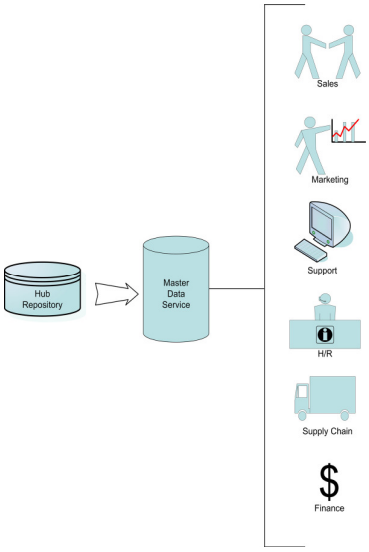
•The result of the identity resolution process is a central data repository of unique entities, with a core set of resolved, “best record” attributes from the participating applications.

•The Central Master is updated in batch on a periodic basis.



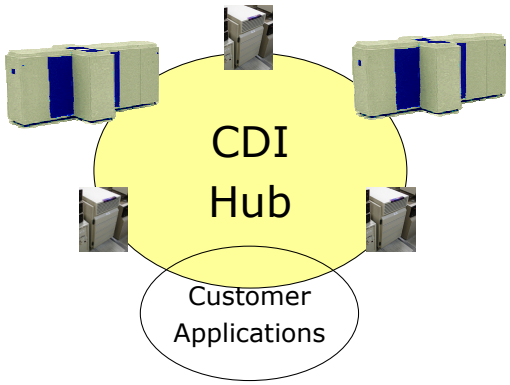
1. An application wants to create a new customer record. It passes the new data to the CDI service, which passes it through to the Central Master
2. The customer demographic data is searched in the Central Master and the corresponding “best record” customer data is passed back to the CDI service
3. The CDI Service forwards the data back to the requesting application.
4. The requesting application can decide to use the forwarded data if it is a match; if not, the application may choose to create a new customer record, which will be synchronized with the Central master the next time it is updated.

Transaction Hub Repository



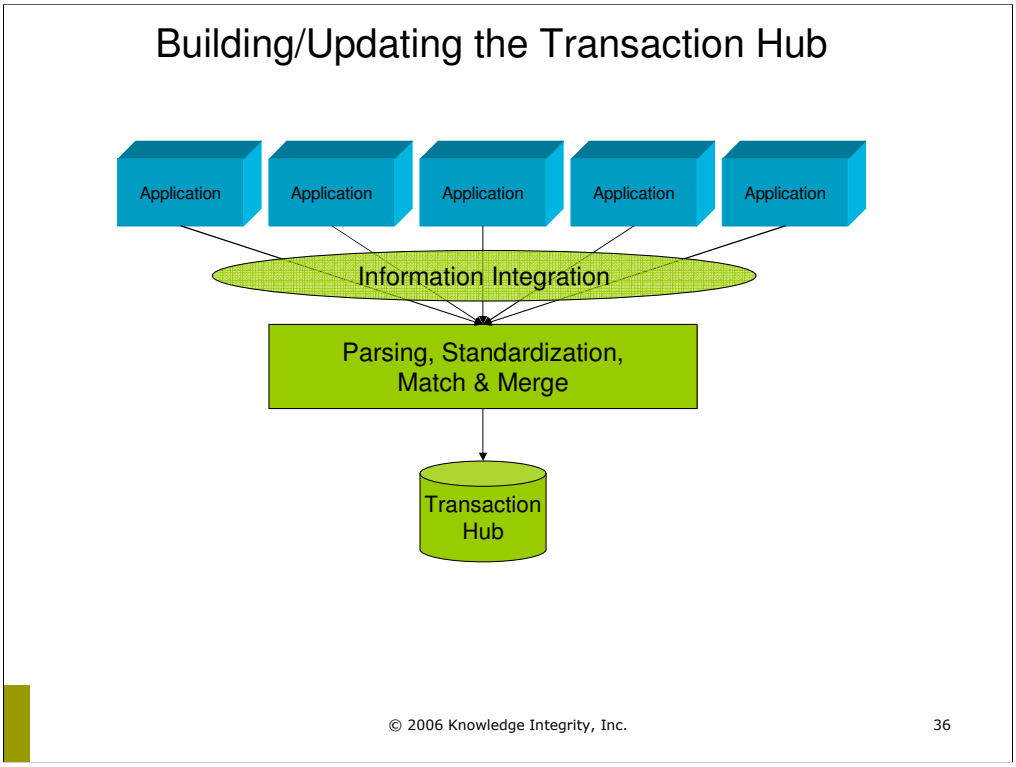
In a **Hub Repository,** a single repository is used to manage the core master system, and data is not replicated to other systems. Applications request information from central hub and provide updates to the central hub. Since there is only one copy, all applications are modified to interact directly with the hub.

Transaction Hub



•A master repository maintains a unique reference for each identified entity

•Through service-oriented architecture, the CDI repository supplants the customer components of the existing applications, providing seamless integration of a unified customer view

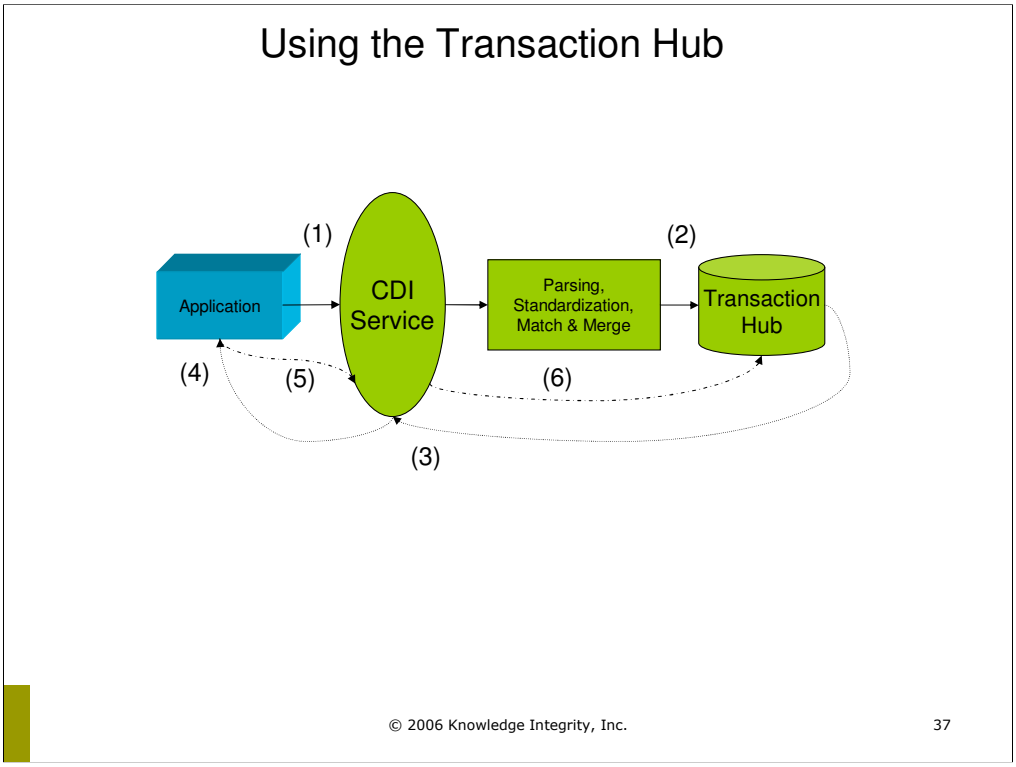


•The Transaction Hub is initially populated with data extracted from source systems.

•The data is subjected to cleansing, and then to an Identity resolution/match & merge process.

•The result of the identity resolution process is a central data repository of unique entities, with a core set of resolved, full set of attributes from all the participating applications.

•The Transaction Hub ultimately replaces application-specific data sets.



1. An application wants to create a new customer record. It passes the new data to the CDI service, which passes it through to the Transaction Hub
2. The customer demographic data is searched in the Transaction Hub and the corresponding customer data is passed back to the CDI service
3. The CDI Service forwards the data back to the requesting application.
4. The requesting application may opt to create a new record or update a returned record; however,
5. The created or updated record is forwarded back to the CDI service, which then
6. Updates the single entry in the Transaction Hub.

Assessment: The ability to identify core data objects that should be incorporated into the master data repository depends on a structured process of assessment that relies on automated tools for analysis. Specifically, data profiling must be employed for empirically identifying reference data sets, primary keys, foreign keys, implicit relational structure and embedded business rules before any data integration can begin.

Integration: The nature of variation associated with master data (e.g., personal names, addresses, telephone numbers, product descriptions) demands that tools be used to help in resolving the variation in representation of specific entities from disparate data sources. Standardization, parsing and matching/linkage techniques have been available as part of any data cleansing tool kit, and the value of using those *technologies* in support of new *methods* is abundantly clear.

Assurance: MDM is not going to be a case of "build it and they will come." Organizational stakeholders will be willing to participate in the integration and consolidation process only as long as they are able to benefit from the process, implying the need for a high degree of confidence in the high quality of master data moving forward. Auditing and monitoring compliance with defined data quality expectations, coupled with effective issue response and tracking – along with strong data stewardship within a consensus-based governance model – will ensure ongoing compliance with application quality objectives.

Questions?

Part 2:
Data Profiling and Data Standards for Master Data Analysis

Review – What is Master Data?

- Core business objects used in the different applications across the organization, along with their associated metadata, attributes, definitions, roles, connections, and taxonomies, e.g.:
  - Customers
  - Suppliers
  - Parts
  - Products
  - Locations
  - Contact mechanisms

Characteristics of Master Data

- Contains reference information
- Referenced in both transaction and analytic system records
- May require specialized application functions to create new instances
- Likely to have models reflected across multiple applications
- May be embedded in legacy data structure models
- Any others?...

Master Data Discovery and Data Profiling

- Objective: analyze data sets to collect useful characterizations that help analysts determine "master data" status
- Use data quality tools to understand underlying metadata
- Data Profiling can be used to:
  - Assess column characteristics
  - Identify reference data objects
  - Determine existence of reference identifiers
- Verify existence of expected master objects
- Inventory knowledge for data standardization processes

For the most part, recognize two important concepts:
1) There will already be a good idea of what data entities are potential master data objects, and so the discovery process will probably support corroborating those suspicions

2) The challenges may revolve around agreement on standardized formats, definitions, attribution, and representations of master data objects, and the data profiling process, as a way to catalog and inventory enterprise data objects as a prelude to the data harmonization and standardization processes

Three (or Four) Aspects of Profiling

- Column profiling
- Cross-column (or, "dependency") analysis
- Cross-table (or, "redundancy") analysis
- Auditing

Column profiling involves looking at characteristics of data values within a single column.

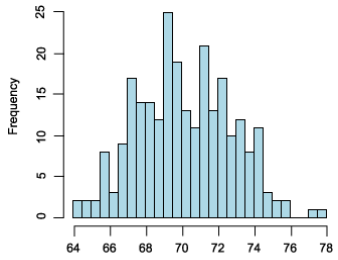
Cross-column (or, dependency) analysis looks for the dependency relationships that exist across columns within a single table.

Cross-table (or, redundancy) analysis looks for relationships that exists across tables

Auditing is used to compare new data instances against those rules discovered during previous profiling runs

Column Profiling Techniques

- Range Analysis
- Sparseness
- Format Evaluation
- Cardinality and Uniqueness
- Frequency Distribution
- Value Absence
- Abstract Type Recognition
- Overloading



Range Analysis, which is used to determine if the values fit within a well-defined range.

Sparseness, which evaluates the percentage of the elements populated.

Format Evaluation, which tries to resolve unrecognized data into defined formats.

Cardinality and Uniqueness, which analyzes the number of distinct values assigned to the attribute, and indicates whether the values assigned to the attribute are unique.

Frequency Distribution, which shows the relative frequency of the assignment of distinct values

Value Absence, which identifies the appearance and number of occurrences of null values.

Abstract Type Recognition, which refines the semantic data type association with a specific attribute.

Overloading, which attempts to determine if an attribute is being used for multiple purposes.

Range Analysis

- Determination of restriction of values within particular value ranges
- Applies to:
  - Integers
  - Numeric
  - Dates
  - Character strings

An example is an integer whose value lies between 0 and 100, or a date that falls between January 1, 1970 and December 31, 1979.

Range analysis can be performed by plotting a value distribution and bounding the range by the endpoints on the graph. In some instances there are multiple subranges within the value set, which might look like “islands” of value frequency interspersed with areas representing values with no occurrences.

Sparseness

- Evaluates the percentage of the elements populated
- May indicate an aging or unused attribute
- Depending on percentage populated, may relate in a consistency relationship to other attributes

A sparsely populated attribute is a candidate for further investigation. A completely empty column is probably a relic from the original data model – an attribute that was expected to be used but never put into production.

An attribute that has a very small number of values may indicate a process problem – for example, if values are floating from one field into the sparse field, it may mean that some input screen is not properly designed, or some update process or query is occasionally updating the wrong field.

Unexpected sparseness is likely to be a process problem as well, if a field is not being updated the way we expect it to be.

Format Evaluation

- A process to identify patterns within sets of values
- Determination of the use of a known standard or format
- Examples: dates, SSNs, telephone numbers, some names
- Also may be extant in organizational definitions (e.g., product codes, SKUs, policy numbers)

Once formats have been identified or defined, they can be used for validation of values through a pattern matching scheme. An example would be to restrict telephone numbers to the form (DDD) DDD-DDDD.

Note that this kind of restriction is structural only and does not validate the values within the attribute. So we may have a valid format for a telephone number even if the area code is 000.

Cardinality and Uniqueness

- Cardinality defines the number of distinct values that the attribute takes
- A cardinality equal to the number of records implies uniqueness
- Low cardinality implies a restricted value set
- Compare with range analysis and sparseness
- Unique value assignment to attribute indicates absence of functional dependency

The most frequent use of this analysis is to determine if a field can be used as a key to the table. An almost unique column may represent a referential integrity problem, where a field is expected to be a key and it turns out that it is not.

This analysis combined with range analysis can be used to determine if a field is a code. For example, if there are a few number of values, each of which appears frequently, and the lower end of the range is 0 or 1, that might indicate that the value relates to a code table.

Frequency Distribution

- Looks at the number of times each value appears
- Low frequency values ("outliers") are potential data quality violations
- Small number of very high frequency values may indicate a flag or code attribute

In some cases, such as names and addresses, the outliers may hold alternate representations of values that you expect to appear only once in the column. These are called approximate duplicates, and can be addressed using data cleansing software.

Value Absence

- Two different issues:
  - Absent values when they are expected
  - Presence of values when absence is expected
- Different meanings for absent values

There are many potential meanings for a null value, and there are different ways null values can be represented. The most common way is as the system null, which is available in most modern RDBMSs. Other ways include defined null or default values, such as 99-9999999 for the null Federal Tax ID.

The absence of a value may not even always mean that there is no value! In some cases, the system null is used as a default value. This is not a recommended practice.

Abstract Type Recognition

- An abstract type is a predefined pattern-based "sub-class" of a character string type.
- This process can exploit format evaluation to try to deduce attribute types

We can make use of the formats standards discovered through a format evaluation process to define patterns that can, together, make up an abstract type. For example, people names may take one of about 200 different forms, depending on use of titles, first names, last names, initials, suffixes, etc.

Overloading

- An overloaded attribute is one that carries more than one piece of information
- Two ways this is manifested:
  - Multiple "bits" of information in single value
  - Use of field for multiple uses depending on context

Overloading is a DBA trick to add a place to store new data when the model or the process is inflexible to change. An example I came across once stuck asterisks at the end of a name field to indicate a status associated with the record. This was fine as long as no one wanted to do anything with those names, but a couple of problems occurred:

- Data cleansing removed the asterisks
- Asterisks overwrote some names

Another aspect of overloading is the composition of multiple values into a single field; the ability to expose composed values requires algorithms with some understanding of string manipulations. For example, often values are prefixed with alternate data codes, such as region codes or price class data. A data profiling tool must be able to arbitrarily dissect a set of strings and evaluate their conformance to known domains to be able to detect composed values.

Cross-Column Analysis

- Key discovery
- Normalization & structure analysis
- Derived-value columns
- Business rule discovery

Primary Key Discovery

- Identify candidate primary keys within a table

A candidate key is one or more columns that can be used to uniquely identify any record in a data set. Key discovery is an interesting exercise, especially when inadvertent duplicate values mask potential keys from being naively exposed. Profiling can be used to identify potential keys based on relative uniqueness across the table.

Functional Dependency Analysis

- Column X is said to be functionally dependent on attribute Y if, for any pair of records R1, R2, if the value of attribute X is the same for R1 and R2, then the value of attribute Y will also be the same
- Y is said to be dependent on X
- Functional dependency analysis is used for normalization

The ability to determine if there are embedded relationships or embedded structures within a data set is enabled through the discovery of functional dependencies. Therefore, it is valuable to review what functional dependencies are and how they can be used as part of the discovery process.

A functional dependency between column X and column Y basically says that, given any two records, if the corresponding values of column X are the same, then the corresponding values of column Y will be the same, or that the value of column X *determines* the value of column

Normalization & Structure Analysis

- Identify embedded table structure based on functional dependencies

COLOR	QTY	PRODNUM	PRICE	TOTAL
RED	1	430001	32.99	32.99
BLUE	3	430002	32.99	65.98
YELLOW	2	430003	32.99	65.98
BLUE	1	430002	32.99	32.99
BLUE	1	430002	32.99	32.99
RED	3	430001	32.99	98.97
RED	2	430001	32.99	65.98
YELLOW	1	430003	32.99	32.99

There are two approaches for evaluating functional dependencies. The first is the clear identification of an existing dependency within the table – in other words, determining that a dependency exists and can be described. The second is the determination of an approximate dependency – in other words, what appears to be a dependent relationship within some margin of error.

Real dependencies are valuable because they can be used to isolate embedded tables that can be extracted and instantiated as part of a relational model. Embedded tables such as the one shown in this and the next slide are magnets for data quality problems, because of the significant redundancy that must be resolved whenever one of the values changes. By isolating the embedded table and extracting it into its own entity, we can reduce the potential for inconsistent data.

Dependencies

- In this example,
  - PRODNUM → PRICE, COLOR

PRODNUM	COLOR	PRICE
430001	RED	32.99
430002	BLUE	32.99
430003	YELLOW	32.99

QTY	PRODNUM	TOTAL
1	430001	32.99
3	430002	65.98
2	430003	65.98
1	430002	32.99
1	430002	32.99
3	430001	98.97
2	430001	65.98
1	430003	32.99

A dependency was present in the previous table, in which the values of both price and color were determined by the product number. The actual dependency analysis process might have first identified two separate dependences first (i.e., PRODNUM → PRICE, PRODNUM → COLOR), and in a second pass determined that the two dependences could be merged.

In this case, we could extract the relevant columns into a separate table (using PRODNUM as the primary key) and establish the foreign key relationship through the PRODNUM attribute.

Derived Columns

- Columns whose values are derived from values in other columns
- TOTAL = QTY * PRICE

COLOR	QTY	PRODNUM	PRICE	TOTAL
RED	1	430001	32.99	32.99
BLUE	3	430002	32.99	65.98
YELLOW	2	430003	32.99	65.98
BLUE	1	430002	32.99	32.99
BLUE	1	430002	32.99	32.99
RED	3	430001	32.99	98.97
RED	2	430001	32.99	65.98
YELLOW	1	430003	32.99	32.99

Using the same table as an example, we can also see that there is a more complex dependence between the QTY and PRICE fields and the TOTAL field, in that the value of TOTAL is derived as a function of QTY and PRICE. This kind of analysis requires smarter algorithms that are able to examine dependencies in the context of arithmetic functions or string manipulation.

Normalization

- Identified embedded tables can be extracted and reduced
- Creates a validity rule for input data for existence of embedded structure

An embedded table is isolated when a set of dependencies call out denormalized data within a single table. There are two issues to be addressed when an embedded table is discovered. The first is to determine whether the information in that embedded table should be extracted into its own separate table. Of course, this is only relevant when migrating to a new data model, or when developing an ETL process. The second issue is determining whether the dependency is the expression of a business rule that should govern the source data set. If so, then the rule should be documented and added to the set of rules for proactive auditing.

Cross-Table Analysis

- Foreign key analysis
- Synonyms
- Reference data coordination
- Business rule discovery

Foreign Key Analysis

- Multiple columns within same table evaluated for key candidacy
- Potential foreign keys are probed and discovered
- Known foreign keys are checked for referential integrity
- Detect orphan records in a child table whose foreign keys do not exist in the parent table

We have already seen that one profiling task is the discovery of candidate keys within a single table. In concert with that process is determining cross-table key relationships. For each table within the data set, each column (or set of columns, if necessary) that is not a primary table key is contrasted against the other table keys to see if a foreign key relationship potentially exists.

If there is a high likelihood of an existing foreign key relationship, that relationship is analyzed to measure consistency and to ensure referential integrity, and to report any violations.

Synonyms

- Evaluation of semantic consistency
- Example:
  - ACCOUNT
  - ACCOUNT_NUMBER
  - ACCT
  - ACCT_NUM
  - ACCT_NUMBER

Two columns are synonymous if they refer to the same conceptual data domain within the same business context. Frequently, though, columns in different tables may have slightly different (or sometimes, extremely different) names. The example in this slide focuses on a number of different ways that an individual's account number might be referred to in different tables.

The goal of synonym analysis is to determine whether two columns in different tables are referring to the same attribute. Identifying synonyms enables semantic clarity (by highlighting multiple names for the same thing) as well as potential dissociated logical entities distributed across multiple physical tables.

Reference Data Coordination

- Master reference tables can be assessed for completeness
- Analyze redundancy between reference table and derived associated tables

A major initiative in many industries is the resolution of master reference data and its consolidation into a single resource. Reference data coordination is a special case of cross-table and structure analysis, whereby the relevant basic entities are identified, reduced to a single representation, and all of their corresponding attributes are collected into a single repository.

An example of a reference data coordination project is embodied within a typical Customer Relationship Management (CRM) application. In some instances, this involves profiling account data to isolate representations of individual customers and then to map the customer information into a revised customer model. In essence, this is a special case of a structure assessment and data migration project.

Business Rule Discovery

- Identify the use of data domains and domain membership rules
- Identify consistency constraints within each record
- Identify uniqueness and dependence constraints within a table
- Identify cross-table consistency and dependence rules

In the context of profiling, business rule discovery is more than just determining data ranges or domain membership. If the goal of profiling is to establish a baseline for ongoing information compliance and auditing, then the business rules are the key components of a proactive validation strategy.

Data profiling helps bridge the gap between business client and implementer by exposing places where additional clarity is required. Profiling will not necessarily tell you what is right or wrong with your data. But when used wisely, profiling should point out opportunities to ask questions about the content under management, and this is a very powerful concept. By using profiling to find those business rules to which the data must comply, within a prioritization based on business impact, serious issues can be identified early in the processing chain and, consequently, their impacts avoided.

Auditing

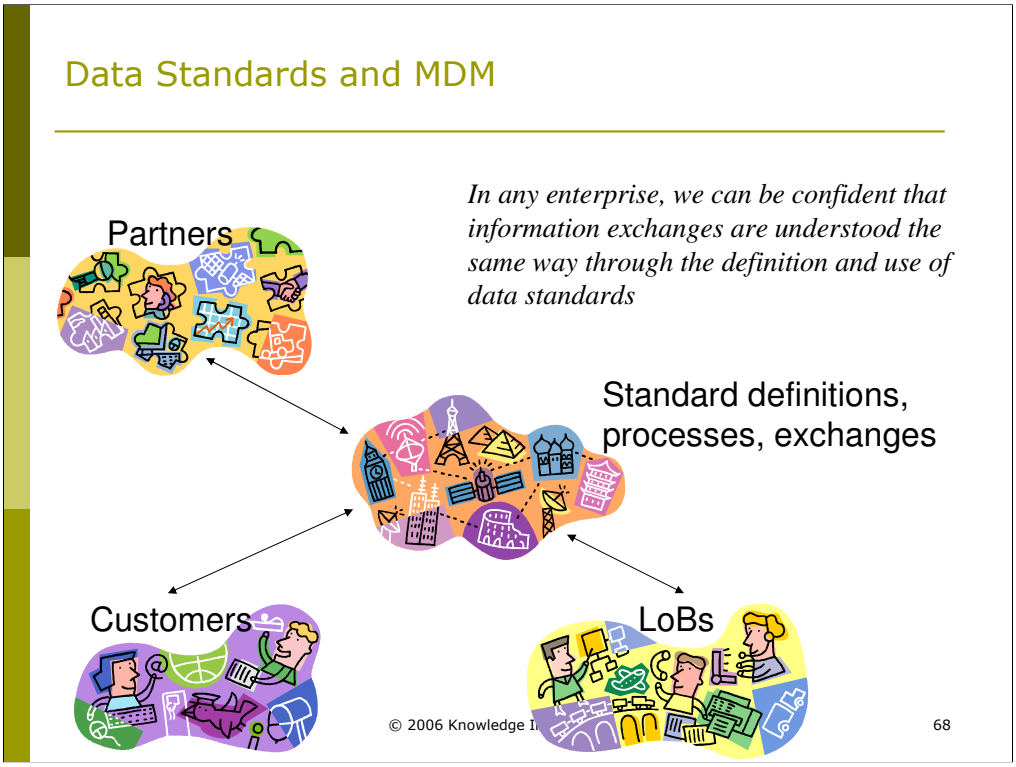
- Not all tools provide this capability
- Required to ensure that knowledge discovered during profiling is both captured and made actionable
- Use the business rules to validate the same (or similar) data sets proactively

As we have discussed, a major goal of profiling is deriving a set of business rules that capture the essence of information validity and using those rules to validate data moving forward. Given a set of rules one can formulate a contextual metric for measuring and monitoring information quality that is relevant within your business context.

Unfortunately, not all tools available on the market provide this capability. If yours does not, it is worthwhile to explore other options for capturing business rules and making them actionable.

Summary

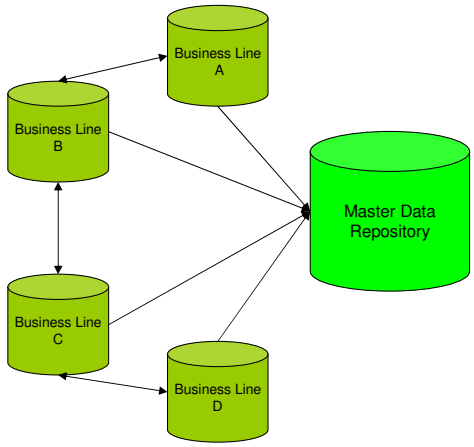
- Profiling can be used to expose different aspects of analytical and statistical artifacts within a collection of data sets
- Deriving metadata from this data involves automation + analyst insight
- Different levels of analysis provide increasingly valuable knowledge



A data standards process is an approach to synchronizing the various metadata aspects of shared or exchanged data objects. By formalizing the process of gaining consensus among the different participants, and enabling their active engagement in both defining and governing the process, we can evolve a collection of well-defined business terms, information object models, information exchange packages, and a means for mapping these shared object definitions into those models ingrained within our legacy environment. In addition, a data standards process can be used to help harmonize common business language terms and data elements to represent those terms as part of a master data management program.

Data Standards and Master Data Management

- Standards are employed in any kind of information exchange
- Aggregating data into an MDM registry or repository involves multiple exchanges
- We can use standards to simplify ETL, EAI, and EII processes



Why apply a standards process when you are not exchanging information with other organizations? The simple answer is that any time data moves from one system to another, it is considered information exchange. Consequently, an organization may employ a data standard to simplify the construction of data movement procedures even within the company, as in these kinds of projects:

•ETL for data warehousing

•Data aggregation

•EAI

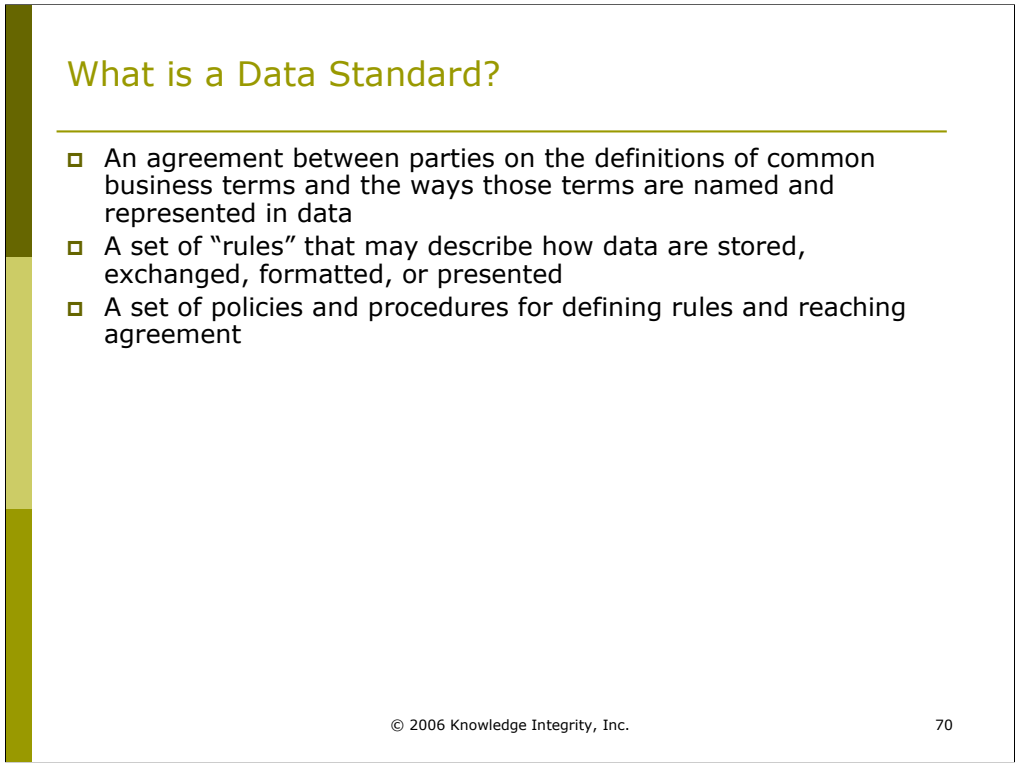
•EII

•Master Data Management

•Customer Data Integration

•Consolidated Reporting

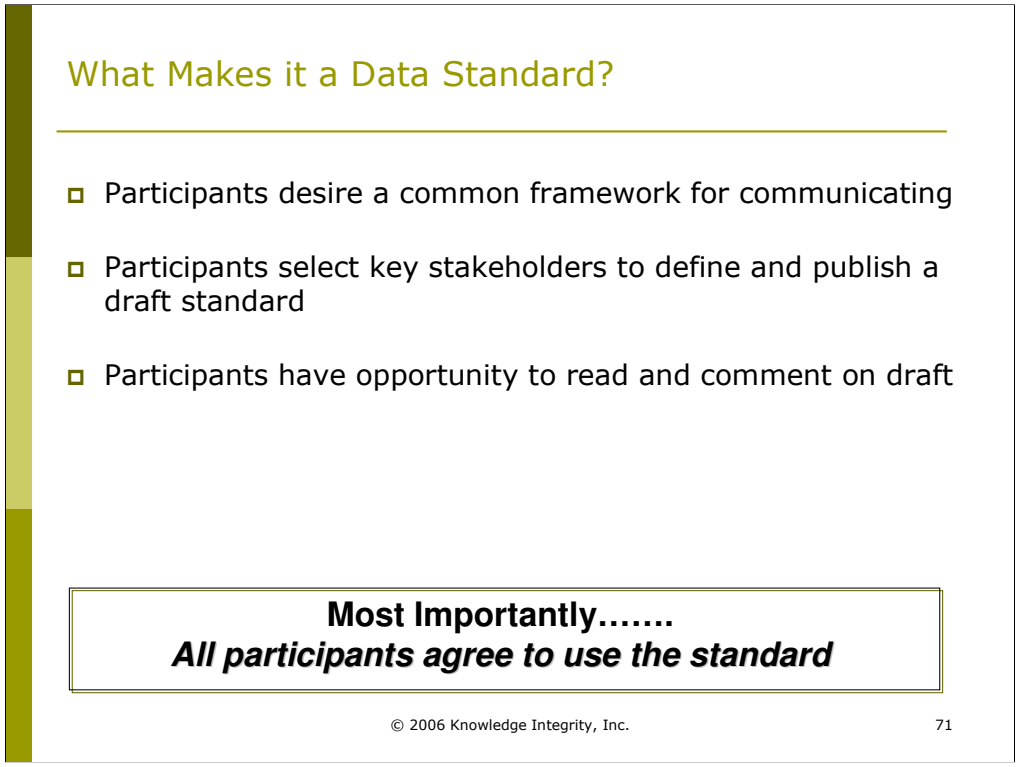
•Performance Metrics



A data standard encompasses the rules by which information is exchanged. This includes:

- The identification and definition of common business terms
- The determination of information objects to be exchanged
- The characterization of data elements as parts of information objects
- Rules for naming data elements and object classes
- Rules for the formats those data elements take on
- Rules for presenting those data elements to users

And mostly, the policies, procedures, and governance frameworks for ensuring compliance with the rules for defining standards and exchanging objects in a way that conforms to that standard.

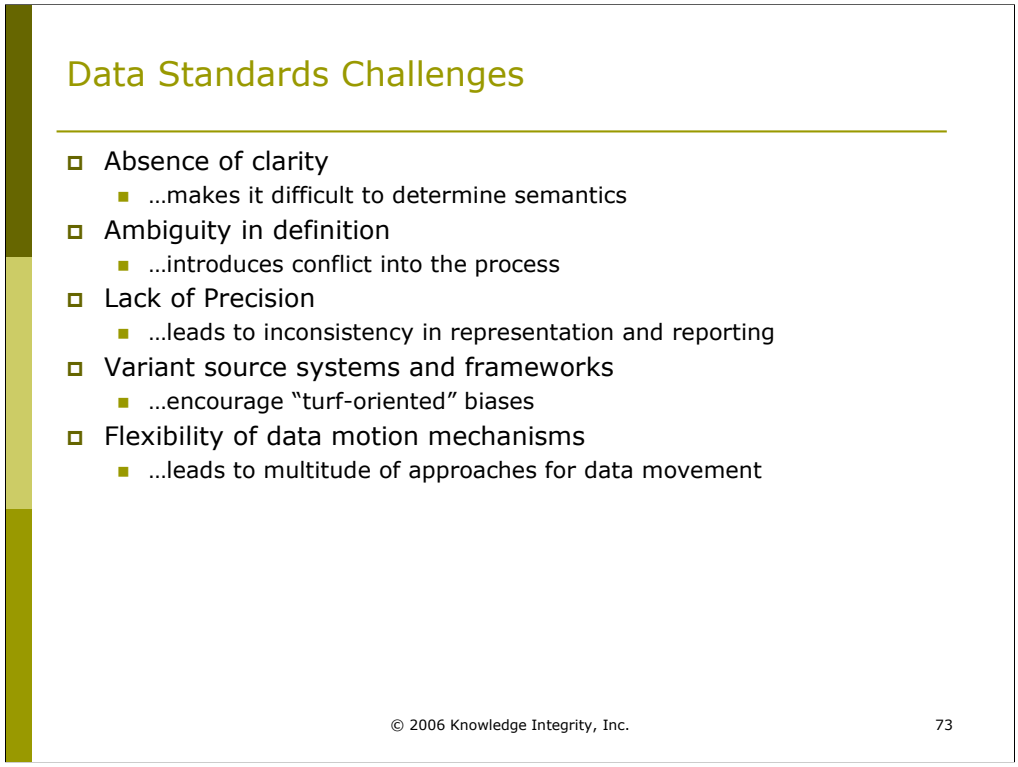


The essence of a standard is not the consolidation of metadata, the formal description framework, or even the rules for definition and governance. Instead, it is the fact that the individual stakeholders desire to work together to develop a common language, that participants are provided with an opportunity to participate in the process through proposing new standards, evaluating proposals, and providing comments, and most importantly, that the participants agree that the defined and endorsed standard is the single set of guidelines to be used for information exchange.

There may be some mitigating circumstances, but for intents and purposes, for any business environment there should be one data standard. Dueling standards create an environment similar to the one you may be trying to fix, in which the same business objects are defined in slightly variant ways. A strong governance framework for data standards management is necessary to ensure that the best value can be obtained from the process.

Benefits of Defining Data Standards

- Enable effective communication between multiple parties expressing an interest in exchanging information
- Reduce manual intervention in data movement processes
- Develop opportunities for increased automation
- Establish a shared catalog for enterprise business terms and related exchanged data elements
- Streamline on-demand client access
- Support ongoing system upgrade requirements
- Improve data quality



The main issues that are introduced when developing a data standards program revolve around putting in place the kinds of practices that are usually ignored during standard application design and implementation:

•Absence of clarity for object semantics: Relying on the implied meanings associated with business terms may be fine when it the system is self-contained, but as soon as there is a need to compare values between two or more environments, subtle differences in meanings become magnified.

•Ambiguity in definition: The ambiguity is typically aligned along application, and subsequently, departmental lines; the exposure of ambiguity will encourage individuals to promote their own semantics to the exclusion of others, and this plants the seeds for organizational conflict.

•Lack of Precision: People tend to be less than precise in standard conversations, because humans can derive understanding through context. However, in an imprecise environment, it is difficult to resolve measurements and metrics into a unified view.

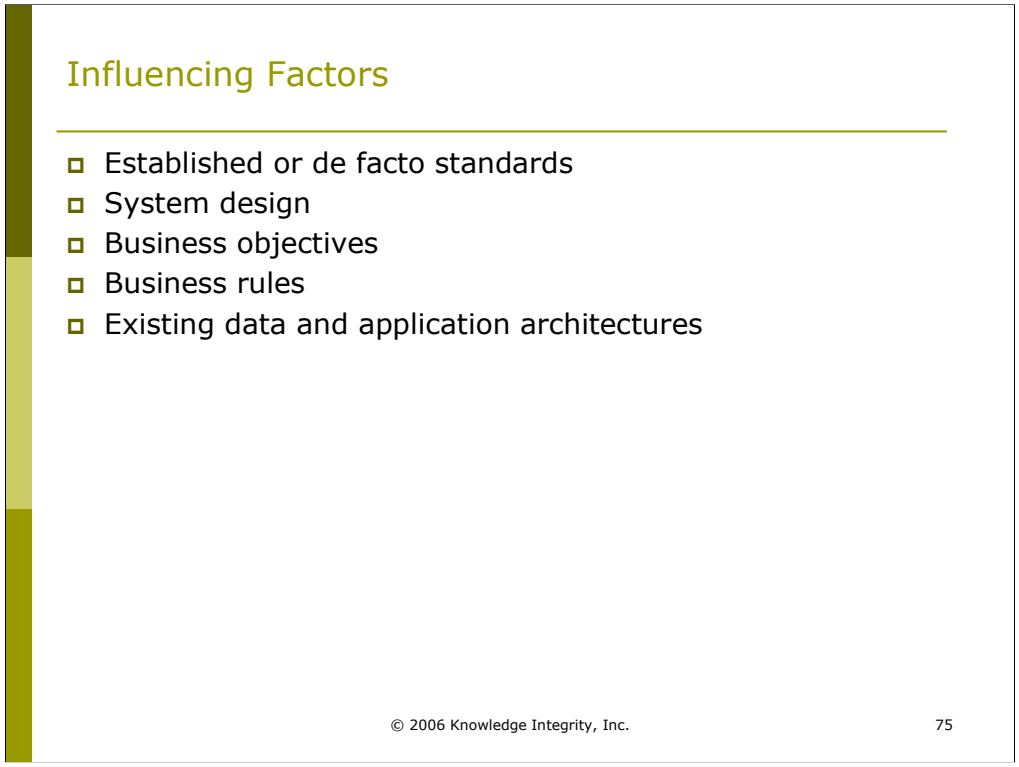
•Variance in Source Systems: Aside from the semantics issues, implementation decisions may create reliance on application frameworks, leading to religious wars (e.g., .NET vs. J2EE, XML vs. Flat data)

•Flexibility of motion mechanisms: The multiple modes by which data is exchanged can expose conflict when trying to create a seamless means for exchange. This may mean creating adapters that can transform data objects between formats, such as between records in flat files and XML documents.

Stakeholders and Participants

- Stakeholders: Those whose business processes rely on consistent, high quality information
- Participants: Those who are involved in the definition, refinement, and endorsement of data standards as well as those involved in the development and deployment of those data standards
- Goal: *Try to make all of them happy*

Ultimately, the goal is to use a governance process to allow creative input into collecting the numerous business terms that are used, correlating those terms to the underlying data objects that represent them, and agreeing to a standardized definition and representation framework within which all can collaborate. This does not mean forcing everyone to change the way they name and use data objects within their own applications, but rather agree to one set of formats and meanings for common exchange.



Before embarking on a data standards program, consider the existing factors that may contribute to the overall success of the program:

•Established standards: There are many established standards, issued by government bodies or by recognized standards bodies, that are in use already within your enterprise. A good example is the use of country codes (US, CA, etc.) that are derived from the International Standards Organization's ISO 3166 standard

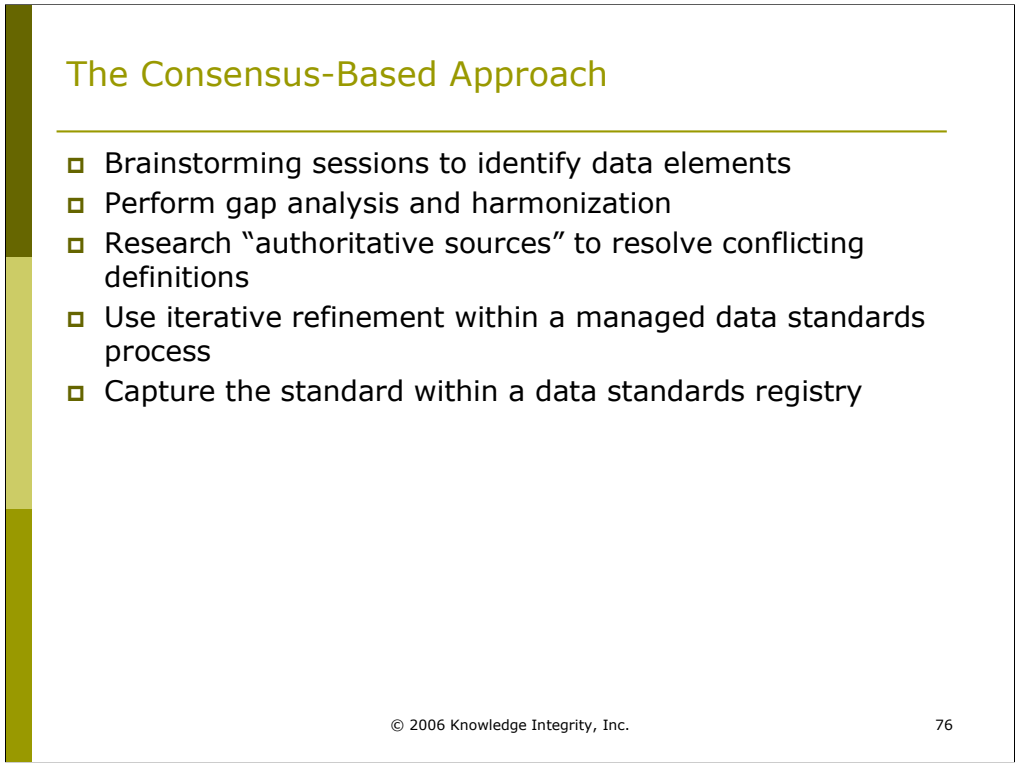
•De Facto Standards: often there are standards that are in use that may not have been "officially" sanctioned, but there is a common understanding that the standard is used. For example, consider the use of USPS Postal State Codes, which are not a standard sanctioned by an official body, but are well-suited for many application purposes and are therefore adopted.

•System Design: The way you application is built may exert influence over the way one expects data to be exchanged. Proponents of batch processing on flat files may be more biased towards flat file exchanges, and less so towards XML documents.

•Business Objectives: A healthy organization drives their development based on their business objectives, and clearly defined objectives will have some influence on the standards process, especially when collaboration is a high-level driver.

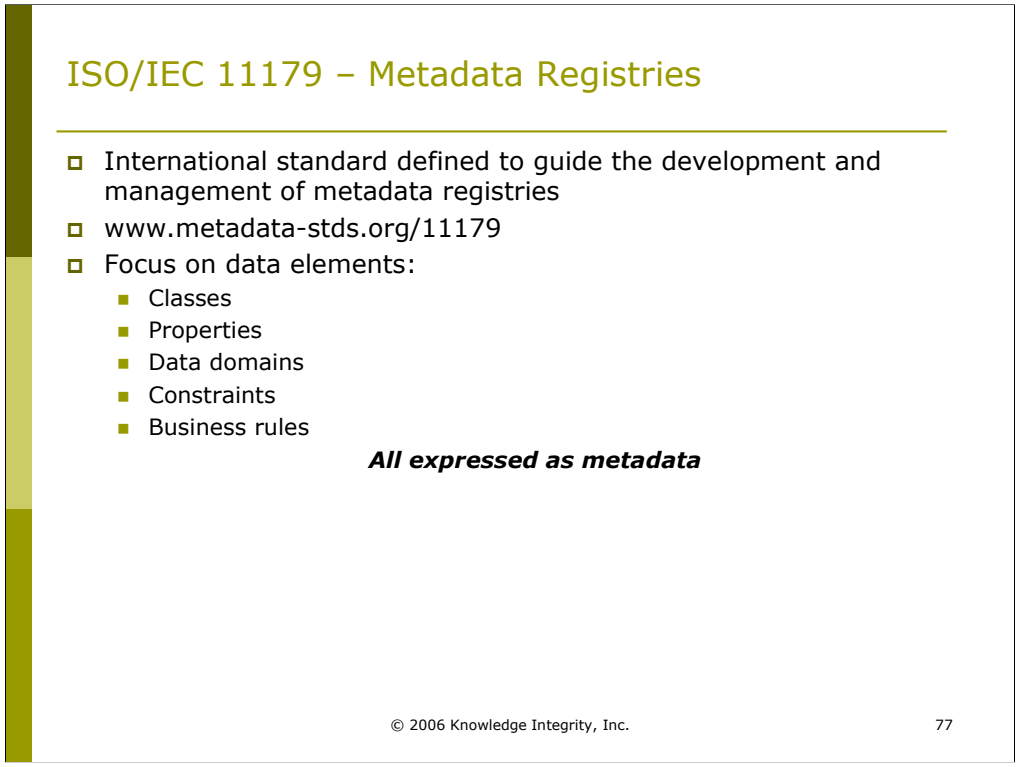
•Business Rules: Guidance that influences activities within the organization (as well as outside of the organization) are critical to the definition of common business terms. In fact, the sources of those business rules may provide the authoritative definitions that can eliminate ambiguity

•Existing Data and Application Architectures: As organization s grow and deploy new functional applications, the distribution of business function and operations management enable creative development that may not reflect a standard set of policies applied across the board. Trying to .



The approach involves gathering the subject matter experts to work together to agree on the terms of the standard. This involves:

- A free-form "brainstorming" session where the terms used are identified, documented, and made available for discussion of a more precise definition.
- A gap analysis to determine whether definitions already exist, and if not, what additional information is needed. At the same time, the terms that are discovered can be categorized within some logical hierarchy.
- Researching the information needed based on the previous step to locate complete and acceptable definitions, if they exist.
- For each term, consensus is sought on behalf of the participants by reviewing all the distinct definitions for each term, and either accepting a definition, revising a definition, or creating a new definition. As the participants reach agreement, the definitions are documented along with any supporting information that influenced the decision.
- The definitions and supporting information are assembled in a glossary.



Underlying any data standards initiative is metadata. Particularly in an environment where emphasis is placed both on information dissemination (i.e., communicating data standards) and on governance (i.e., overseeing the definition an endorsement process), it is worthwhile to employ a registry as an information resource. Registries are similar to repositories, capturing metadata at the data element level, except that the registry can capture workflow status. In addition, providing a browsing capability enables a greater potential for reusing business terms and definitions that already have some degree of acceptance. We will briefly explore the ISO standard 11179 on Metadata Registries as the basis for our Data Standards Registry.

11179 Overview

Part 1: Framework	Fundamental ideas of data elements, value domains, data element concepts, conceptual domains, and classification schemes
Part 2: Classification	Provides a conceptual model for managing classification schemes
Part 3: Registry Metamodel	Specifies a conceptual model for a metadata registry. It is limited to a set of basic attributes for data elements, data element concepts, value domains, conceptual domains, classification schemes, and other related classes, called administered items
Part 4: Formulation of Data Definitions	Provides guidance on how to develop unambiguous data definitions
Part 5: Naming and Identification	Provides guidance for the identification of administered items
Part 6: Registration	Provides instruction on how a registration applicant may register a data item with a central Registration Authority

ISO 11179 & Data Standards

Object Class	A set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning and whose properties and behavior follow the same rules
Property	A characteristic common to all members of an ***Object Class***.
Data Element Concept	A concept that can be represented in the form of a ***Data Element***, described independently of any particular representation.
Data Element	A basic unit of data of interest to an organization (formed when a ***Data Element Concept*** is assigned a representation).
Representation Class	A ***Classification Scheme*** for representation (a mechanism for conveying to an individual functional or presentational categories).
Conceptual Domain	A set of ***Value Meanings***, which may either be enumerated or expressed via a description.
Value Domain	Provides representation, but has no implication as to what data element concept the values are associated nor what the values mean.
Glossary	**Not in ISO 11179**; a set of controlled terms with their definitions as used within a data standards environment.

Data Type	A set of distinct values, characterized by properties of those values and by operations on those values.
Unit of Measure	The unit in which any associated ***Data Element*** values are specified.
Non-Enumerated Conceptual Domain	A ***Conceptual Domain*** that cannot be expressed as a finite set of ***Value Meanings***. It can be expressed via a description or specification.
Enumerated Conceptual Domain	A ***Conceptual Domain*** containing a finite allowed inventory of notions (expressed as a set of ***Value Meanings***).
Non-Enumerated Value Domain	An expression of a ***Value Domain*** as description or specification such as a rule, a procedure, or a range.
Enumerated Value Domain	An expression of a ***Value Domain*** as an explicit set of two or more ***Permissible Values***.
Permissible Value	An expression of a ***Value Meaning*** within an ***Enumerated Value Domain***.
Value Meaning	An expression that provides distinction between one member and the other members in an ***Enumerated Conceptual Domain***.

Conceptual Domain

"A set of valid value meanings."

Example:

US States – The set of primary governmental divisions of the United States.

A conceptual domain encapsulates a set of "value meanings," which represents a set of things that belong together within some business context. The domain is conceptual when it is not specifically bound to a representation. The example we will use is that of states of the United States.

Note that the definition we use may not necessarily be exactly aligned with the way the conceptual values are always used, and this in itself presents interesting food for thought. This definition focuses on "governmental divisions," as opposed to "geographic divisions." This might imply slightly different semantics in practice, such as if an application were intended to aggregate individuals by location. One might say that this definition is not sufficient to allow its use in that business context…

Value Domain

"A set of permissible values"

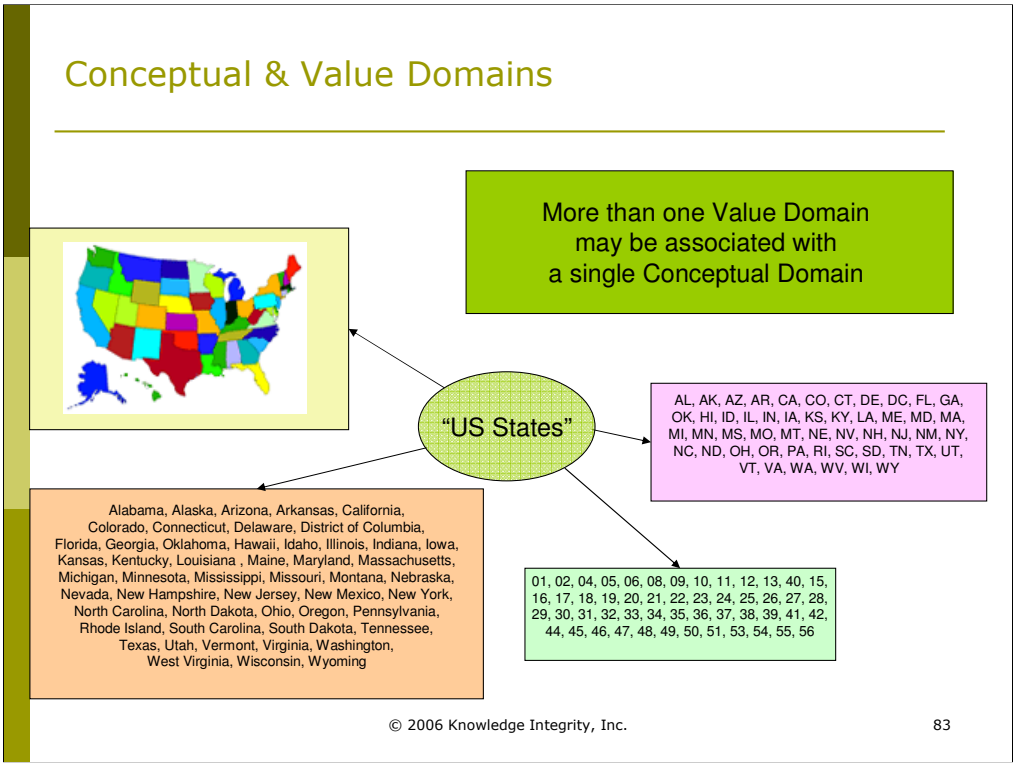
Example:
"Alabama", "Alaska", "Arizona", ..,
"Wisconsin", "Wyoming"

A value domain is a set of permissible values, but it also implies a mapping between the actual representative values and the conceptual meaning that backs it up. For example, the character string "Alabama" is mapped to the conceptual object Alabama that is one of the primary governmental divisions of the United States.

Domain Associations

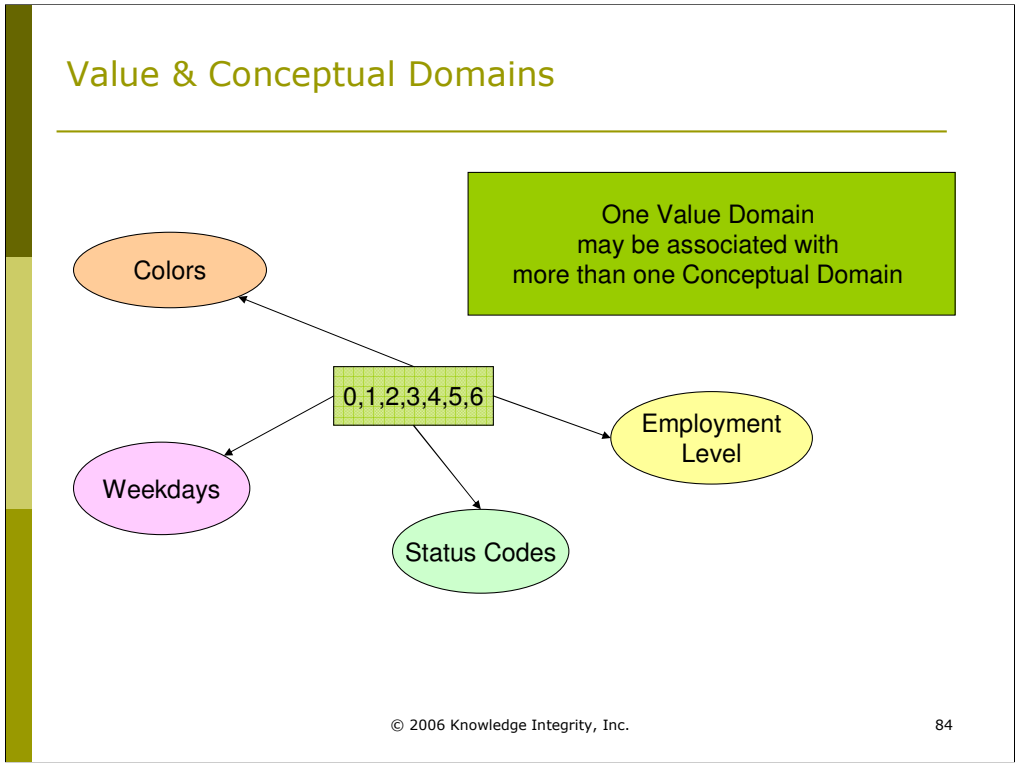
- More than one Value Domain may be associated with a single Conceptual Domain
- One Value Domain may be associated with more than one Conceptual Domain

What is becoming apparent is that there may be many ways that the same conceptual domain is represented across a collection of organizations. The important thing to remember is to synchronize those value domains that refer to the same conceptual domain. Within a metadata registry, one may define the conceptual domain and associate it with a number of value domains. It would be worthwhile to provide a mapping between each value domain and the “things” in the conceptual domain, which could be used to relate the value domains to each other as well. We look at some examples on the next 2 slides.



This slide demonstrates a number of value domains related to the single conceptual domain "US States." They include:

- Postal Abbreviations
- NIST Federal Information Processing Standard 2-digit codes
- State name strings
- Geographical images



Alternatively, and this may be a source of confusion, the same value domain may be used to represent different conceptual domains. In this example, an enumeration of integers is used as a code to represent 4 different conceptual domains.

Data Element Concept

"A concept that can be represented in the form of a **Data Element**, described independently of any particular representation."

Example:

Issuing Jurisdiction – The governing body in which a license was issued

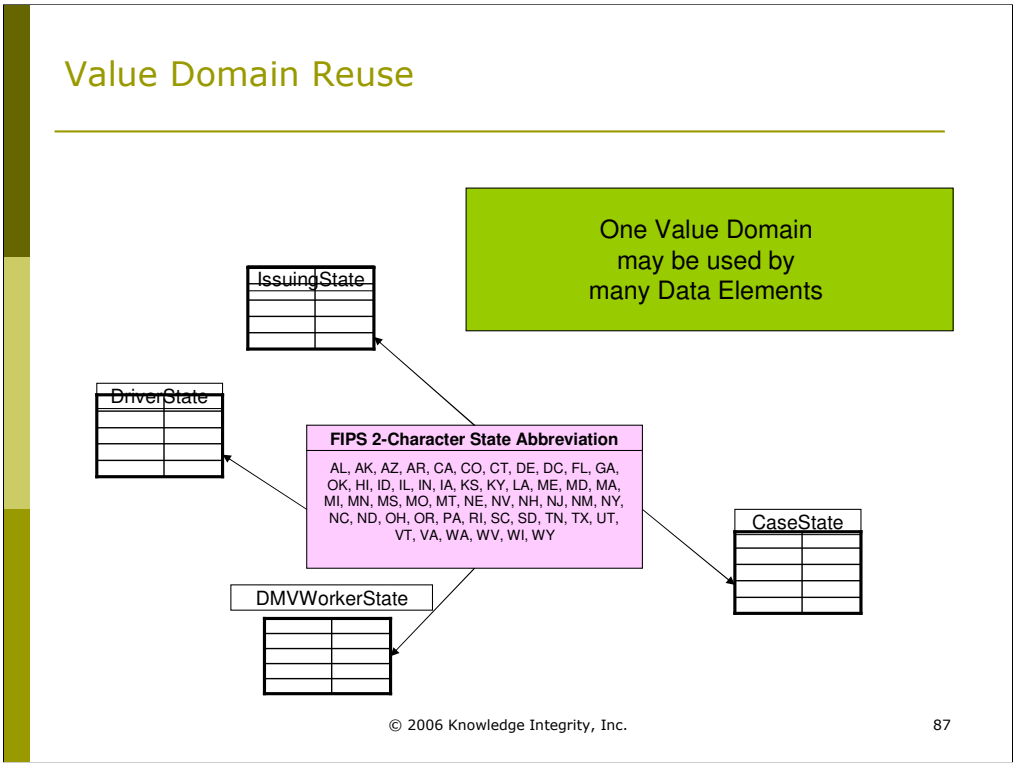
Within 11179, a "data element concept" represents a concept and its description, but independent of any specific representation. Data element concepts employ conceptual domains; the actual applications are intended to select an appropriate value domain that is valid within their business contexts.

Data Element

"A unit of data for which the definition, identification, representation, and permissible values are specified by means of a set of attributes."

Name	**IssuingState**
Data Element Concept	Issuing State
Used By	E-DMV
Definition	The state in which the license was issued
Data Domain	FIPS 2-Character State Abbreviation
Data Type	CHAR(2)
Presentation	CHAR(2)
Business Rules	Not null
Unit of Measure	N/A
Data Steward	John Doe
Meta Tag Name	<IssuingState>

Within the framework of data element concepts, a data element encapsulates the definition, representation, and permissible values associated with its data element concept. It is through this relationship that we start to see how business terms are associated down into the data layer. The data element captures representation metadata, such as name, data type, presentation type, data domain, etc.



Value domains essentially become subclassed data types, which can then be incorporated into a growing directory of available types. Once the domain has gained approval through the standards process, these standardized value domains may be reused in many different environments.

Summary

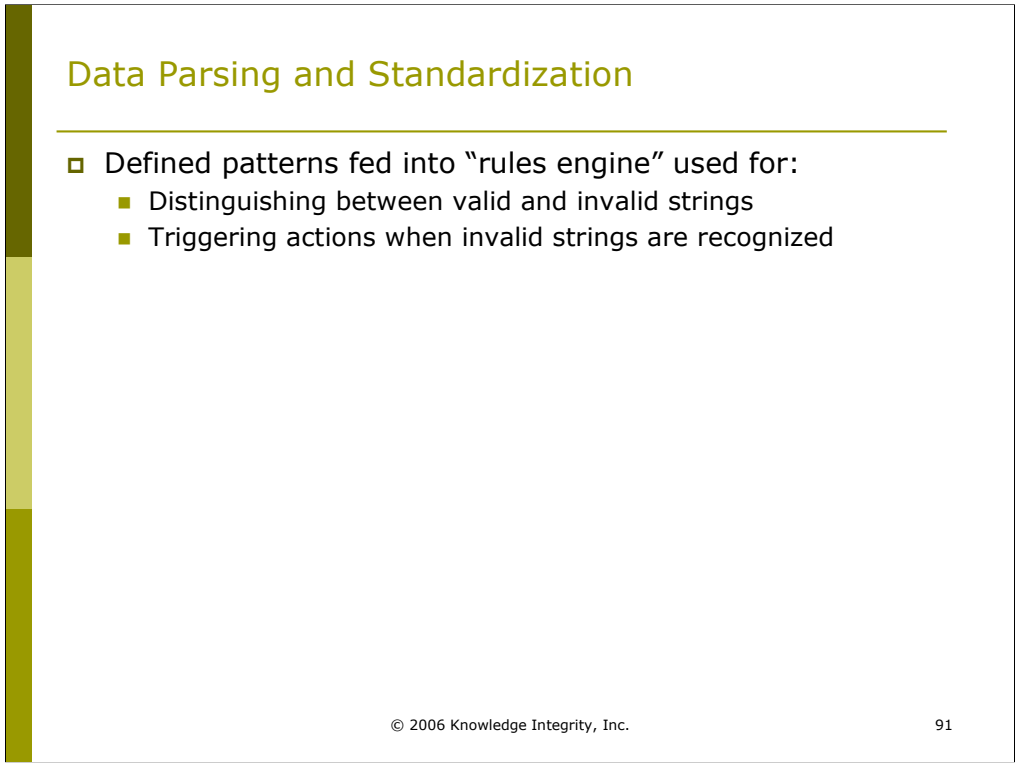
- Data quality tools can be used to:
  - Catalog table metadata from vertical application architectures
  - Discover sources of master data from across the enterprise
  - Verify expectations of existing master data
- Data standards processes support:
  - Clarification of semantics
  - Harmonization of detail
  - Unifying a model for sharing data

Part 3:

Data Quality Tools and Technology

Data Quality and MDM

- Data profiling
- Data parsing and standardization
- Record Linkage/Matching
- Data scrubbing/cleansing
- Data enhancement
- Data auditing/monitoring



To address the frequency of variations due to data entry (and other) errors in names, addresses, and other root causes, organizations employ tools to recognize data values that conform to known patterns, identify those that do not conform to known patterns, and to transform those values into a standard form for further analysis and review. The general approach is to use defined patterns, consisting of low-level strings that can be recognized as tokens, and then higher-level patterns of those tokens, that can be fed into a pattern analyzer. Each token, or string component, may optionally be recognized and tagged, and the engine may allow for a transformation to be applied as a standardization step, either modifying the tokens themselves, or even the ordering of the individual tokens.

For example, a pattern might recognize that "Smith, John A." has three tokens, a last name ("Smith"), a first name ("John"), and a middle initial ("A."). If the business process requires a standardized format for names consisting of first name, middle initial, and last name, these three recognized tokens can be rearranged as "John A. Smith."

When a data value is fed into the application, the pattern recognition process attempts to match the value to one (or more) patterns. There are three possible outcomes:

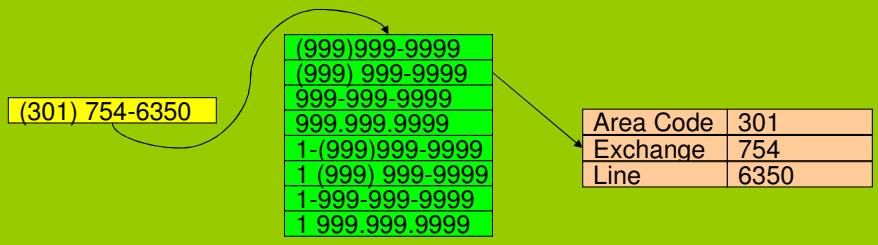
The value matches a single pattern

The value matches more than one pattern

The value does not match any patterns

In the first case, the parts of the data value that matched the tokens in the pattern can be transformed into the standardized format for output. In the second case, where more than one pattern is matched, there must be some prioritization applied to automatically perform the transformation. In the third case, the data value is identified as being one that does not fit into any known patterns, and should be presented for manual review.

Parsing and Standardization Example

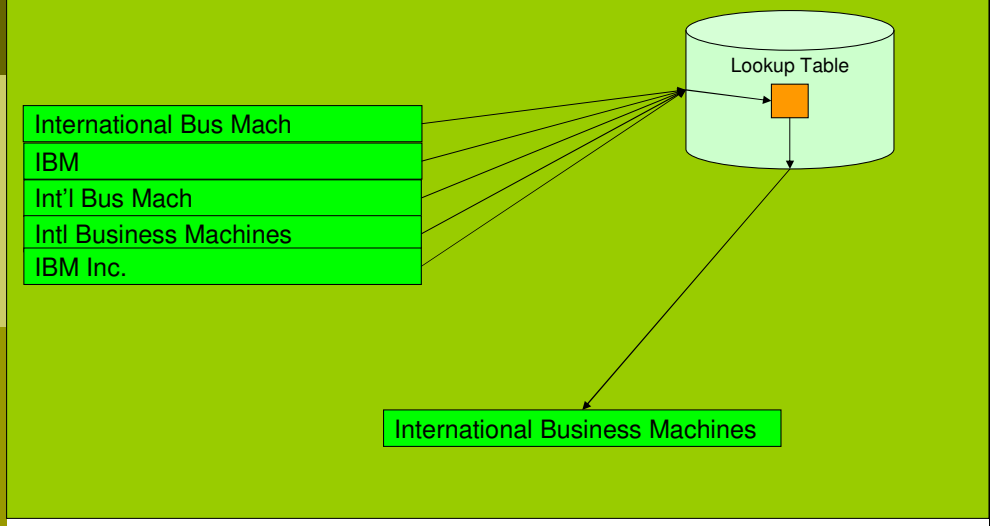


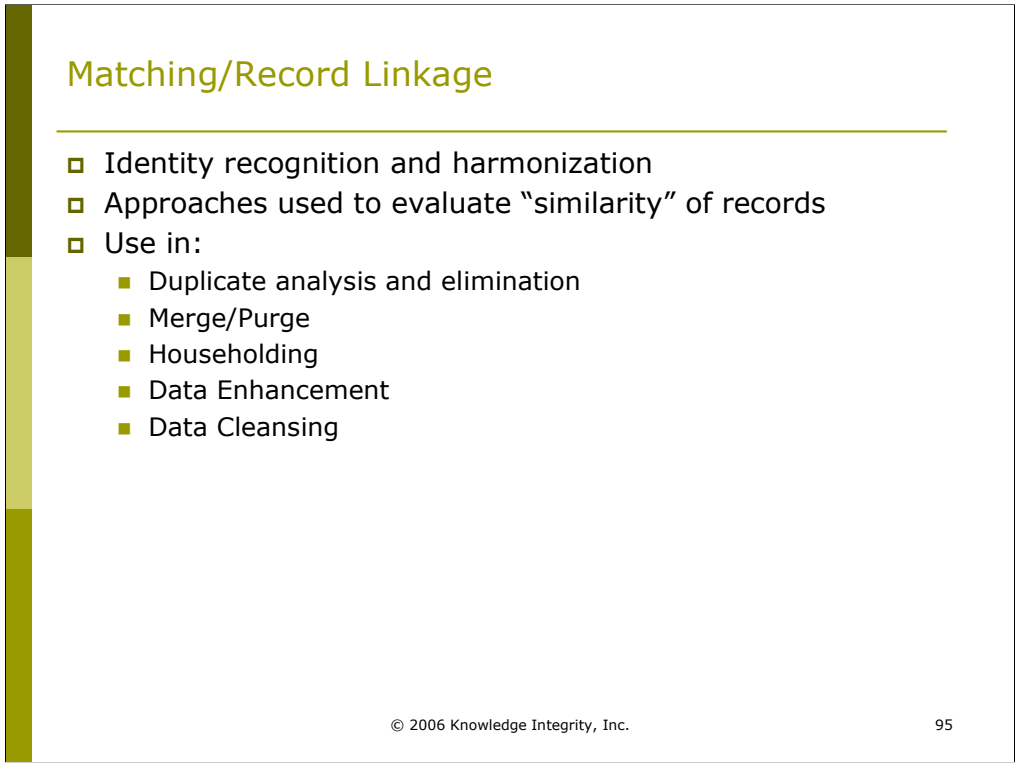
Data Scrubbing/Cleansing

- Identify and correct flawed data
  - Data imputation
  - Address correction
  - Elimination of extraneous data
  - Duplicate elimination
  - Pattern-based transformations
- Complements (and relies on) parsing and standardization

As opposed to parsing and standardization, which recognizes patterns and transforms the data into a standardized form, data scrubbing and cleansing are processes to identify and correct flawed data. This includes tasks such as data imputation, value correction, elimination of extraneous data, duplicate elimination, as well as pattern-based transformations. Incorporating a data cleansing process, and integrating cleansing with matching at the points of data entry may allow for recognition of variations of individual or practitioner names, provide suggestions for corrections, and thereby significantly reduce the number of potential replication early in the information flow.

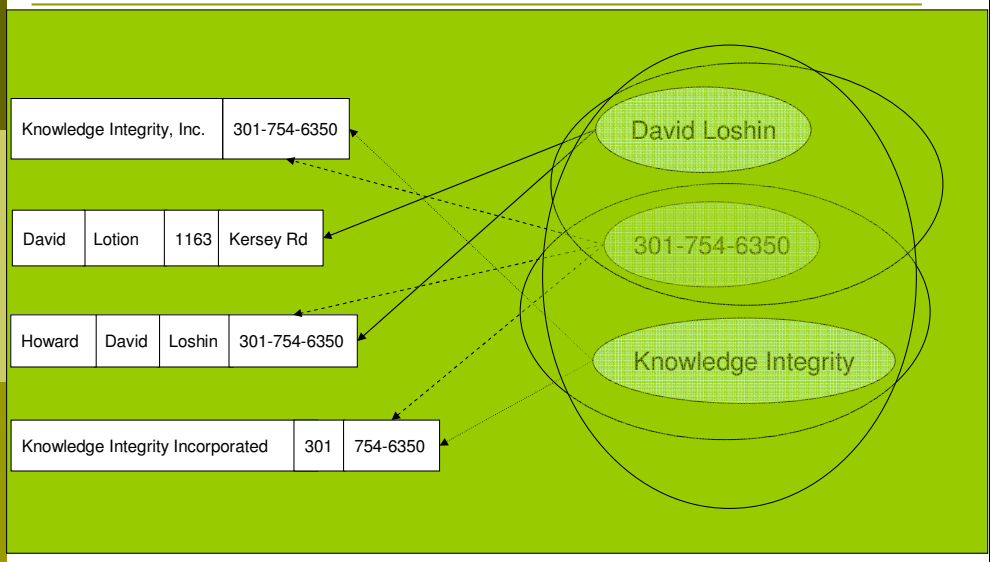
Cleansing Example





One of the key aspects of an MDM program is the need to uniquely resolve the identity of any specific individual or organization. However, it is common that variations of both person and organization names prevent the determination of duplicate data, since two similar names may refer to the same person or organization. Record linkage and identity resolution tools are employed in identity recognition and resolution, and incorporate approaches used to evaluate "similarity" of records for use in duplicate analysis and elimination, merge/purge, householding, data enhancement, and cleansing.

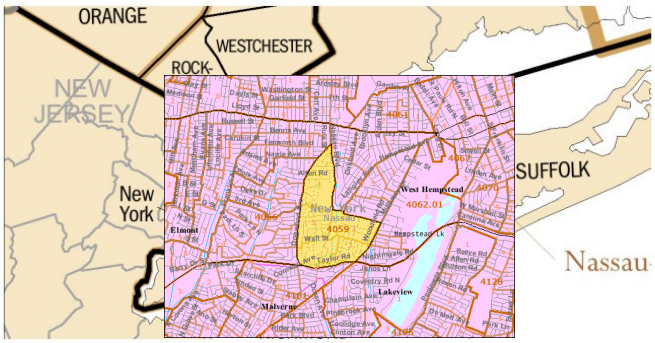
Matching/Record Linkage - Example



Data Enhancement

- Data improvement process that relies on record linkage
- Value-added improvement from third-party data sets:
  - Address correction
  - Geo-Demographic/Psychographic imports
  - List append
- Typically partnered with data providers

Data enhancement is a data improvement process that relies on record linkage between provided data and value-added third-party data sets to improve the provided data. Examples include address correction, data cleansing, and Geo-Demographic/Psychographic and list appends.

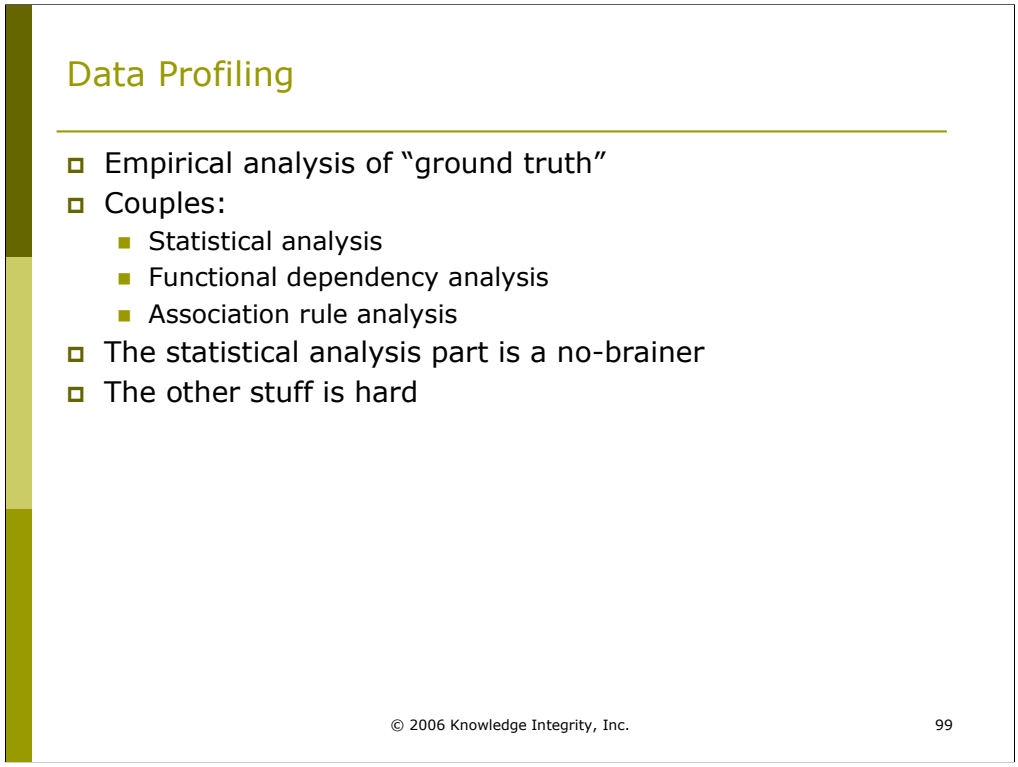


Subject	Number	Percent
OCCUPANCY STATUS		
Total housing units	**1,704**	**100.0**
Occupied housing units	**1,681**	**98.7**
Vacant housing units	**23**	**1.3**
Tenure		
Occupied housing units	**1,681**	**100.0**
Owner-occupied housing units	**1,566**	**93.2**
Renter-occupied housing units	**115**	**6.8**

Subject	Number	Percent
SCHOOL ENROLLMENT		
Population 3 years and over enrolled in school	**1,571**	**100**
Nursery school, preschool	142	9
Kindergarten	102	6.5
Elementary school (grades 1-8)	662	42.1
High school (grades 9-12)	335	21.3
College or graduate school	330	21
EDUCATIONAL ATTAINMENT		
Population 25 years and over	**3,438**	**100**
Less than 9th grade	122	3.5
9th to 12th grade, no diploma	249	7.2
High school graduate (includes equivalency)	996	29
Some college, no degree	674	19.6
Associate degree	180	5.2
Bachelor's degree	676	19.7
Graduate or professional degree	541	15.7
Percent high school graduate or higher	89.2	(X)
Percent bachelor's degree or higher	35.4	(X)

The specific area we are going to look at is a neighborhood in the town of West Hempstead, NY. The area is contained within what is called a census tract, specifically tract 4059. In this picture, the area in yellow is the area under investigation, and the subsequent census data refers to this area. As a matter of detail, this area is approximately four square miles, and as of the Year 2000 census, contained about 1,700 houses, and covered about 5,200 people. If you look more closely, you will see within the yellow area a partitioning into five smaller areas (called block groups) labeled 1, 2, 3, 4, and 5. Furthermore, block groups are then again divided into smaller areas called census blocks. A census block group may cover around 1,000 people; a block may cover a few hundred (or even fewer).

It is interesting to note the degree of granularity that is associated with a census tract. The fact that the data covers a relatively small population is valuable because it allows us to draw our conclusions in the context of a small enough geographic space. To refer back to our theme that birds of a feather flock together, we can exploit the small granularity to provide generalizations about those individuals that live within each census area.

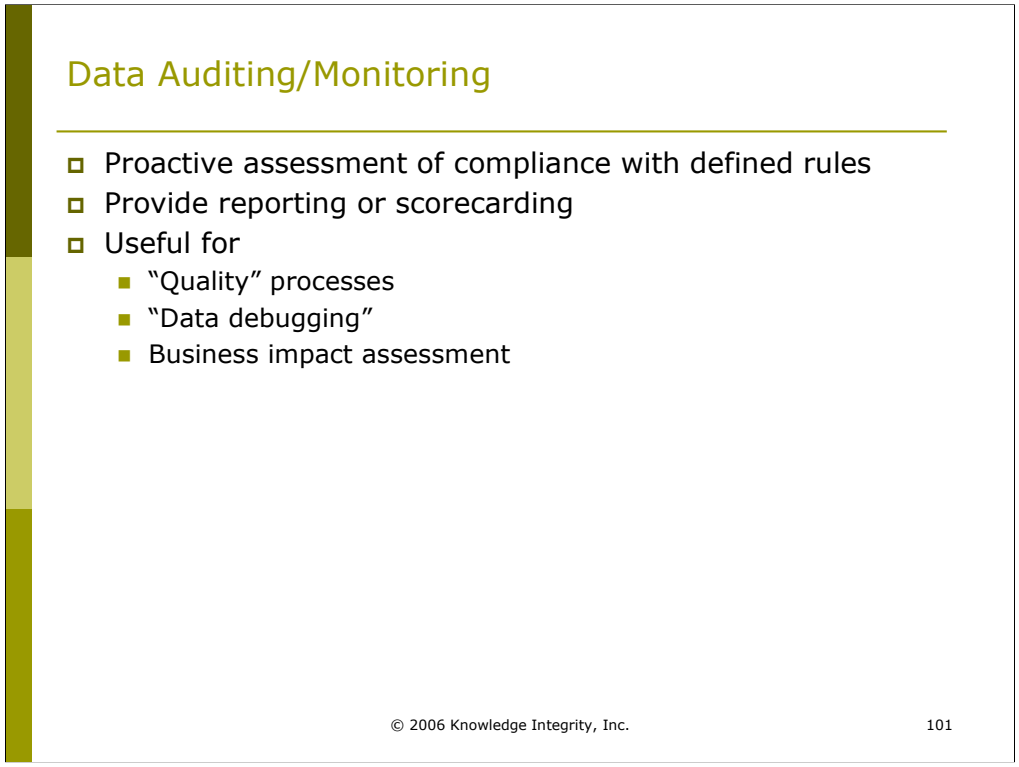


Data profiling is a set of processes that provide empirical analysis of what is in the data, employing statistical analysis, functional dependency analysis, and other data mining techniques such as association rule analysis. The results of data profiling provide insight into metadata and data models, as well as data typing, data domains, valid ranges, and other business rules that are embedded within data sets.

Discovered anomalies and dependencies can be reviewed and analysts can identify key data quality rules for validation. These rules can be deployed from within the data profiling tools for proactively validating data against data quality expectations. Consequently, profiling tools will be used for both analyzing data to discover data rules and subsequently auditing data against a set of discovered or defined rules.

Data Profiling – Approach

- Column Profiling/Frequency Analysis (Easy)
  - Done in place using database capability (e.g., SQL), or using hash tables
- Cross-Table/Redundancy (Harder)
  - May be done using queries, or using set intersection algorithms
- Cross-Column/Dependency (Hardest)
  - Uses fast iterative set partitioning algorithms
  - A priori, TANE, FD_MINE



Auditing and monitoring are processes that evaluate data to assess and report conformance with defined data quality rules. This proactive assessment can either be applied to a set of data values contained within a managed data set, or can be applied though the information flow to validate data before it is transmitted between processing stages. Data quality auditing and monitoring activities include:

•Assessing data set conformance to defined business rules,

•Logging the performance scores on a periodic basis,

•Identifying if the levels of quality do not meet the business expectations and triggering the generation of notification events, and

•Tracking data quality levels from day to day to provide an ongoing view of measured data quality.

Data profiling tools are increasingly incorporating the functionality for defining data quality rules and a means for testing data records against those rules. This capability is useful for auditing data rule conformance at different points in the information flow to provide quality analysis, data debugging, and business impact analysis.

Summary

- Customer Data Integration is an approach to enabling organizations to build a 360° view of their customers
- Historical degradation of the quality of data impacts the success of a CDI implementation
- Combining sound information management practices with information quality tools will help you devise a strategy for CDI success!

Part 4:

Data Governance

Fundamental MDM Challenges

- Organizational
  - Ownership
  - Governance
  - Change Management
- Technical:
  - Consolidation and integration of data
  - Entity identification
  - Master data management architecture
  - Application Integration

Data Ownership

- Most organizations lack clearly defined data ownership and data stewardship policies
- Questionable responsibility, accountability, and authority associated with information quality
- Lack of consistency and standards regarding data conventions, naming, and usage

Data Governance

- Instituting responsibilities and accountability for sound information management practices
- Information valuation at organizational level helps in establishing proper "information asset management"
- Establishing long-term information exploitation strategy

Change Management

- Challenge:

Quality assessments and governance ➨

  - Fear
  - Obstruction
  - Lack of cooperation
  - …

- Before embarking on an MDM program, plan to address the behavioral and change management issues you are bound to face

Technical Challenges

- Finding your master data
- Figuring out the consolidated view
- Managing the metadata
- Managing the policies and procedure for data quality management

- These are *all* data quality opportunities!

Example – Financial Industry

- Bank with distributed, unconnected application infrastructure
- Challenges:
  - Inconsistent and incomplete customer data
  - Duplication of customer data
  - Duplication of functionality
  - No information architecture
- Impact:
  - Inability to exploit customer analytics
  - Low product to customer ratio
  - Inability to manage attrition
  - Exposure to compliance risk
- Recommendations:
  - Introduce CIO, Information architect, Data Governance
  - Master Data Management for customer data

Example: Retail

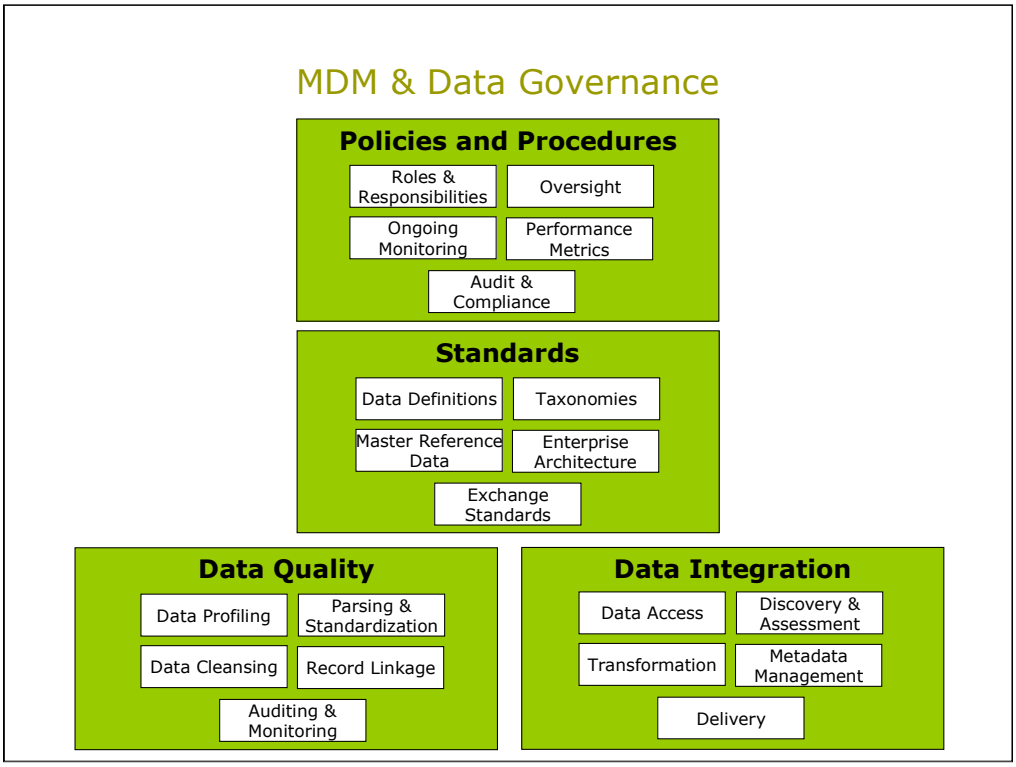
- Consumer electronics – assembles devices from catalog of parts
- Challenges:
  - Lack of consolidated view of Global inventory/catalog
- Impacts:
  - Inconsistency leads to design and manufacturing rework
  - Limited component reuse
  - Commodity inventory depreciation
- Recommendations:
  - Apply a Data Standards approach to resolution and standardization of product names
  - Create a Master Data Management program for products
- Similar to other industries with large catalog of similar products with variance in attribution across suppliers (e.g., footwear)
- Customer, supply chain, vendor analytics

Example: Public Sector

- National healthcare agency
- Challenges:
  - Distributed management of healthcare individual data
  - Federal/Jurisdictional environment imposes different policies, regulations
- Impacts:
  - Difficult to assemble a complete healthcare record for any specific individual
  - Privacy constraints restrict ability to browse information
- Recommendation:
  - Develop a unique healthcare identifier as a global key, indexed across all participating data sources (MDM registry)

Example: Retail

- Supermarket chain
- Challenges:
  - Multiple data warehouses addressing separate operational and analytical applications
  - Distribution of resources and data sets
- Impacts:
  - Limited ability to perform
    - Customer analytics
    - Customer profiling
    - Market basket analysis
    - Shrinkage analysis
- Recommendation:
  - Create a master repository populated from all applications to feed all data warehouses


MDM & Data Governance
Policies and Procedures
Roles & Responsibilities
Oversight
Ongoing Monitoring
Performance Metrics
Audit & Compliance
Standards
Data Definitions
Taxonomies
Master Reference Data
Enterprise Architecture
Exchange Standards
Data Quality
Data Profiling
Parsing & Standardization
Data Cleansing
Record Linkage
Auditing & Monitoring
Data Integration
Data Access
Discovery & Assessment
Transformation
Metadata Management
Delivery

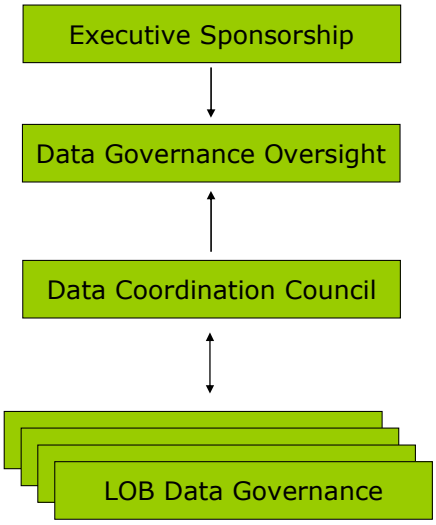
Oversight: Maintaining High Quality Data

- Is an MDM program just another database?
- Consensus from all fronts:
  - MDM must incorporate governance and oversight framework to ensure quality
  - MDM is driven by collaboration and data sharing, implying the need for policies and protocols

Characterizing and Maintaining High Quality Data

- Identify the critical dimensions of data quality that are relevant within the business context
- Establish objective metrics for evaluating data quality
- Identify levels of expectation for data quality conformance
- Characterize means for measuring conformance to those expectations
- Provide a framework for reporting, tracking, and escalating data quality issues
- Report and communicate levels of data quality across the enterprise

Roles and Responsibilities



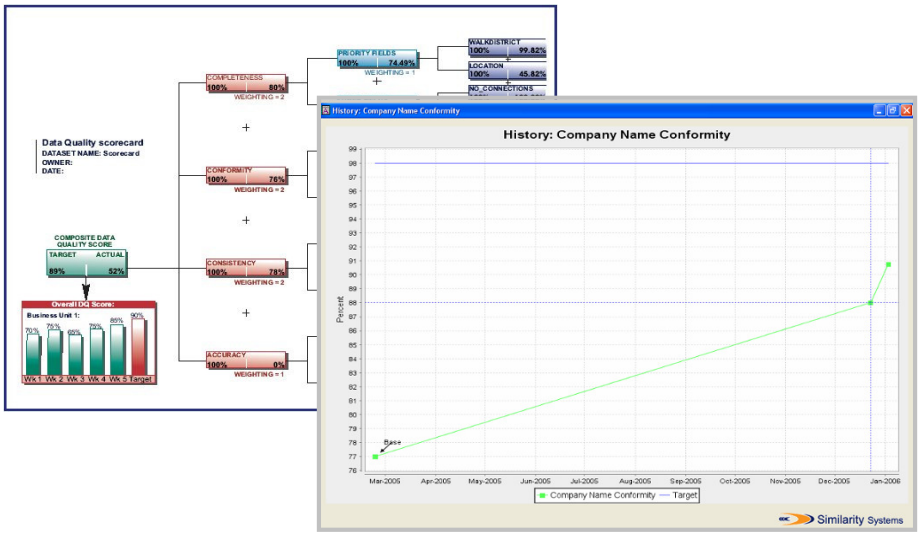
Provide senior management support at the C-level, warrants the enterprise adoption of measurably high quality data, and negotiates quality SLAs with external data suppliers.

Strategic committee composed of business clients to oversee the governance program, ensure that governance priorities are set and abided by, delineates data accountability.

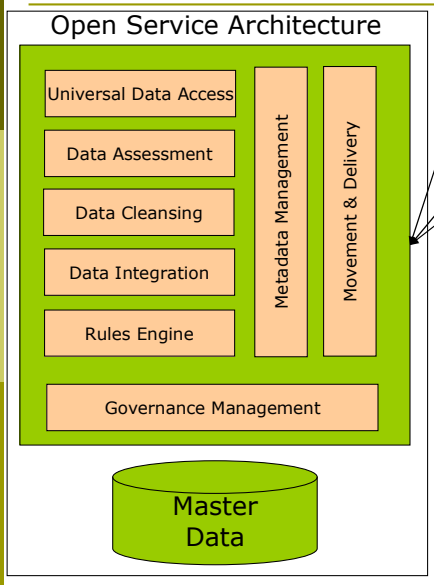
Tactical team tasked with ensuring that data activities have defined metrics and acceptance thresholds for quality meeting business client expectations, manages governance across lines of business, sets priorities for LOBs and communicates opportunities to the Governance Oversight committee.

Data governance structure at the line of business level, defines data quality criteria for LOB applications, delineates stewardship roles, reports activities and issues to Data Coordination Council

Link a "DQ Scorecard" to Business Performance



The Service-Oriented Approach

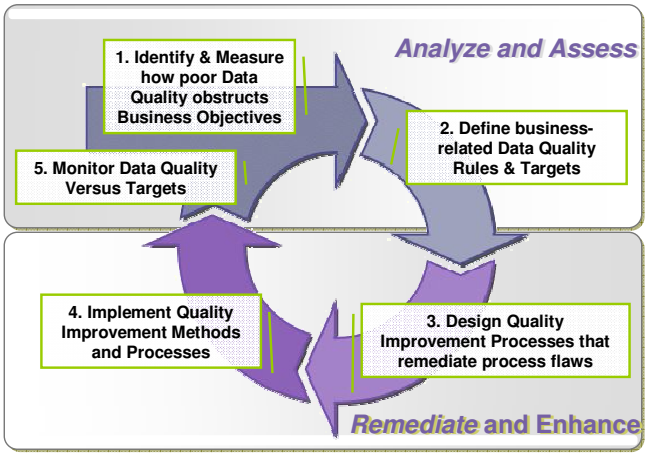


Accessibility	Provides universal data access
Availability	High availability
Quality	Cleansing and matching against master reference entities
Consistency	Parsing, standardization, and transformation using metadata
Auditability	Information flow, audit trail, and data lineage captured
Security	Access is secure and limited appropriately

Data and *Process* Reuse: Master Data Integration

- Processes used to integrate a centralized, services-oriented, high quality master data table can be employed for any master data set:
  - Customers
  - Products
  - Partners
  - Regulations
- Consolidating reference sets improves transactional productivity and operational efficiency;
- Establishing bi-directional universal data access, data integration, and relational connectivity between centralized reference sets exposes potential for business performance improvement through
  - Reduction in complexity
  - Customer intelligence through analysis
  - Increased sensitivity to risk

A Process-Driven Approach to Data Quality



 Source: Informatica

DQ Center of Excellence: Benefits

- Standardizing the methods & tools used for DQ improvement
- Achieving economies of scale in SW and HW acquisition
- Providing a DQ improvement service model
- Amortizing program development investments across the enterprise
- Consolidating and documenting best practices performed across the enterprise and allow everyone to benefit from common experience
- Establishing a forum for developing and agreeing to data standards
- Coordinating and synchronizing professional training in both the use of the tools and methods
- Reducing overall project TCO

Data Quality COE: Management Initiatives

- Establishing data quality monitoring and improvement as a business imperative
- Acquiring, then deploying the proper tools, methods, and expertise to improve the exploitation of reference information
- Transitioning from a reactive to a proactive organization with respect to data quality
- Prepare the organization to be a high-quality Master Data Integration environment

Building the Center of Excellence

- Identify senior management champion to ensure organizational support
- Create an enterprise data quality management group to oversee and manage DQ activities
- Institute data governance policies and procedures whose accountabilities cross the boundary between IT and Business
- Establish expertise in use of data quality tools and methods for multiple purposes
- Create a service model to deploy best practices
- Educate, educate, educate!

Example : Center of Excellence

- Insurance Company:
  - Forms a data quality team with representatives from most subsidiary organizations
  - Gains senior management support
  - Identifies best practices and integrates them across the enterprise
  - Consolidates SW and HW purchases for economies of scale
- … Is rewarded with recognition with the Data Warehousing Institute's Best Practices Award

Summary

- MDM/CDI/PDI… is both driven by the need for data quality, and the ability to let applications exploit high quality data
- Traditional data quality technology is being integrated into MDM solutions
- MDM/CDI and data quality are enabling technologies, making other (*failed?*) initiatives successful
- MDM is a *program* for marrying data governance, technology, and change management to achieve business objectives

DQ Management Goals

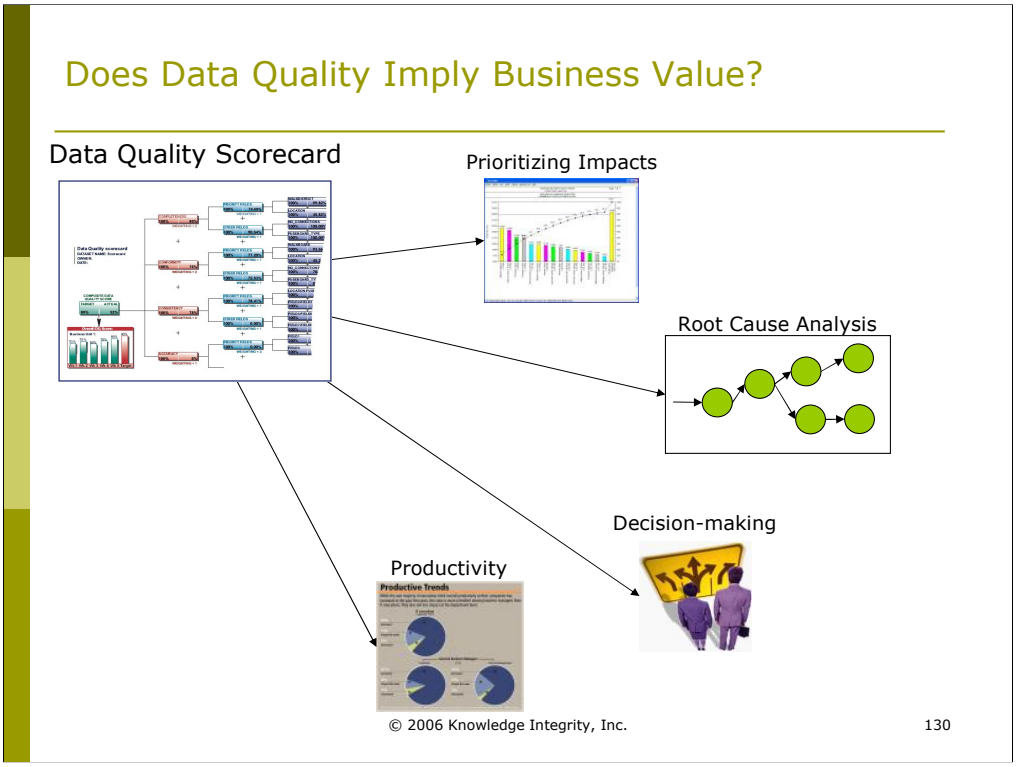
- Evaluate **business impact** of poor data quality and develop ROI models for Data Quality activities
- Document the **information architecture** showing data models, metadata, information usage, and information flow throughout enterprise
- Identify, document, and validate **Data Quality expectations**
- **Educate** your staff in ways to integrate Data Quality as an integral component of system development lifecycle
- **Management framework** for Data Quality event tracking and ongoing Data Quality measurement, monitoring, and reporting of compliance with customer expectations
- Consolidate current and planned **Data Quality guidelines**, policies, and activities

Turning Data Quality into Process Quality

- Institute a data governance framework
- Use business-driven data validity assessment to baseline current state and to measure ongoing improvement
- Establish data quality issues tracking to improve internal remediation within an accountability chain
- Develop a services-based approach to your centralized reference master(s)
- Establish best practices for data management for other enterprise data sets

Governance and Oversight

- Awareness of the problem
- Establishing a value proposition
- Engaging the appropriate sponsors
- Data Governance frameworks



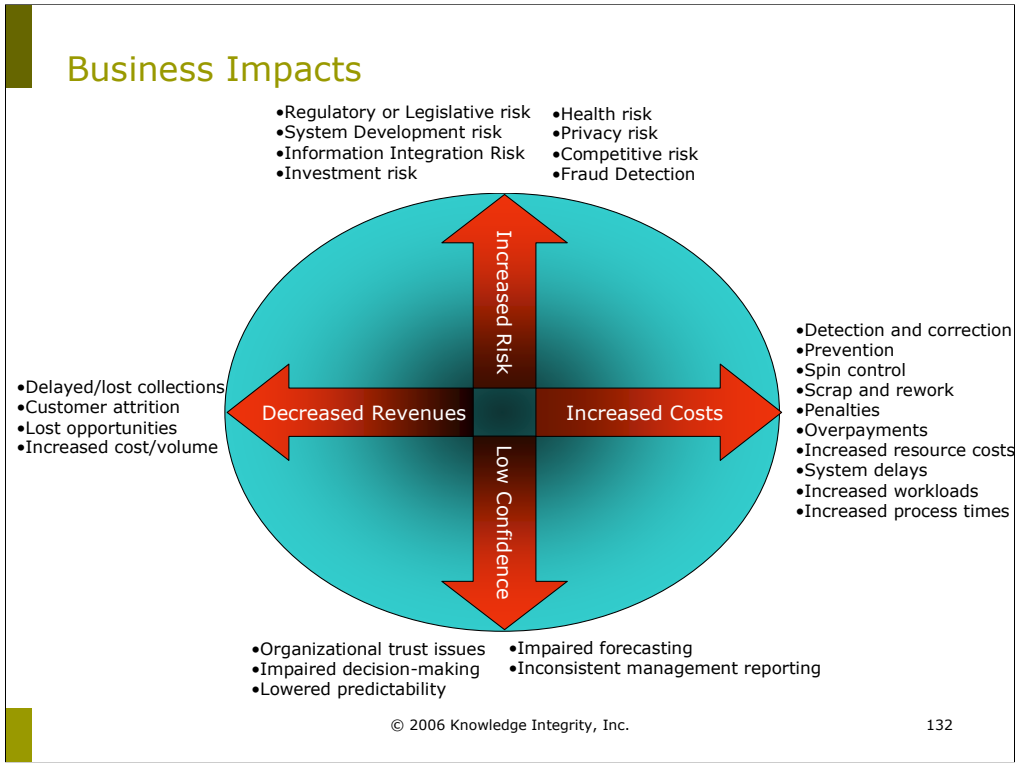
Assumption: improved data quality, “golden copy,” and integrated reference data sets all imply business value

However,

- How are data quality metrics tied to business performance?
- How do you distinguish high impact from low impact data integrity issues?
- How do you isolate the source of the introduction of data flaws to fix the process instead of correcting the data?
- How do you correlate business value with source data integrity?
- What is the best way to employ data integration best practices to address these questions?

Finding Business Relevance

- Identify key business performance criteria related to information quality assurance
- Review how data problems contribute to each business impact
- Determine the frequency that each impact occurs
- Sum the measurable costs associated with each impact incurred by a data quality issue
- Assign an average cost to each occurrence of the problem
- Validate the evaluation within a data governance forum



The best way to start thinking about a return on an IT investment is to evaluate the real pains associated with running the business. The goal of running any business is to have the most productive employees develop a high-value product to be sold at the optimal price to the largest set of good customers with the lowest reasonable investment and the least amount of acceptable risk.

This statement implies a strong dependence between each of these embedded variables:

-Increased Costs

-Decreased Revenues

-Decreased confidence

-Increased Risk

Note that the statement effectively defines an optimization problem in terms of maximizing the value of the result based on impacts associated with each dimension. Our job is to determine when and where poor information quality affects one or more of these variables.

Cost:

Some of these costs are incurred when try to address information quality issues, while others are incurred by ignoring them. Detection and correction costs are incurred when a problem has been identified, and these may be relatively large but infrequent. Alternatively, prevention costs may be incremental costs that are ongoing, and may diminish in expense as time goes on.

Spin control costs are those costs associated with ensuring that any information quality impacts that are exposed outside of the organization are mitigated. An example might be a discovery (by an external organization like a newspaper) that decisions about which medical procedures are approved by the health insurer are based on faulty data, which might indicate that the needs of the member are not always being met properly. The cost of spin control includes the costs of publicity to address the discovery, plus any acute costs incurred to immediately modify procedures in place to close the perceived gap exposed by the discovery.

Scrap and rework refers to costs associated with rolling back computations, undoing what had been done, and starting again.

Information quality problems can impact timeliness of service; if there are well-defined service-level agreements that are not being met, penalties for missing objective targets may be incurred.

The inability to properly track all representational and contact information related to financial agreements with business partners can result in accounting failures, leading to potential overpayments or duplicated invoice payments.

When there is limited control or oversight over the costs associated with procurement, or when there is limited ability to see a consolidated view of all provider interactions, inaccuracy, and subsequently, missed opportunities are bound to occur.

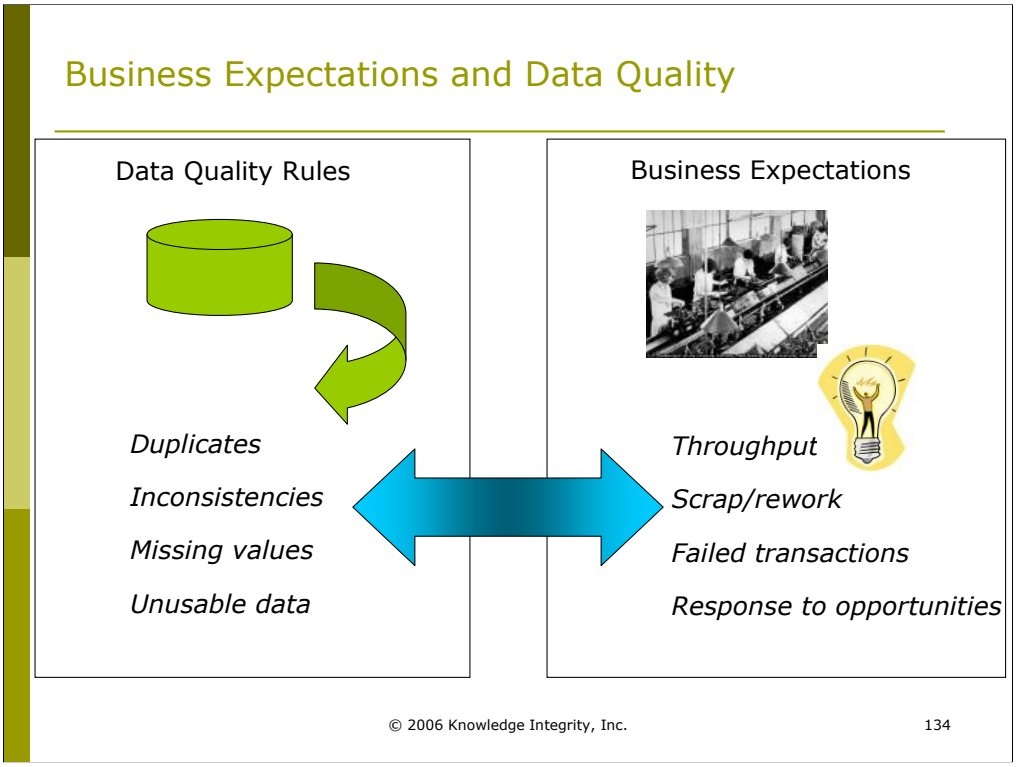
The Challenge of Sponsorship

- Business partners will expect business justification and a reasonable business case ("ROI") before funding an ongoing program

But…

- The right people in the organization who "get it" can transcend the need for "ROI"

- Objective: Socialize the value of improved data quality to support ongoing business productivity improvement as well as operational process improvement
  - Engage stakeholders with implicit or explicit "ownership"
  - Foster a collaborative approach to oversight and management



Data quality expectations are expressed as rules measuring completeness, consistency, validity *of data values*

What data is missing or unusable?

Which data values are in conflict?

Which records are duplicated?

What linkages are missing?

Business expectations are expressed as rules measuring performance, productivity, efficiency of *processes*

How has throughput decreased due to errors?

What percentage of time is spent in scrap and rework?

What is the loss in value of transactions that failed due to missing data?

How quickly can we respond to business opportunities?

Yet, to determine the true value added by data integrity programs, conformance to business expectations should be measured in relation to its component data integrity rules

Business Expectations and Data Quality

- Data quality expectations are expressed as rules measuring completeness, consistency, validity *of data values*
  - *What data is missing or unusable?*
  - *Which data values are in conflict?*
  - *Which records are duplicated?*
  - *What linkages are missing?*
- Business expectations are expressed as rules measuring performance, productivity, efficiency of *processes*
  - *How has throughput decreased due to errors?*
  - *What percentage of time is spent in scrap and rework?*
  - *What is the loss in value of transactions that failed due to missing data?*
  - *How quickly can we respond to business opportunities?*
- Yet, to determine the true value added by data integrity programs, conformance to business expectations should be measured in relation to its component data integrity rules

This requires collaboration between the technical and business teams, supported by senior management sponsorship

Dimensions of Data Quality

- Three conceptual categories:
  - Intrinsic (related to data values out of context)
  - Contextual (related to pertinence to application)
  - Conformance (related to meeting defined expectations)
- Should be:
  - Measurable
  - Related to achievement of business objectives
- Measuring against data quality dimensions should provide a snapshot scorecard of the quality of data as well as guide improvement

Data Quality Dimensions - Examples

- Accuracy
- Provenance
- Syntactic Consistency
- Semantic Consistency
- Timeliness
- Currency
- Completeness
- Value/Attribute Consistency
- Conformance

Using Data Quality Dimensions for Baselining

- Assess relevance of data quality to business objectives and determine which dimensions are of value
- Assess the quality of data as it relates to each of the dimensions
- Correlate data quality measurements to business impacts

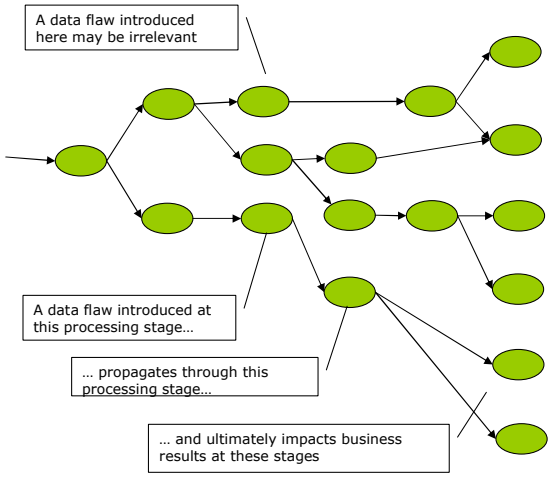
- Some suggestions:
  - Determine your technical needs for measuring these metrics at this time
  - Create a protocol for capturing measurements and tracking over time for the purposes of statistical process control

Policies and Protocols

- Root Cause Analysis
- The Data Quality Life Cycle
- Validation and Certification
- Data Stewardship
- Remediation and Manual Intervention

Cause and Effect(s)

The Information Flow Graph



Determining the value of fixing the process where the flaw is introduced must be correlated to the cost of the eventual business impacts.

But you also have to find out where the flaw is introduced!

Root-Cause Analysis

- Impacts are typically associated with the *discovery location* of a data quality problem
- In fact, one impact may be related to a combination of problems
- Alternately, a single problem may have multiple impacts
- A key to improving information quality is to identify the root cause of problems and eliminate them at their sources
- A key to managing information quality include:
  - Setting policies for data quality issue remediation
  - Establishing best practices for data management
  - Describing protocols and service level agreements for documenting, tracking, and eliminating data quality issues

The issue here is that in order to determine the costs to improve, one must first identify the problem. Identification and isolation of problems is the first step in determining how to attack them. Once a problem has been isolated, naming each one provides a rational to managing and controlling it.

Multiple impacts may be incurred by one problem, and multiple problems may contribute to a single impact. And since the impacts associated with poor data quality may not be felt until later stages of the information flow, part of the ROI/business case process is to determine what problems are actually related to specific business impacts.

Root cause analysis is critical to this aspect of information quality ROI, because this will help in determining what kind of investment is needed to alleviate the business impacts.

Root Cause Analysis

- Map information flow to understand how data moves across the organization
- Use the dimensions of data quality to define probing measurements to assess the quality of data at points along the information data
- When data failures occur, employ the probes to determine at which stage the flaw was introduced

Developing Metrics

- Develop metrics based on relationship of information to relevant business activities
  - Master reference information
  - Human capital productivity
  - Business productivity
  - Sales channel
  - Service level compliance
  - Vision compliance
  - Behavior
  - Risk

We can develop metrics based on the relevant business activities and how they translate to the bottom line. Since we are positing that poor information quality affects one or more of these aspects of the business, we should be able to baseline selected key indicators and measure how well we are doing with respect to those indicators.

Keep the following ideas in mind when thinking about metrics:

•Master reference information: Diagnostic codes, provider codes, procedure codes, customers, Loyalty

•Human capital productivity: Call center utilization, call center optimization, help desk, production effectiveness, “scrap and rework”

•Business productivity: Capacity planning, financial reporting, asset management, resource planning

•Sales channel: Marketing, opportunity identification, sales performance

•Service level compliance: provider and partner performance, timeliness of service, amounts of penalties

•Vision compliance: how well does the organization comply with its corporate vision

•Behavior: Trend analysis, member behavior patterns, anomalous behavior diagnosis, network/connectivity analysis

•Risk: Legislative/regulatory risk, investment risk

For example, if we believe that customer satisfaction is impacted by poor information quality, a key indicator may be related to the number, length, and topics of incoming call center calls. As another example, if poor information quality is believed to impact the precision with which prescription orders are filled, we can measure the number of duplicate items sent out based on customer reports on non-delivery.

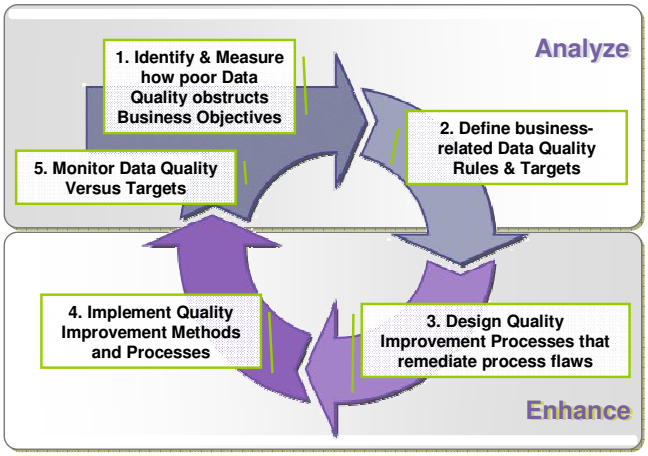
Key Information Quality Indicators

- A key indicator reflects a rolled-up summary of some important aspect of the current state of the organization's information quality
- Sample indicators:
  - Number of unique reference data objects (e.g., customers, vendors, products) vs. duplicate entries
  - Number of transaction "back outs"
  - Financial inconsistencies
  - Null or missing data values
  - Exposures to risk
  - …

Once a problem has been isolated and named, the next desire is to control it. In the words of Tom DeMarco, we cannot control what we cannot measure. Therefore, for each problem, we want to understand the key issues that characterize the problem as well as how that characterization relates to the business impact.

The concept of a key indicator is a measurement of some important characteristic that provides insight into the current state. The value of the indicator is that it is a metric that is bound directly to the problem, and the visualization of its measurement relates directly to an understanding of the impact.

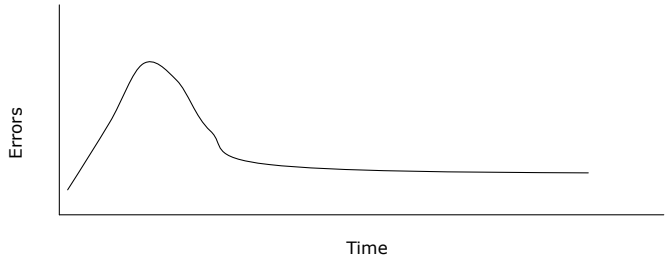
A Process-Driven Approach to Data Quality



 Source: Informatica

Data Quality Life Cycle

- Initially, many new issues will be exposed
- Over time, identifying root causes and eliminating the source of problems will significantly reduce failure load
- Change from an organization that is "fighting fires" to one that is building data quality "firewalls"
- Transition from a reactive environment to a proactive one facilitates change management among data quality clients



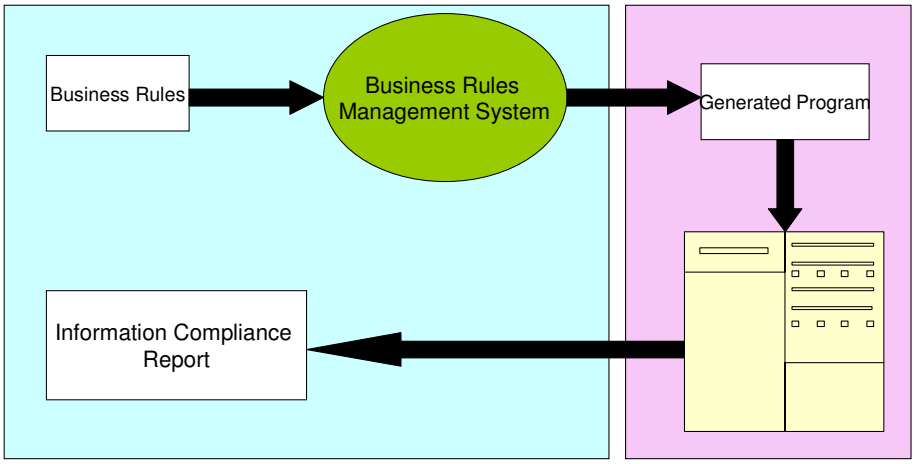
Data Validation

- Use business rules to assert data quality expectations
- Assertions can reflect expected measurements for dimensions of data quality
- Note:
  - Data quality tools may be useful for this process
  - If so, determine business needs and document them as a prelude to acquisition

Business Rules and Validation

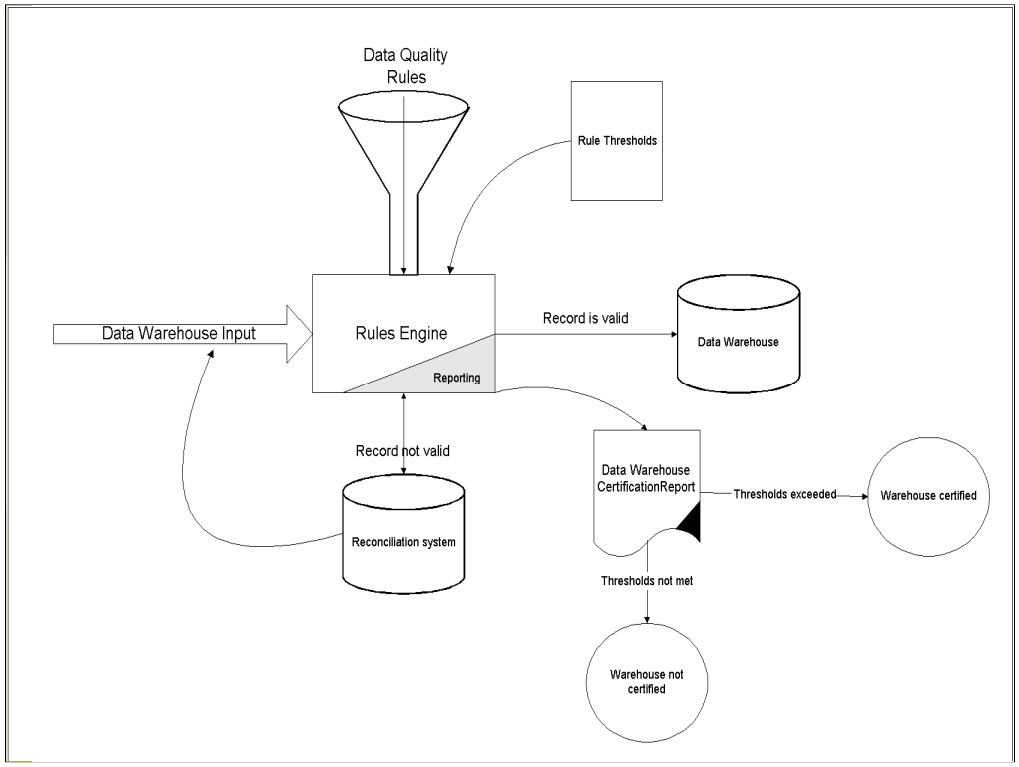
- Defined client expectations can be captured using a formal language
- Formal rules embed business logic for distinguishing between valid and invalid data
- The business logic can be easily transferred into program templates to create data validity filters
- Validity filters can be used for one-time probes or ongoing monitoring

Business Rules and Validation



Certification

- Ensuring that data meets defined expectation levels
- For each area of data quality dimension:
  - Assert expectations based on defined business needs
  - Identify an acceptance threshold
  - Ensure that the metrics can be measured
  - Assemble a set of processes to measure against business rules
  - Assemble a process to collect and report on levels of data quality



Here is a diagram showing how we incorporate a rules engine into a data warehouse validation system. The rules coupled with a rules engine are used as a filter at the point where records are streamed into the warehouse.

If the record is valid, it is forwarded directly into the warehouse. If not, the record is tagged with the rule (or rules) violated and is shunted into a reconciliation system. Records can be aggregated by violated rule, both for faster correction and for root cause analysis.

At all times, the filter maintains reporting information based on the number of rules violated, and the number of records successfully gated through. Comparing these numbers against input success thresholds will allow a determination of “certification” for the data in the warehouse, which can be used as a measure for warehouse users as to trustworthiness of the data!

Questions?

- If you have questions, comments, or suggestions, please contact me

David Loshin
301-754-6350
loshin@knowledge-integrity.com