Creating the Golden Record
Better Data through Chemistry

Donald J. Soulsby
metaWright.com
Agenda

• The Golden Record
• **Master Data**
  • Discovery
  • Integration
  • Quality
• **Master Data Strategy**
C. Lwanga Yonke
- Information Quality Practitioner

...Spewak advocated using data dependency to determine the ideal sequence in which applications should be developed and implemented: “Develop the applications that create data before those that need to use that data” (p.10).
Architecture Advocates

William Smith
   - Entity Lifecycle

Clive Finkelstein
   - Information Engineering – CRUD

Ron Ross
   - Resource Life Cycle Analysis
# CRUD in a Perfect World

<table>
<thead>
<tr>
<th>Entity Process</th>
<th>Customer</th>
<th>Customer Order</th>
<th>Customer Account</th>
<th>Customer Invoice</th>
<th>Vendor Invoice</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Receive CustomerOrder</td>
<td>R</td>
<td>C</td>
<td>CR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Process CustomerOrder</td>
<td>CRU</td>
<td>R</td>
<td>RU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maintain CustomerOrder</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td></td>
<td>RU</td>
<td>R</td>
</tr>
<tr>
<td>Terminate CustomerAccount</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td></td>
<td>RU</td>
<td></td>
</tr>
<tr>
<td>Fill CustomerOrder</td>
<td>RU</td>
<td>RU</td>
<td></td>
<td></td>
<td>RU</td>
<td>RU</td>
</tr>
<tr>
<td>Ship CustomerOrder</td>
<td>U</td>
<td>U</td>
<td></td>
<td></td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>Validate Vendor Invoices</td>
<td>U</td>
<td>U</td>
<td></td>
<td></td>
<td>R</td>
<td></td>
</tr>
<tr>
<td>Pay Vendor Invoices</td>
<td>U</td>
<td>U</td>
<td></td>
<td></td>
<td>RU</td>
<td></td>
</tr>
<tr>
<td>Invoice Customer</td>
<td>RU</td>
<td>RU</td>
<td></td>
<td></td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>Maintain Inventory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CRUD</td>
</tr>
</tbody>
</table>
Broadly speaking, materials scientists investigate two types of phenomena. Both are based on the microstructures of materials:

... 

ii. How do these microstructures influence the properties of the material (such as strength, electrical conductivity, or high frequency electromagnetic absorption)?

http://www.its.caltech.edu/~matsci/WhatIs2.html
Business VS Development Life Cycles

Zachman Framework for Enterprise Architecture

<table>
<thead>
<tr>
<th>WHAT</th>
<th>HOW</th>
<th>WHERE</th>
<th>WHO</th>
<th>WHEN</th>
<th>WHY</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTEXTUAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>List of Business Goals/Stat.</td>
</tr>
<tr>
<td>Scope</td>
<td>List of things</td>
<td>List of Processes</td>
<td>List of Locations</td>
<td>List of Organization Units</td>
<td>List of Events</td>
</tr>
<tr>
<td>Business Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOGICAL</td>
<td>Logical Data Model</td>
<td>Application Process Model</td>
<td>System Network Model</td>
<td>Human Interface Paradigm</td>
<td>Processing Structure</td>
</tr>
<tr>
<td>System Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHYSICAL</td>
<td>Physical Data Model</td>
<td>Application Structure Chart</td>
<td>Network Technology Model</td>
<td>Presentation Architecture</td>
<td>Control Structure</td>
</tr>
<tr>
<td>Technology Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OUT-OF-CONTEXT</td>
<td>Data Definition</td>
<td>Program</td>
<td>Network Components</td>
<td>Interface Components</td>
<td>Timing Definition</td>
</tr>
<tr>
<td>Components</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRODUCT</td>
<td>DATABASE</td>
<td>APPLICATION</td>
<td>NETWORK</td>
<td>ORGANIZATION</td>
<td>SCHEDULE</td>
</tr>
<tr>
<td>Functioning System</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

WHAT, HOW, WHERE, WHO, WHEN, WHY

CONTEXTUAL Scope

CONCEPTUAL Business Model

LOGICAL System Model

PHYSICAL Technology Model

OUT-OF-CONTEXT Components

PRODUCT Functioning System

DATABASE APPLICATION NETWORK ORGANIZATION SCHEDULE STRATEGY

META WRIGHT
Tibetan Proverb

“ If upstream is dirty, downstream will be muddy”
Data Warehousing - ETL
Transmutation
Bernard Trevisan, a 15th century alchemist, spent much of his life and a sizable fortune in search of the secret of turning base metals into gold, realized while dying was...

“To make gold, one must start with gold.”
DAMA Guide to the Data Management Body of Knowledge

Reference and Master Data Management: Managing golden versions and replicas
Data Management Body of Knowledge
Reference Data

Any data used to categorize other data, or for relating data to information beyond the boundaries of the enterprise. See master data.

Master Data

Synonymous with reference data. The data that provides the context for transaction data. It includes the details (definitions and identifiers) of internal and external objects involved in business transactions. Includes data about customers, products, employees, vendors, and controlled domains (code values).
Master Data Management (MDM)

Processes that ensure that reference data is kept up to date and coordinated across an enterprise. The organization, management and distribution of corporately adjudicated data with widespread use in the organization.

Reference & Master Data Management

Ensuring consistency with a “golden version” of data values. One of nine data management functions identified in the DAMA-DMBOK Functional Framework.
Top 13 MDM buzzwords

1. **Metadata** *(Don’t say metadata)*
2. Product information management (PIM)
3. **Enterprise master patient index (EMPI)**
4. Data governance
5. Customer data integration (CDI)
6. MDM hub
7. MDM architecture
8. "Collaborative" vs. "analytical" MDM
9. MDM return on investment (ROI)
10. MDM stakeholders
11. Enterprise hierarchy management
12. MDM metrics
13. **Data competency center**
The Quest

Bloor Research: the discovery of relationships between data elements, regardless of where the data is stored

dDIQ

DAMA: The degree to which data is accurate, complete, timely, consistent with all requirements and business rules, and relevant for a given use.

DAMA: The planned and controlled transformation and flow of data across databases, for operational and/or analytical use.
Atomic Data of the Enterprise

Provide and maintain a consistent view of an enterprise's core business information assets

ALCHEMY VS. CHEMISTRY
What is Master Data?

A Transaction is relationship of

**Master Data + Reference Files**

+ time stamp + volumes + secondary descriptors
Master Data - Classification

- Master Data
- Cross Reference Data
- Reference Data

- Customer Code
- AR Customer ID
- Inventory

- Status Code

- ISO 3166 Country Code

- Universal
- Enterprise
- Community
- Local

Operational  Analytical  Unified
Master Data Types

Operational Master Data
Definition, creation, and synchronization of Master Data required for transactional systems and delivered via service-oriented architecture (SOA).

Analytical Master Data
Definition, creation, and integration, including multiple historical versions, of Master Data required for Enterprise Reporting or Data Warehousing applications.

Unified Master Data
Master Data that is defined and created to apply to both Operational and Analytical applications. Implementation of the “single view of the truth”. It requires achieving agreement on a complex topic among a group of people.
Master Data Integration

Universal
Master Data that is created and maintained by an external organization. For reference data it is often developed by a standards organization such as the ISO. Master Data Models are often available to members of an industry trade or professional association.

Enterprise
Master Data that represents the common business information assets that need to be agreed on and shared throughout the enterprise.

Community
Master Data that is shared between two or more applications. Typically data is replicated from the application that contains the system of record for the Master Data, by federation (key linkage) or propagation (materialized views).

Local
Master or Reference Data that is created and maintained for a single operational application. Much of the Reference Data will remain local, such as Status Codes, for the transactions within the application.
Master Data - Content

Master Data

Transactional  Analytical  Unified

Definition

Operational  Technical  Business

Unified

Product  Place  Process  Period  Person
Atomic Primitives

PIM

PRODUCT

PROCESS

PURPOSE

PARTY

PLACE

PERIOD

CDI
Master Data Classification

Artifact

Person (Who)
- Internal Organization
- Customer

Product (What)
- Product Hierarchy

Place (Where)
- Geography

Period (When)
- Time Slice

Process (How)
- Business Activity
- Go To Market

Product (What)
- Service

Place (Where)
- Geo-Political

Period (When)
- Seasonality
MDM – Integration Techniques

**Data Propagation (Today)**
Copies Master Data from one location to another (materialized view).

**Data Consolidation**
Captures data from multiple data sources and integrates into a single Master Data hub.

**Data Federation**
Provides a single virtual Master Data view of one or more data sources. No data is stored in the Master Data hub.
Architecture choices

- Independent Data Mart
- Centralized Data Warehouse
- Data Mart Bus Architecture
- Hub & Spoke Architecture (Corporate Information Factory)
- Federated Data Stores
Are master data files metadata?
Need for central repository

Inside Metadata Repository

Outside Metadata Repository
Master Data Architecture – Repository

- Comprehensive, single source for Master Data
- Centrally managed, highly effective data governance
- Fixed data model – typically Customer & Product
- Extensive and sophisticated data integration process
- All applications need to change to conform to Hubs
- Expensive to modify to new business requirements
- Hard to adapt to rapid changes
Master Data Architecture – Registry

- Linkages to other data stores with transform logic
- Low Cost, small footprint, only unique key kept in DB
- Minimal disruption of current applications
- Limited Data Governance, no merged Master Data
- No resolution of semantic disintegrity
- No History, System of Record linked to Data Source
- Query performance - data availability from multiple sources
Master Data Architecture – Hybrid

- Relatively low cost solution
- Extensible Data Model based on industry templates
- “Best we have” Master Data Model
- Generalized governance & maintenance facilities
- ETL to load HUB may be complex
- Data redundancy may lead to update conflicts and data latency issues
Technology Adoption

"The Chasm"

Technology Adoption Lifecycle

Innovators
Early Adopters
Early Majority
Late Majority
Laggards

Area under the curve represents number of customers

GEOFFREY A. MOORE
Author of跨越the Trenches and Living on the Front Line
A BusinessWeek Bestseller

CROSSING THE CHASM
MARKETING AND SELLING HIGH-TECH PRODUCTS TO MAINSTREAM CUSTOMERS
HarperBusiness Essentials
Product Life Cycle

1. **Market introduction**
   - Costs are high
   - Slow sales volumes to start
   - Demand has to be created
   - Customers have to be prompted to try the product
   - Makes no money at this stage

2. **Growth**
   - Costs reduced due to economies of scale
   - Sales volume increases significantly
   - Public awareness increases
   - Profitability begins to rise
Market Strategy

- MDM Proof of Concept
- Other Subject Area
- More of same Subject Area

metaWright
By any other name...

Analytical Master Data
Subject Area Integration

PRODUCT-ID
Product Identification

POWER ASSISTED DATA RATIONALIZATION

Prod-ID
Char 9

PROD_ID
Packed 9

Install_base
Char 9

Prod Num
Packed 5

Unique Data Elements

Source Data Structures Field Occurrences
# Data Modeling vs Data Profiling

Data Modeling documents metadata or columns.

<table>
<thead>
<tr>
<th>Customer Name</th>
<th>Company</th>
<th>Address</th>
<th>Year Purchased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe Smith</td>
<td>Komputers R Us</td>
<td>Germany</td>
<td>1970</td>
</tr>
<tr>
<td>Mary Jones</td>
<td>Big Bank Co</td>
<td>10 Gulf Road</td>
<td>1999</td>
</tr>
<tr>
<td>Proful Bishwal</td>
<td>Little Bank Inc</td>
<td>Manhattan</td>
<td>prospect</td>
</tr>
<tr>
<td>Ming Lee</td>
<td>My Favorite Store</td>
<td>PO Box 987</td>
<td>10/31/1984</td>
</tr>
<tr>
<td>Giovanni Petra</td>
<td>(*#&amp;$JKLJ</td>
<td>10 Main St</td>
<td>2000</td>
</tr>
</tbody>
</table>

Data Profiling analyzes instance data or rows.
Traditional Meta-driven approach

*Design does not always match Reality!*

<table>
<thead>
<tr>
<th>Customer</th>
<th>Gen. Mot.</th>
<th>GM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9238423984</td>
<td>G-M 2342</td>
</tr>
<tr>
<td></td>
<td>**********</td>
<td></td>
</tr>
</tbody>
</table>
Data Discovery

• Profile data
• Discover primary-foreign keys
• Identify orphan rows
• Find overlapping columns

Data Profiler

• Import inferred structures from Data Profiler
• Visualize information in an intuitive data model
• Integrate with other modeling efforts

Data Modeler

metaWright
Data Profiling – Match, Merge and Cleanse

**Master Data Quality Analysis**
- Identify critical data elements
- Review the cardinality, range, mode, and null rules
- Review the value, pattern & length frequencies
- Review parent-child relationships
- Identify orphaned rows and values

**Cross-System Analysis**
- Identify reference data overlap between data sources
- Identify data content discrepancies between data sources
- Validate mappings to MDM target

**Master Data Error Types**
- Synonyms (False negatives) - two representations of the same entity are incorrectly defined as two separate entities
Appearing next month...

Why re-invent the wheel? that apply to well over 50 percent of most data model constructs and that can be reused. It is this 50 percent that I address in this presentation.

Paul Agnew, Universal Patterns, how they can help you build your data models

November 19, 2009
**Universal Models**

**Basic Party - Role**

- **Party**
  - **Person**
  - **Organization**

By Party Sub-Type:

- **Party Role**
  - 1:M

For each Party Role:

- **Party Relationship**
  - 1:M

- **Role Type**
  - 1:M

*The Data Model Resource Book: Len Silverston*
Architecture Models

- **Process based**
  Hierarchy mode

- **Topical or subject based**
  Network model
Data Governance

‘The act or process of directing, leading and assuring that information is managed effectively as an enterprise resource, including resolving information conflicts, across the enterprise.’

Larry English
Data Governance – Building Trust in Data

People

Data Stewards
Data Architects
Data Modelers

Process

Data Security / Access
Data Quality Assurance
Master Data Management
Metadata Management
Enterprise Model Management
Data Quality Assurance

Data Stewardship
Data Profiling/Cleansing
System of Record – (Definition)
  - Data Model
    Versioning
    Change Management
    Health and Welfare

System of Reference – (Lifecycle)
  - Workflow
    Synchronization and Integration
    Business Rules and Standards
    External Referencing
You can’t manage what you don’t measure!
# Scorecard – TDWI.org

<table>
<thead>
<tr>
<th>Data Quality Dimensions</th>
<th>Completeness</th>
<th>Conformity</th>
<th>Accuracy</th>
<th>Consistency</th>
<th>Timeliness</th>
<th>Uniqueness</th>
<th>Redundancy</th>
<th>Duplication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Element List</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obligor Identifier</td>
<td>100%</td>
<td>76%</td>
<td>94%</td>
<td>83%</td>
<td>95%</td>
<td>99%</td>
<td>97%</td>
<td>100%</td>
</tr>
<tr>
<td>Obligor Description</td>
<td>94%</td>
<td>85%</td>
<td>83%</td>
<td>96%</td>
<td>79%</td>
<td>99%</td>
<td>85%</td>
<td>86%</td>
</tr>
<tr>
<td>Agreement Identifier</td>
<td>80%</td>
<td>84%</td>
<td>84%</td>
<td>88%</td>
<td>87%</td>
<td>81%</td>
<td>79%</td>
<td>96%</td>
</tr>
<tr>
<td>Facility Default Grade</td>
<td>78%</td>
<td>76%</td>
<td>80%</td>
<td>77%</td>
<td>86%</td>
<td>95%</td>
<td>86%</td>
<td>87%</td>
</tr>
<tr>
<td>Obligor Default Grade</td>
<td>87%</td>
<td>90%</td>
<td>77%</td>
<td>97%</td>
<td>89%</td>
<td>87%</td>
<td>79%</td>
<td>94%</td>
</tr>
<tr>
<td>Probability of Default</td>
<td>100%</td>
<td>98%</td>
<td>93%</td>
<td>77%</td>
<td>75%</td>
<td>83%</td>
<td>96%</td>
<td>81%</td>
</tr>
<tr>
<td>Loss Given Default</td>
<td>87%</td>
<td>91%</td>
<td>83%</td>
<td>77%</td>
<td>86%</td>
<td>90%</td>
<td>97%</td>
<td>88%</td>
</tr>
<tr>
<td>Total Exposure Amount</td>
<td>79%</td>
<td>94%</td>
<td>90%</td>
<td>80%</td>
<td>76%</td>
<td>91%</td>
<td>89%</td>
<td>88%</td>
</tr>
<tr>
<td>Expected Loss Amount</td>
<td>97%</td>
<td>95%</td>
<td>98%</td>
<td>91%</td>
<td>94%</td>
<td>97%</td>
<td>79%</td>
<td>83%</td>
</tr>
<tr>
<td>Maturity Date</td>
<td>88%</td>
<td>82%</td>
<td>92%</td>
<td>81%</td>
<td>81%</td>
<td>88%</td>
<td>96%</td>
<td>96%</td>
</tr>
<tr>
<td>Collateralized Exposure</td>
<td>88%</td>
<td>100%</td>
<td>84%</td>
<td>96%</td>
<td>79%</td>
<td>92%</td>
<td>89%</td>
<td>84%</td>
</tr>
<tr>
<td>Collateral Acquire Date</td>
<td>88%</td>
<td>86%</td>
<td>82%</td>
<td>87%</td>
<td>95%</td>
<td>95%</td>
<td>81%</td>
<td>84%</td>
</tr>
</tbody>
</table>
Data Stewardship

- **Difficult to assign, not process based**
- **Cross-functional**
- **Not a Single DATA Steward**

Business, Technical, Operational Accountability, Authority, Responsibility

<table>
<thead>
<tr>
<th>Rows</th>
<th>Cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columns</td>
<td></td>
</tr>
</tbody>
</table>
Master Data Stewardship

Difficult to assign, not process based
Infrastructure — cross-functional
Business stewards - Content Descriptions
Technical stewards - Structure Codes and structure
Data Management Body of Knowledge

- Data Architecture Management
- Data Quality Management
- Data Development
- Database Operations Management
- Data Security Management
- Meta Data Management
- Document & Content Management
- Data Warehousing & Business Intelligence Management
- Reference & Master Data Management

Data Governance

© 2008 DAMA International
References

Managing Reference Data in Enterprise Databases: Binding Corporate Data to the Wider World
Malcolm Chisholm - Morgan Kaufmann

The Data Model Resource Book: A Library of Universal Data Models for All Enterprises (VOL. 1,2)
Len Silverston - John Wiley & Sons

The Data Model Resource Book: Universal Patterns for Data Modeling
Len Silverston - John Wiley & Sons

Master Data Management & Customer Data Integration for a Global Enterprise
Alex Berson & Larry Dubov - McGraw Hill
Thank You!

Donald J. Soulsby
DSoulsby@MetaWright.com