Creating Simplicity out of Complexity

Achieving End-to-end Metadata Lineage – Best Practices Learned

January 16, 2014
Gordon DeGrandis, Senior Consultant
Gordon.degrandis@adaptive.com

Copyright© Adaptive, Inc. 2014
Confidentiality & Disclosure Agreement

This is an unpublished work, the copyright in which vests in Adaptive, Inc. (“Adaptive”). All rights reserved.

The information contained herein is Copyright© and the property of Adaptive, Inc. and is supplied without liability for errors or commissions. No part may be reproduced, disclosed or used, except as authorized by contract or other written permission. The copyright and the foregoing restriction on reproduction and use extend to all media in which the information may be embodied.

All product names used herein are for identification purposes only and may be trademarks of their respective companies.
Metadata and Data Lineage is the way to track data movement between data sources and destinations across different tools, environments, and technologies. Data Lineage answers the questions that many organizations are facing today due to regulatory compliance, impact analysis and data quality efforts. Metadata lineage provides the means to merge business and technical requirements for traceability, lineage and impact analysis for the enterprise. The process to build Metadata lineage can be long but the results can be impressive. This session will highlight the best practices of completing a data governance project achieving end-to-end data lineage.
1. A lineage project – best practices, where to start
2. Definitions - Business glossary / information map, what is lineage, what is traceability
3. Updating the Information Model – Governance
4. Bringing in the metadata sources, Business Terms, Design models, Actual databases
5. Documenting which Business Terms are contained in which data sources, which databases a Business Term exists in
6. Stitching and Establishing lineage – tying it all together
7. Adding value - Being able to relate Business Terms back to regulatory objectives and risks, tracking which data are being used and which aren't, enabling users to 'tag' Business Terms or data sources with their views on value, quality, etc., and tracking data quality metrics
8. Summary - Benefits of effective metadata management
What is lineage

• How information moves through different systems and databases that create, manage, and use data
  – So it’s often about relating the same types of item
  – E.g. Table to Table, but also Table to XML Element to Dimension
What is lineage
What is lineage – example summary lineage
What is lineage – example
complete lineage
What is lineage – Why does it matter

• Impact analysis
  – If a source data item changes what will be affected downstream
  – Direct or indirect

• Focus data quality efforts

• Starting from critical business reports, work back to where the data originates

• Is the right source being used?
Starting a lineage project

• Start small then build on success
  – Old saying but still true today
  – Pick a business area of metadata that has a small scope with high visibility
  – Typically BI reporting is a good choice
  – Give preference to business area that uses industry tools

• Limit expectations, this is a complex process
Select which type of lineage project

• Technical lineage project
  – Minimal business involvement
  – Typically an IT project
  – Provides end to end lineage but no references to the Business Glossary
Select which type of lineage project

- Lineage project with Business Glossary
  - Business involvement
  - Typically includes many more stakeholders
  - Provides traceability from the Business Glossary to a report field
  - Provides the same end to end lineage as Technical lineage project
Select which type of lineage project

- This presentation is based on the project type ‘Linage project with Business Glossary’
- Differences between the project types are
  - Usage of a Business Glossary
  - Usage of Design models
  - No mapping for traceability between the Business Glossary, Design Models and actual database schema
Produce a project plan with Phases

• Project broken into phases
  – Planning and inventory of metadata
  – Planning and implementing the repository configuration
  – Validating and Populating the business glossary
  – Populating the Design Models and mapping to the business glossary
  – Populating the Actual database schema and mapping to the physical design models
  – Importing the tool metadata and stitching to Actual database schema
  – Keeping the repository up to date, ever greening
## Example Project Plan

<table>
<thead>
<tr>
<th>Name</th>
<th>Adaptive Metadata Repository Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Survey information sources (and integration possibility)</td>
</tr>
<tr>
<td></td>
<td>Survey information tool versions</td>
</tr>
<tr>
<td></td>
<td>Survey organization and stewardship processes</td>
</tr>
<tr>
<td></td>
<td>Survey and validate non-functional requirements (i.e. security, performance)</td>
</tr>
<tr>
<td>Installation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Install Adaptive</td>
</tr>
<tr>
<td></td>
<td>Configure LDAP/SSO and SMTP integration</td>
</tr>
<tr>
<td></td>
<td>Install Adaptive Metadata Integrator Module</td>
</tr>
<tr>
<td></td>
<td>Metadata test loads and Excel spreadsheets</td>
</tr>
<tr>
<td></td>
<td>Identify organization/workforce data and extract</td>
</tr>
<tr>
<td>Requirements Workshop</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Requirements Specification and Architectural Design Workshop</td>
</tr>
<tr>
<td></td>
<td>Define User Acceptance Criteria and Measurement</td>
</tr>
<tr>
<td>Full Design</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Produce Metadata Information Architecture</td>
</tr>
<tr>
<td></td>
<td>Define application presentation aspects</td>
</tr>
<tr>
<td></td>
<td>Define store architecture and security model</td>
</tr>
<tr>
<td></td>
<td>Define user group profiles</td>
</tr>
<tr>
<td></td>
<td>Define roles security and access control profiles</td>
</tr>
<tr>
<td></td>
<td>Define metadata report queries</td>
</tr>
<tr>
<td></td>
<td>Define stitching/weaving metadata strategy</td>
</tr>
<tr>
<td></td>
<td>Define workflow for metadata management</td>
</tr>
</tbody>
</table>

| Implementation | | |
| | Import organization/workforce and link to LDAP/SSO |
| | Create metadata versioning and security model (for workflow) |
| | Implement workflow |
| | Define transformation processes |
| | Import and link metadata |
| | Import and link data dictionary and assign stewardship |
| | Configure User Interface based upon User Groups |
| | Configure Access Control based upon User Groups |
| | Create Reports |
| | Implement navigation maps |
| | Integration and Testing |
| | Integrate and test |
| | Validation and Production Deployment |
| | Pre-production test deployment |
| | Validation Testing |
| | Production deployment |
Best Practice

• Partner with the Metadata repository vendor
  – To capture project requirements
  – Identify the metadata sources and tools
  – Identify any gaps that might exist in the metadata
  – Start with basic capabilities of the metadata repository – minimal customization
  – To produce realistic project schedule
Lineage Project Benefits

• Metadata information not lost with people
  – Collaborative capabilities allow people to enrich the metadata
  – Information that was local can be merged into the repository for everyone

• Enterprise Collaboration
  – Repository users can have discussions, add artifacts, create assessments or add a subscription
Lineage Project Benefits

• One stop metadata access - metadata from different tools in one place
  – Importing tool metadata into the repository enables access to the tool metadata
  – No need for users to have different tools to view tool metadata
• Merges silos
  – Similar to one stop metadata access
  – Silos created by tools not being able to collaborate
• Sharing metadata where it matters
  – Only need a Web browser for user access, no clients, no software to install
  – People and tools have access to the metadata
Lineage Project Benefits

• Enterprise wide metadata governance
  – Being able to govern metadata for the enterprise

• Regulatory metadata requirements can be addressed
  – Lineage and traceability of the metadata can answer nearly any requirement
  – Reduced time for preparing regulatory documents
DEFINITIONS
Business Glossary

• Definition of attributes, entities or relationships or indeed any words or phrases used by the business in business terms
• Can contain business rules, domains or list of values
• Terms have Data Owners and possibly Stewards
  – These are used for the governance process
• Typical sources are Excel workbooks Access Database or purpose built application
Data Owner and Stewards

• Data Owner
  – Person, that approves changes
  – in this presentation approves changes to the business glossary
  – Is not an organization

• Steward
  – Accountable for quality of data definition
  – A person, that approves changes to data definitions
Information Map

• The metadata map used to navigate the “information Ocean”
  – Provides linkages to metadata items and their relationships

• Diagram that provides information for users of the metadata repository

• Can be used as a clickable diagram for navigation of the metadata repository
Example Business Catalogue Information Map

DATA GOVERNANCE BUSINESS GLOSSARY

- My Subscriptions
- My Roles
- My Information
- Classification Matrix
- Help
- Object Counts
- Catalog Owners
- Terms with No Rules
- Terms not Related to EIP
- Advanced Search
- Abbreviations
- My Workflow Messages
- My Active Workflows
- My Workflow History

Business Catalogs
- Search
- Create

Owned or Used By

Business Terms
- Search
- Create

Exist As

EIP Warehouse Data
- Search by Entity
- Search by Table
- Search by Attribute
- Search by Column

Govern

Business Rules
- Search
- Create

Classified By

Rule Sets
- Search
- Create

Governed By

Own or Use
Information Map
Example Diagram

Metadata Repository

Portfolios
- Business Infrastructure
- Care Delivery
- Health Plan
- IT Infrastructure
- Facilities Svcs.

Conceptual

Deployed

Project

Business Concept

Business Term/Metric

e.g. Care Delivery POC

e.g. Hospital Services

e.g. Patient Administration

e.g. Encounter Type

Portfolio

Sub-Portfolio

Data Field

View

Table

Schema/Database

ETL Job

ETL Transformation

e.g. ENC_TYPE_C

e.g. PAT_ENC

e.g. HCCNLC

Report Object

BI Universe

Report
What is Traceability

• Traceability is how more implementation-oriented items trace back to more business-oriented items as part of a development lifecycle
  – So it’s generally about relating different types of items
  – E.g. Table to Entity to Business Concept
Why Traceability

• The Lineage project with Business Glossary requires the use of traceability
  – Mapping the Actual Schema to the Design model
  – Mapping Design model to the Logical model
  – The Logical model to the business glossary
What is Traceability

Producers and Consumers of a Business Glossary

Subject Matter Experts
Define the P&C Concepts, Rules

Legal, Compliance

Business Process Modelers

Business Analysts,
Data Modelers,
DBAs, Data Integration,
Data Quality, ETL Users

Business Glossary, Rules

Application, Service Developers

Vendors, System Integrators, Partners...

Business Glossary <> Information Models

Business Glossary, Rules

Conceptual Model / Ontology

Logical Data Model (LDM)

Relational Model

Dimensional Model

XML Schema Model

Traceability

DDL

DDL

XSD

OWL, RDF...

XMI

Copyright© Adaptive, Inc. 2014
Example – Term to actual database schema
What is Traceability - Why does it matter

• Impact analysis
  – If the business requirement changes what will be affected?
• System of record
  – What’s the best database table to get hold of a customer’s address?
• Continuity across different tools and technologies
  – The same business meaning should apply whether in a database or an XML Schema or a program
• Business visibility of technology
  – How something is implemented
• IT visibility of business meaning
  – What does a column mean?
Using Traceability and lineage

• Provides starting points for users to browse lineage from a business perspective
Governance

• Do you really want to allow anyone to update your metadata?

• Provides auditing and approvals of changes
  – To the Business Glossary
  – Relationships to the Design and actual data models

• Based on Data Ownership or stewards
  – The ultimate approver of any changes
  – Can be different for data and business glossary
• Can also be used when importing tool metadata
  – Governance is much simpler than for the business glossary
  – Work with vendor to demonstrate what is possible
  – Is similar to source control using check out/check in
  – Can work with or without approvals
  – Allows for tool metadata validation and comparison before making the tool metadata available
• **Best Practice**
  
  – Setup a governance process based on project requirements to protect your organization’s investment
  
  – Work with stakeholders and vendor to demonstrate what is possible
POPULATING THE REPOSITORY
Metadata Inventory

• Best practice
  – create an inventory of metadata sources based on project requirements

• Capture
  – Metadata type – database, ETL, BI, business metadata
  – Ownership and development information stakeholders such as DBAs, project managers, developers, business owners
  – Configuration information such as log ins, passwords, schema names, hostnames, database names
  – For metadata in file form – capture file location, source control location, passwords, hostname
  – Development Life cycle frequency (day, month, quarter etc.)
## Example Inventory

<table>
<thead>
<tr>
<th>App#</th>
<th>App Name</th>
<th>Short Name /Data Store or ETL</th>
<th>Hostname</th>
<th>Database</th>
<th>Schema</th>
<th>Port #</th>
<th>Service Name</th>
<th>User ID</th>
<th>Password</th>
<th>Instance</th>
<th>File Location</th>
<th>AppManager</th>
</tr>
</thead>
<tbody>
<tr>
<td>5670</td>
<td>Finance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FINAPP - Oracle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FINAPP - PL/SQL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FINAPP - S-Plus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FINAPP - Unix Shell Scripts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FINAPP - DataStage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55332</td>
<td>FDW (FINANCE DATA WAREHOUSE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FDW - Oracle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FDW - PL/SQL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FDW - Unix Shell Scripts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55889</td>
<td>Gateway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FIG - Data Store Not Listed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FIG - PL/SQL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>FIG - Informatica</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Best Practice

- Best practice
  - Get agreements for access to metadata from all stakeholders in the inventory
  - Consult with database DBAs for access to databases
  - Verify that all metadata sources are accessible and take action if they are not
  - Provide regular updates to stakeholders on progress
Order of import

• Using the metadata inventory populating the repository can be started

• Top Down order
  – Business metadata
  – Design models
  – Actual database
  – Tool metadata for ETL
  – Tool metadata for Business Intelligence
Business Glossary

• Business Glossary
  – Work with repository vendor to provide import templates
  – Usually imported using Excel workbooks

• Best practice
  – If more than one business glossary import them separately into different glossaries they can be reconciled once in the repository using the Governance process
  – Document the source of the business glossary and the stakeholder
Logical/Physical design models

• Logical and physical Design models populated from industry tools
  – CA Erwin, Embarcadero ER/Studio, PowerDesigner and etc.
  – Only of value if they are maintained

• Best practice
  – Use the latest available design models that match the actual running databases
Operational Databases

• Operational databases
  – Parse database DDL
  – Access via direct access to the database metadata
  – Relational Views are most important since they can be the start of lineage

• Best Practice
  – Import the operational database metadata based on the latest design model
Tool metadata

- Using the tool metadata inventory process the remaining tools that have metadata.
- XSD – XML Schema
- Business Intelligence
  - Tools such as Business Objects and MicroStrategy require network access for metadata access
  - Some tools require parsing programs
  - Software installation for access via an API (Business Objects and MicroStrategy)
Tool metadata

• Using the tool metadata inventory process the remaining tools that have metadata.
• ETL tools such as Datastage, Informatica, Microsoft Integration Services
  – Most of these tools are file based
  – These tools might require a variable substitution file that assigns a value to variable in order to replace variables by real values
  – These tools require a higher level of knowledge as they are much most complex
Tool metadata

• Other tool metadata sources that provide lineage
  – Source code scanners for programming languages such as
    • COBOL and mainframe sources (JCL, Copybooks, IMS Database)
    • PL/SQL
    • ELT (Shell scripts and SQL statements)
STITCHING AND ESTABLISHING LINEAGE
Establishing lineage – Tool metadata

• This project task will populate the repository from metadata tools that use databases in the inventory
  – Tools could be ETL, BI, Reports, application source code
  – Provides standalone lineage for each tool

• Best Practice
  – Capture the tool version in the inventory list
  – Get familiar with the tool to understand the metadata it is providing
Establishing - importing

• Lineage is usually imported
  – More frequently than for traceability
• Typically from the tool responsible for the movement
  – Informatica PowerCenter, IBM Datastage, Microsoft Integration Services and other data warehousing tools
  – BI and Reporting such as Microsoft Reporting Services, SAP BusinessObjects, Cognos and other BI and reporting tools
Establishing lineage – other means

• May need to use other means if the movement technology is not integrated e.g. a bespoke program
  – Excel workbook with lineage mapping
  – Transforming XML files to provide lineage mapping
Stitching and Establishing lineage

• For this project task a stitching operation will be performed to stitch the tool metadata to source and target metadata
  – Sources/target are usually relational tables, views, record files, XML schema

• Stitching is a process that will stitch standalone lineage into end to end lineage
Stitching Process

Steps:
1. Import Source Schema
2. Import Target Schema
3. Import ETL tool mapping to source and targets (specifying source and target instances)
Stitching Process

• Before the stitching process
  – Tool metadata contains standalone database items (tdb1)
  – Tool lineage references standalone database items

• The stitching process
  – Match tool standalone database items to repository database items (schema, Table, View, Column)
    (Match tdb1 to exiting gdb1)
  – Once matched delete the tool standalone lineage information (tdb1)

• After the stitching process
  – Tool metadata references repository database items (gdb1)
  – All tools that are stitched refer to the same repository database items
  – End to end lineage is established
Establishing lineage and stitching

• Best practices
  – Verify the stitching operation to validate all sources/targets have been stitched
  – Stitching that is not complete is usually source/target or metadata tool that is not at the same release (newer schema, older tool or older schema newer tool)
Establishing lineage and stitching

• Best Practice
  – As each metadata tool is stitched record the tool source/target name
  – Tool database names do not always match names for source/target data
    • Record these names
  – Record any exceptions or workarounds in the project document or the repository
DOCUMENTING THE METADATA
• Documenting business glossary items to their relevant data sources
  – Resolves any ambiguity when mapping to attributes in the data environment
  – Can be a difficult and labor intensive task if no mapping documents exist
  – Requires expertise with the business domain expert of the data and Business Glossary (which attribute to which business Term?)
  – Mapping information is usually a human intensive activity
  – Mapping can be cross checked by using the lineage information present in the repository
    • Lineage should display the target field contained within a BI Report.
Business Glossary

• Best Practice
  – Allocate enough time in the project plan for this work
  – Work with stakeholders for planning time and access to subject matter experts.
  – Work with vendor to make available a mapping template
  – Work with Logical Design models – Will save time and money
  – Take advantage of Logical Design models to use automated mapping for name matching entity/attribute to Business Glossary
Using the metadata

• Goal in this part of the project is to provide a navigation layer for business users
  – Access to metadata via language business users understand
  – Reduce complexity by offering one stop lookup of business related metadata

• Advantage for technical users
  – The inverse relationships exist
  – Access to business metadata via technical metadata that technical users understand
  – Open up the business view to technical users
Using the metadata

• Provide a feedback mechanism for users via the repository to comment and provide additional metadata information
Evergreening – keeping things up to date

• Keeping the metadata repository up to date
• Updating the repository
  – Updated metadata tool files and sources
  – Updated database sources
  – Development releases
  – New tool files and sources
Ever greening – keeping things up to date

• Manual or Automatic
• Triggering Event
  – Scheduled imports (e.g. batch command file run every weekend)
  – Model update/save (repository updated only when model changed, or when model changes complete change control process)
• Push or Pull
  – Import initiated by modeling tool (e.g. on save of changes)
  – Import process reads from metadata tool repository (e.g. ERwin modelmart, Business Objects and Informatica repositories)
  – Metadata tool saves to file store location, import process polls for updated files to be imported
• Best Practice
  – Plan ahead for ever greening when preparing the metadata inventory
WHAT NEXT
What Next

• Managing success
  – Increasing number of users
  – Increasing number of item searches
  – Feedback from repository users
• Add more metadata sources with more complex architecture
• You have experience now from the first project
• Follow the same process
• Provide regular updates and keep stakeholders informed
Collaborative Capabilities

• Make use of the collaborative capabilities of the repository
  – Increases user collaboration
  – Makes for a better repository

• Capabilities are
  – Email notifications
  – Adding artifacts to repository items (Documents, PDF files, Excel workbooks)
  – Assessments for repository items
  – Discussion groups
  – Subscriptions and notifications
• Add versioning for lineage
  – can be used to validate releases
  – Releases are isolated until validation takes place
  – Validation can use lineage to validate table/column changes
    • Impact analysis using lineage will help identify changes that could affect users of the items that changed
Workflow

• Add a Workflow to control repository versioning with an approval workflow
  – Works with the versioning to allow only certain users to initiate a metadata refresh workflow
  – Workflow works very much like check out/check in with source control
  – Assigns an approver to approve the version once it has been validated, a version can also be rejected
• Controls the visibility of new metadata releases in the repository
• Versions can be Compared with previous versions
ADDING VALUE
Adding value

• Use the rich repository content to advantage
  – Regulatory objectives
  – Tracking metadata usage

• Get users involved
  – Track users views on value and quality
  – Track data quality
  – Enrich the business metadata
Adding value

• Enable users to 'tag' metadata such as
  – Business Terms
  – Data sources

• with their views on value, quality, availability and documentation
Summary

• End to end lineage of metadata from different metadata tools in a repository is not an easy challenge, the road can be long and difficult
• Choose your first project wisely
• Document everything
• Get stakeholders involved
• Keep everyone informed
Summary

• Benefits
  – Metadata information not lost with people
  – Enterprise Collaboration
  – One stop metadata access - metadata from different tools in one place
  – Merges silos
  – Sharing metadata where it matters
  – Enterprise wide metadata governance
  – Regulatory metadata requirements can be addressed
THANK YOU