White Paper

Building an Enterprise Metadata Repository

Ron Lewis, CDO Technologies

December 2009

Corporate Headquarters
100 California Street, 12th Floor
San Francisco, California 94111

EMEA Headquarters
York House
18 York Road
Maidenhead, Berkshire
SL6 1SF, United Kingdom

Asia-Pacific Headquarters
L7. 313 La Trobe Street
Melbourne VIC 3000
Australia

CONTENTS

INTRODUCTION

The importance of capturing metadata has been a topic of many webinars, teleconferences, and white papers over the last several years. There's has also been an increasing emphasis on "building metadata repositories". To avoid being just another white paper describing the "metadata trend" or promoting a particular metadata repository solution – the intent of this whitepaper is to provide basic definitions for the concepts of metadata and metadata repositories, as well as to provide a basic methodology for collecting metadata and populating an enterprise metadata repository.

IT ALL BOILS DOWN TO THE DATA

A good starting point is defining the difference between data and information. Data is nothing more than collections of raw facts. *Data* is transformed into *information* as result of the manipulating or processing of these "raw facts" by putting them into meaningful contexts. Data is often referred to as the "crown jewels" of an organization. This is a valid analogy; an organization's data is typically priceless to that organization. The financial health and profitability is often tied to how well a business entity utilizes its data resources.

As technology has evolved, executive leaders have become increasingly more aware of the importance of data to the efficiency of an organization. A lot of emphasis is being placed on methodologies and frameworks—a couple good examples would be LEAN Engineering Principles and the Zachman Framework—to provide additional insight into how to streamline a company or business' effectiveness. There's a heavy focus on mapping business processes and to the corporate data upon which the processes rely. This focus introduced the need for metadata management.

METADATA PROVIDES CONTEXT

Metadata describes data and is used to enhance the effectiveness of data use. Metadata is often categorized as either business or technical metadata. Business metadata describes taxonomies, articulates business rules, and establishes common vocabularies. This helps align data with its business context. Conversely, technical metadata describes data sources, attributes, domains, nomenclature, movement, and consumption rules. The bottom line: Metadata is used to identify the context in which data becomes meaningful.

METADATA REPOSITORIES STORE ALL YOUR PROVERBIAL EGGS IN ONE BASKET

When it comes to metadata, having all your proverbial eggs in one basket is a good thing. Metadata, especially when aggregated properly, can expose new ways of exploiting data to significantly increase efficiency and profitability. In order to glean the full value from collected metadata it's important to store metadata in a manner where it can be easily indexed, cataloged, and searched. A metadata repository facilitates this. A metadata repository is a system for aggregating, indexing, cataloging, protecting and providing access to corporate metadata.

This is no small task. Storing metadata collectively is challenging because there are many different types of metadata, and many different means by which metadata can be expressed.

As stated above, there are two basic categories of metadata: business and technical. Below are a few examples of crucial metadata and a few of the forms by which it is normally expressed. This is by no means an exhaustive list, but is provided to help scope the metadata collection task.

Crucial business metadata:

- Business Vocabularies: describe terms common to the organization and are built of Business Definitions
- Business Definitions: provide a common meaning to a common term
- Business Processes: groups of business activities that center around a business purpose and governed by business rules
- Business Rules: define the organization and how it achieves it business goals

Critical technical metadata:

- Data Models
- System Catalogs
- Extract, Transform, and Load Scripts
- Data Lineage
- Data Consumption Rules

A few common means by which metadata is articulated:

- Use Cases: Often used to describe business processes in software development terms
- Business Models: Used to facilitate communication between business and systems analysts
- Data Models: Used to describe relationships between data elements

A metadata repository solution should be capable of collecting all of these bits of data in a readily searchable, protected form. Quick rule of thumb concerning metadata repository security: The value of the metadata is proportionate to the perceived quality and reliability of the metadata repository contents. The metadata is incredibly valuable and should be adequately protected.

First Step: Decide What Needs to be Collected

There is obviously lots of metadata that can be collected and housed in a repository. Knowing what needs to be collected is definitely a huge challenge. Deciding what needs to be collected is viewed as an enterprise engineering task. Enterprise architects typically describe an organization using a formal, structured framework, such as Zachman, to define the "why, how, what, who, where, and when" associated with an enterprise. This is an expensive and time consuming endeavor but is also well worth the investment.

Reverse Engineering Data Collection Needs From Key Reporting Requirements

A simple method for getting started for diverse, well-structured organization is to iteratively collect data associated with the business processes that define the enterprise reverse engineering from the reports that the enterprise relies upon. For example, a medical practice can be decomposed into several large building blocks—such as doctors, patients, visits, and perhaps treatments. A medical practice requires several different types of key reports, such as billing, treatment record, medical record requests, and HIPAA release forms-- to name a few (disclaimer: this is only for illustrative purposes). Start with the most frequently used "reports" and identify the information necessary to complete the tasks associated with the large building blocks. Reports encapsulate critical business rules, and key business attributes. A lesson learned from the security industry is *reports often put business data in its most meaningful context and are prime targets of cyber-attackers*. This method for determining metadata needs leverages this security knowledge. Since reports provide easy-to-understand, readily accessible mappings between business needs and the supporting data--the answer to these two questions: "What information do the reports describe" and "what is necessary to build them" provide a good starting point for identifying the most important metadata to be collected.

Second Step: Collecting the Metadata

The Traditional Method

Many enterprise architects start with building business process models, and then map the business process models to conceptual data models. This requires business architects to interview key business analysts to derive and validate the business models. It also requires data architects to interview key technical personnel to build and validate the conceptual data models. Once the models and mappings have been built—enterprise architects then map these to systems and data repositories identifying key dependencies between systems. The correlating metadata defining the AS-IS is then collected, cataloged, indexed, and housed in a metadata repository.

Key Challenges

Managing Cost: Initial metadata collection activities associated with the manual effort for describing the existing environment (e.g., the "AS-IS") are expensive, both in terms of cost and time. Manual interview processes are disruptive. Interview processes have a hidden cost in lost production; they pull key personnel away from revenue generating tasks. In rapidly evolving, high-demand environments—this loss can be devastating.

Managing Scope: There are two common scope-related tendencies during initial data collection efforts: to focus on collecting everything, or to get buried in the details. Both of these cause a loss of momentum and end up increasing cost significantly.

Staying focused on the data collection: It's tempting, especially when glaring inefficiencies are discovered, to stop the data collection activities and focus on addressing and correcting the discrepancies. This can have huge negative impact due to the ripple effect changes can cause across an organization.

SOLUTIONS

Summing up what seems to work best in most environments to counter these challenges:

- Focus on what's important. Start with the organization's main focus and then quantify each activity based on the percentage of revenue for the organization.
- Leverage automation to the greatest degree. Eliminate as many of the manual interviews as possible and glean as much business information by decomposing reports, reverse engineering applications, and current software development documentation. Caution: Not all software development artifacts will match what's been deployed into production.
- Do not perform data re-engineering or process improvement during data collection. Focus the effort on data collection. When all the metadata has been collected, cataloged, and indexed, business analysts will often get a clear picture of inefficiencies across the enterprise.

A BETTER WAY – USE TOOLS

A prime solution to reduce the time and cost while increasing the overall effectiveness of the data collection activities is to leverage technology. It's best to start with the reports and work backwards towards deriving processes. The reports and associated metadata collected provide a good means for validating the processes, provide a common frame of reference for all involved in the interview process, and minimizes number of interviews necessary to fully capture the enterprise view.

Perform the following actions using tools:

- Identify as many of the database servers within the enterprise and reverse engineer the physical schemas contained within each database server.
- Build a single enterprise data model and include each reverse engineered physical schema in the master model.
- Export the schema metadata to a spreadsheet and have the appropriate data architects annotate what is known about each entity and associated attributes.
- Instrument and profile each application with an application protocol monitor to capture business rules captured as programmatic logic-thereby capturing the data consumption rules.
- Extract business logic from application source code
- Capture transformation rules and data lineage metadata by profiling extract, transform, and load (ETL) scripts. This is accomplished by instrumenting each database server and capturing use with a database profiling tool.
- Import any existing business process models (BPM) into a standard BPM tool.
- Profile data use, focusing on identification of dormant data as well as high-utilization data—databases, tables, and elements.
- Evaluate the metadata captured for consistency. Identify and annotate any discrepancies discovered.

Third Step: Populating the Metadata Repository

Once the above steps have been completed, there will be a significant amount of metadata. To glean the value from the metadata, it needs to be imported in a common repository, cataloged, indexed, and made available to the appropriate business, system, and data analysts.

There are plenty of pre-built metadata repositories. The metadata repository solution selected must have the following critical capabilities:

- Manages unstructured content.
- Provides appropriate access controls and auditing.
- Imports myriad formats and exports in a standard format such as XML.
- Allow the creation and associations of descriptions to each metadata asset.
- Provides a searchable and intuitive means for finding and evaluating the appropriate metadata.

The steps for populating a metadata repository will vary based on the particular solution; however, the methodology is pretty standard. Each solution should provide import and export mechanisms that allow data bridging between metadata collection tools. For collection tools not natively supported by the repository solution, a commonly understood format, such as XML, can be used as an intermediary bridge. For example--if metadata describing data nomenclature has been captured in a loosely structured tool, such as Microsoft (MS) Excel, the metadata can be exported as a comma-separated-value (CSV) file and imported back into a supported data modeling tool. Most data modeling tools allow users to export metadata in XML format. While an MS Excel file would most likely normally be captured in the repository as "unstructured content"—it can be converted to structured content by importing the appropriate metadata into the data modeling tool and exporting as schema metadata. This example, of course, assumes that the metadata articulated in the spreadsheet is associated with either logical or physical data models.

Selecting the Right Tools

Selecting the right tools for collecting metadata is critical. The tools should be intuitive, and easy-to-use. More importantly, the tools need to integrate well with one another to minimize the number of tool bridging required. (Note: each time data is exported from one tool and imported to another there is a risk that metadata will be lost or altered.) The tools should also be accepted as an industry standard; this ensures that whatever metadata repository solution selected, the tools will be supported.

This Author's Tool Choice

It's important to have tools that are comfortable, reliable, and easy-to-demonstrate. I really like the Embarcadero solution set. The integration provided by the new Embarcadero All Access solution further solidified my choice. I've supplemented the toolset with NitroSecurity's NitroView Application Protocol Monitor (APM) to allow the visibility necessary to profile legacy applications to extract business logic articulated as .NET and Java Code.

Metadata Collection Tools

Embarcadero® ER/Studio® Data Architect for reverse engineering physical schemas, building the associated conceptual and logical models, and for managing schema related metadata. It is well suited for capturing and managing enterprise transformation, consumption, and data lineage metadata. Each reverse-engineered schema can be added to a master data model as a sub-model to the master and the master can be auto-generate a master data catalog. ER/Studio Data Architect supports creation of meta-data tags which are useful for meeting regulatory requirements, such as identify PHI and PII for HIPAA or FISMA respectively. It also supports creation and management of aliases, and provides a good foundation for data governance.

Embarcadero® ER/Studio® Business Architect for importing existing business process models and for building new ones.

Embarcadero® ER/Studio® Enterprise. The latest version of ER Studio allows Data Architect and Business architect to integrate into the ER Studio Repository. This allows business process metadata and data metadata to coexist in a web-accessible repository.

Database Polling Tools

Embarcadero® DB Optimizer™. This provides the ability to capture sessions and filter them by database and table. It is well suited for capturing key performance parameters that are critical for data warehousing construction. It also provides a means of identifying which ETL scripts are being utilized and for validating data lineage.

Application Profiling Tools

Embarcadero® JBuilder™ for evaluating legacy J2EE applications and mapping to business cases

NitroView APM for mapping application roles to business functions and for exposing business rules in near-real time. The output from APM can be correlated with the DBOptimizer output to provide a very accurate picture of which users are accessing which data, even when legacy applications use connection pooling or shared application user accounts.

Most of the metadata is stored in the ER/Studio Repository and made available via the ER Studio Portal included in the ER/Studio Enterprise. The integration of Data Architect and Business Architect into a single repository, as well as the functionality provided in the All Access Toolkit makes All Access a good fit for metadata collection and management.

Summary

The intent of this whitepaper was to provide basic definitions for the concepts of metadata and metadata repositories, as well as to provide a basic methodology for collecting metadata and populating an enterprise metadata repository. Three key points hopefully gleaned:

In order to reap the full benefit from enterprise data assets, an organization must properly collect and manage enterprise metadata.

Collecting the metadata across an enterprise can be expensive, although worthwhile endeavor; the cost can be significantly reduced through proper use of the right tools; and the tangible

benefits associated with the construction of an enterprise view should yield a high Return on Investment.

Metadata Repositories are crucial tools to facilitate metadata management; metadata repository selection is important and the repository should provide key functionality such as management of unstructured content, import/export for a well-known set of tools, and the ability to import and export standard formats such as XML.

About the Author

Ron Lewis is an analyst who specializes in application security for CDO Technologies, a systems integrator that delivers technology-based solutions to government agencies and customers in the private sector. He has worked in the government and commercial security arena for more than 15 years identifying and providing guidance for remediating application vulnerabilities. Ron is considered an industry authority, having authored numerous articles on hardening applications and the hacker mindset. He is also actively involved in industry organizations and efforts such as the Open Web Application Security Project (OWASP) and the Oracle Development Tools User Group (ODTUG).

Embarcadero Technologies, Inc. is a leading provider of award-winning tools for application developers and database professionals so they can design systems right, build them faster and run them better, regardless of their platform or programming language. Ninety of the Fortune 100 and an active community of more than three million users worldwide rely on Embarcadero products to increase productivity, reduce costs, simplify change management and compliance and accelerate innovation. The company's flagship tools include: Embarcadero® Change Manager™, Embarcadero® RAD Studio, DBArtisan®, Delphi®, ER/Studio®, JBuilder® and Rapid SQL®. Founded in 1993, Embarcadero is headquartered in San Francisco, with offices located around the world. Embarcadero is online at www.embarcadero.com.