



Active Metadata Integration

A Cerebra Whitepaper

Executive Summary

Data management is broken. Historic pitfalls and the emerging requirements of an SOA environment have created an unprecedented demand for a robust 'active' metadata layer, while highlighting the inadequacies of current metadata tools.

The Cerebra approach is not a typical metadata solution. Cerebra's active metadata approach is designed to directly benefit business and technical users, to ensure that customers' data and metadata never get locked into proprietary systems, and to never compromise the business's ability to change.



Table of Contents

ACTIVE METADATA INTEGRATION	1
SUMMARY	3
INTRODUCTION	3
THE SOURCE OF THE PROBLEM	4
WHAT'S WRONG WITH THE CURRENT "SOLUTION"	4
ENTER THE SEMANTIC WEB (OR "A NEW WAY OF LOOKING AT METADATA")	6
EXAMPLES OF TODAY'S "ACTIVE METADATA" IMPLEMENTATIONS:	6
CEREBRA'S METADATA REPOSITORY SOLUTION	7
BENEFITS.....	10
CONCLUSION	11



Summary

Data management is broken. Typical large organizations have hundreds of separately developed and implemented applications, with each system using its own concepts, local data definitions and business constraints. The usual attempts to solve this heterogeneity problem only exacerbate the problem, by creating more silos of localized data. These historic pitfalls, and the emerging requirements of an SOA environment, which actually delivers the promise of reuse, have created an unprecedented demand for a robust metadata layer. They have also highlighted the inadequacy of current metadata tools.

The cost to maintain consistent and coherent data interpretation using point-to-point interfaces has given rise to the pursuit of semantic technologies in tackling the problem. Forward-looking commercial and government organizations, especially those interested in building a loosely-coupled service-oriented architecture, have already begun to employ semantic technology standards to build an active metadata solution that can improve the efficiency and agility of their information systems.

A recent report from Gartner highlights the drivers behind this trend:

“As the boundaries between departments and application domains break down, there is a growing need to create an enterprise-wide information strategy to ensure semantic consistency and persistence for all users, applications and services. This is very different from the current practice of adding a new application stack and merging its reference data definition into the current enterprise model. This practice really only integrates data and does not rationalize semantics.”

Gartner Group, 17 January 2005

“Enterprise Information Management Is a Core Element of Your IT Architecture”

Today, companies are using Cerebra products to build an active metadata layer to overcome these issues highlighted by Gartner, using semantic technology to build adaptive ERP, content management and impact analysis applications, primarily in the manufacturing, life sciences, healthcare and defense-intelligence industries.

Introduction

Organizations have invested a great deal of money and many man-years profiling, cleansing, extracting, transforming, aggregating, and loading source data into data warehouses so they can slice and dice it any way they want, get accurate reports, and have a common view of the business. They have added user friendly query tools and near real-time updating. The data is timely, accurate, and accessible. And yet, according to a survey that Cerebra ran at the 2005 DAMA/Metadata conference, fewer than 35% of respondents said that they have accurate automated financial reports. And “when integrating data from multiple departments for reporting purposes,” 43% said they “must resort to the use of specific code and/or manual processes for cross-functional queries” vs. just querying the warehouse.

Even in well-implemented highly centralized data warehouse environments, “inconsistent meanings have created barriers to reliable analytics.” The data architects in corporate IT departments report that the inability of their systems to reconcile inconsistent business



terminology is undermining attempts to implement data management and data integration.

In their own words...

- 👉 "It takes 8 weeks to reclassify products into different market categories"
- 👉 "It takes 3-6 weeks to get a customer list for a mailing, which limits the number of offers we can make"
- 👉 "Mis-categorization of customers by sales throws off revenue accounting by \$1M every reporting period"
- 👉 "Compliance officers are only concerned about the accuracy of the reports. They don't care how many resources were wasted getting them"
- 👉 "The compliance reports only show what events occurred, not whether or not we're in compliance. That's a manual process"
- 👉 "You can't measure KPIs across departments to get a corporate rollup"
- 👉 "I'm pretty sure the rules conflict from one application to another"
- 👉 "The customer data is consolidated, but it's not integrated"
- 👉 "We're on our 3rd attempt at data management and I predict a 3rd failure"

They also suspect that business users may not realize the severity of the problem or just attribute reporting errors to poor data quality.

The Source of the Problem

In fact, the problem is not so much with the data as it is with the *metadata*. If any metadata exists at all, it can't be found. And even if it could, it could not be used to effect the semantic reconciliation required to maintain consistent interpretation of business data by all stakeholders (users, applications processes, services). Organizations that have invested in metadata management in recent years complain of having generated metadata silos to go with their data and application silos. This is an unacceptable return on the investment made in software tools, but more importantly on the investment made by enterprises to share their knowledge with the software tools through configuration and capture.

The tools themselves (ETL, data modeling, BI, DBMS) follow broad industry standards, but their proprietary extensions create semantic problems among the tools. After years of acquisitions, reorganizations and autonomous application development, there are huge discrepancies among departments in the definition of business terms and the application of business policies.

Clearly the need exists in most organizations for an enterprise-wide metadata solution to facilitate the ongoing data integration projects. Exacerbating that need is the focus on SOA. According to Gartner's Yefim Natis, a metadata repository is a key enabling technology for SOA, and no long-term enterprise SOA initiative can succeed without an integrated and searchable repository or registry. (*"The Key to Success with SOA," ebizQ, 7/31/2005*).

What's Wrong With the Current "Solution"

This helps explain the "renewed interest in metadata management" that Gartner is seeing. (*"Magic Quadrant for Metadata Repositories, 2H05 to 1H06," 30 June 2005, ID: G00129274*).



While it may be clear that the next step needed in data and application integration is to solve the metadata conundrum, it is less clear how to do it. Around the turn of the century, what was considered best practice in metadata management was to use traditional metadata tools to build a physically integrated corporate repository. Few successes have been reported, leaving data warehouse users with a lot of good quality data that they still have to reconcile manually (or with single function explicit code or ETL scripts) in order to integrate it in a meaningful way. This brittle and manual effort inhibits further leveraging the warehouse to support business users. Forcing agreement on a common set of terms, even if it were feasible to do once, won't last past the first reorganization, product reclassification, or revision in regulatory requirements.

This approach and the 1990's era metadata tools can not begin to address the needs of today's explosion in data availability driven by an environment of outsourcing, regulatory compliance and widespread adoption of SOA composite applications.

The key problem is that traditional metadata offerings are not able to:

- ☛ manage business vocabulary along with technical syntax
- ☛ actively resolve differences in vocabulary across different departments, applications and services
- ☛ support consistent assignment of business policies and constraints across applications and users
- ☛ accurately reflect all logical consequences of change and dynamically reflect change in affected areas of the business
- ☛ assure and manage component reuse in an SOA environment

Side-by-side, the differences between traditional syntactic metadata repositories and the next generation of standards-based active semantic metadata integration become apparent:

	Traditional Metadata Repository	Active Metadata Integration
Focus	Syntactic descriptions based on DDL extraction and data structure	Adds: Semantic definitions based on business usage; includes terminology and rules
Structure	Multiple repositories, each specific to its application and/or data source. Requires consolidation into single repository for sharing.	Sharable model linking and reconciling multiple existing repositories. Provides reconciliation across multiple repositories for sharing via a hierarchical graph.
Approach to inconsistent vocabulary	Depends on agreement on single standard data definition	Standardize where possible, then use sharable model to reconcile remaining different data definitions

	Traditional Metadata Repository	Active Metadata Integration
Data Reclassification via:	Query-specific ETL scripts	Done dynamically based on use of expressive hierarchical concept-oriented structures
Approach to business rules	Handled via separate tool. Must explicitly code all rules. Managed by IT	Rules included in model; Eliminates rules conflict between applications. Uses inference to infer implicit consequences from explicit information Managed by business users
Support for multiple users and ad hoc query	Supports pre-defined queries and user perspectives. Additional use requires explicit coding.	Extends support to ad hoc queries and new user perspectives w/out requiring extensive manual processes and explicit coding

Enter the Semantic Web (or “A New Way of Looking at Metadata”)

Recent articles cast semantics as a disruptive technology spawned by the vision of the Semantic Web. What does this have to do with metadata? While the full vision of the Semantic Web may be several years away, its promise has fueled the development and commercialization of semantic technologies and the adoption of new semantic data languages by the W3C.

Semantic technologies are being deployed today as the basis of an active metadata layer which aligns technical metadata from multiple applications and data sources and more importantly, manages and aligns *business vocabulary and business policies* within the same model-based environment. And it does not require physical consolidation into a common repository or pre-agreed common definitions.

Examples of Today’s “Active Metadata” Implementations:

Manufacturing companies are using semantics-based metadata today to create “adaptive ERP” systems, as they call them, to provide:

- 👉 automatic reclassification of products by market
- 👉 consistent and accurate application of pricing and other business rules to speed launch of pricing and discount programs

Health care providers are using semantics-based metadata today to:

- 👉 ensure accurate and consistent diagnosis and treatment
- 👉 enable accurate and automated web-based provision of customized document-based information to improve care while reducing cost

Defense and Intelligence agencies worldwide are using semantics-based metadata to:

- 👉 consolidate financial reporting across diverse organizations



- deliver a single view of coalition battle-space data for tactical operations
- discover linkages in data to help fight terrorism

Business integration companies, such as IBM, webMethods, and Software AG, are partnering with semantic technology companies to provide dependency analysis in the management and reconfiguration of software assets and also to assist with smarter data integration.

Oracle's middle layer of grid computing, the Information Grid, enables dynamic joining of information resources. Semantic technologies from Cerebra provide the critical semantic reconciliation for the Information Grid, using active metadata.

Federal agencies, Systems Integrators, large commercial organizations, enterprise software companies and analysts all recognize the potential for semantic technologies to enable a new class of metadata. Active business metadata. They've now begun to realize the benefits.

Cerebra's Metadata Repository Solution

As the leading provider of enterprise-ready semantic technologies, Cerebra, with its standards-based repository, integration and inferencing solutions, takes a holistic approach to unify and harmonize metadata, data and services throughout the enterprise.

The Cerebra solution is natively based on two modern data languages that improve upon and extend relational and XML formats, offering a much simpler path to achieving the business goals of reducing stand-alone solution development.

RDF (Resource Description Framework) is a W3C standard for encoding graph-based data. The graph nature of RDF is fundamentally different from relational, XML, or object data structures, but can also represent the semantics of each. RDF's key technical benefit is flexibility, allowing data structures to be extended dynamically without causing costly system-wide impacts to queries, mappings, and schema.

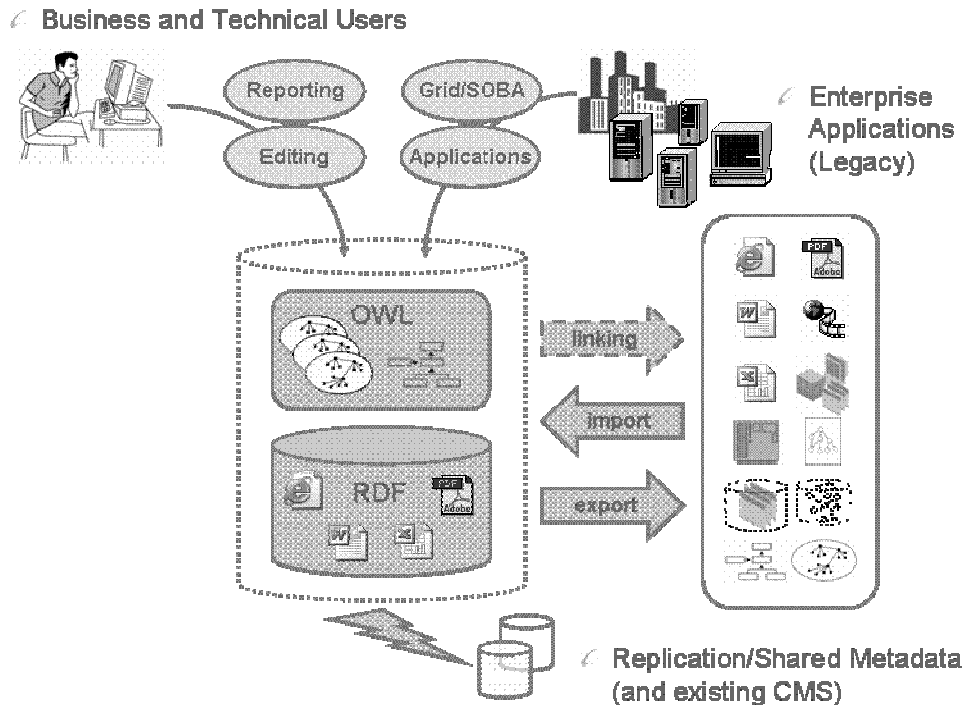
OWL (Web Ontology Language) is a W3C standard for encoding schema-driven constraints and rules on graph data. Data classification can be substantially automated using OWL by taking advantage of inference on the rules and constraints within the schema. Thus contents can be classified dynamically according to business policies and searched directly from the metadata, even if the content is still within a legacy system.

The benefits of a system that natively uses OWL and RDF are federation of metadata and simple integration. Business analysts are able to associate information to the metadata without having to store all of it in a warehouse – directly mapping to relational databases, web services, and other IT assets.

This approach uses an RDF layer as the physical storage system for content that can be centralized. OWL is used to constrain the RDF and link to external sources. The core query mechanisms and metadata markup would be handled by Cerebra's native OWL platform.

This approach enables all the typical use cases you would expect: reporting, editing, mapping, and management. It also provides the infrastructure connectivity to draw from

and link to legacy databases, applications, and services. The repository can be published on the network as a simple web service offering both query and functional APIs for application access.



From an enterprise scalability standpoint, the system is hardened. The RDF subsystem makes use of all the typical database performance features like clustering, partitions, and failover. The Cerebra system has similar capabilities and has proven to scale linearly with respect to query times under load. From the IT perspective, the system is managed as a single unit.

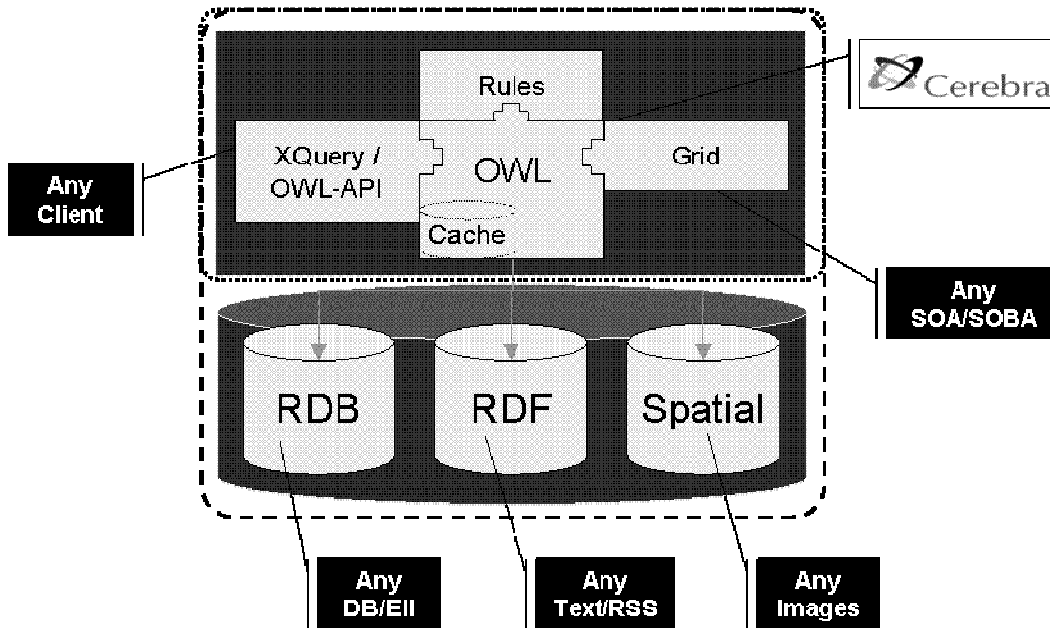
The Cerebra solution goes beyond traditional systems, providing dynamic content classification, highly expressive metadata, queries using an inference engine, massive scalability, and rich content support that spans XML and RDB types. The combination of these strengths cannot be found in any other metadata platform.

In response to customers' needs for a metadata solution to meet increasingly demanding requirements, the Cerebra solution brings best-of-breed capabilities to the following key areas:

- Consistent corporate-wide integration – standards-based OWL/RDF
- Cross-functional information sharing – using dynamic data classification
- Improved management and security – sound/correct metadata constraints
- Unified information governance – due to active metadata OWL

An example of the Cerebra metadata repository is shown below. It is designed to serve as a comprehensive platform to drastically reduce development and deployment of stand-alone and siloed applications. Note that each layer of the repository solution

allows for standard application server and database extensions suited for specific business needs.



From a technical standpoint, this approach is able to leverage active metadata internal to an IT organization. In effect, creating what Cerebra has termed "The Enterprise Semantic Web" and what Gartner is now calling "The Corporate Semantic Web." Restricted to behind-the-firewall applications, this active metadata repository is capable of integrating, disseminating, linking, and managing IT information assets. The basis of the technology lies in the inference engine's (a reasoner) ability to change the categories under which content is organized. From an IT perspective, the process of organizing content is no different than with traditional systems. However, when the time comes to change how the data is organized, and it will come, this active metadata approach can actually help automate widespread change by simplifying the process of re-categorizing data, applying policies (such as security and regulatory), and integrating new data. In short, the application of modern data management techniques has transformed the way IT practitioners should think of metadata – from something that was "nice to have" to something that is a strategic part of working with corporate information.

Key technical differentiation, when compared with traditional metadata management approaches is summarized by the following table:

What Technology	Why it's Important
Standards-based Metadata	To prevent creating a metadata silo and preserve investment
- OWL	Describe and classify content
- RDF	Most flexible core data structure

What Technology	Why it's Important
Inference / XQuery	Reclassify dynamically Secure with correct provenance Perform "what-if" scenarios Native XML query approach
Graph Management DB	Apply classic RDB management abilities to graph data model
Grid / SOA Service	Metadata to describe and query Web services – legacy transition is easy

Benefits

This approach to metadata repository serves traditional needs such as reporting and mapping. Additional capabilities in the areas of vocabulary management, business rule management, and enterprise search are provided by the technical enhancements made possible by Cerebra's unique core data formats. Importantly, all the metadata is persisted in standard data languages, ensuring that the meaning of information (not just the syntax) can outlive any particular software solution. Cerebra's metadata capabilities include:

- 🍌 **Information Asset Management**
 - 🍌 Track assets with global or distributed model
 - 🍌 Embed assets as metadata models
 - 🍌 Link assets as external resources
 - 🍌 *Asset = documents, structures, rules, programs, elements, stewardship etc.*
- 🍌 **Vocabulary Management**
 - 🍌 Federated taxonomy and graph management
 - 🍌 Global impact analysis for change management
 - 🍌 Automatic and extensible re-classification of content
- 🍌 **Lifecycle and Versioning**
 - 🍌 Secure assets with global registry and control list
 - 🍌 Steward assets with quality-centric ownership properties
- 🍌 **Enterprise Search & Retrieval**
 - 🍌 Query using metadata (across RDB, Web Services, KB)
 - 🍌 Search Grid/SOA services using metadata as a virtual repository
- 🍌 **Source/Target Mapping and Scanning**
 - 🍌 Annotate SOA Web Services and DBs with expressive metadata
 - 🍌 Import/Export metadata models to 3rd party tools and formats
- 🍌 **Reporting and Composite Applications**
 - 🍌 Manage and maintain reporting/stewardship hierarchies
 - 🍌 Present intuitive interface for business users and technical users
 - 🍌 Expose application and query API for new applications to source data

Cerebra's metadata solution provides unique differentiation and capabilities that go far beyond traditional systems. Most importantly, the metadata has a life outside of a particular vendor's solution. The standard OWL/RDF format for the metadata also provides expanded capabilities that are limited only by the imagination of business needs.



- 👉 The metadata is encoded in an actionable model, not DB tables
- 👉 Users may query directly through the metadata, not via DB procedures
- 👉 The metadata applies validate-able policies, not custom functions
- 👉 Reclassification of meta/data is automated, not ETL or transformation
- 👉 Reliability of views is DB quality, not a "probabilistic" inference

A significant benefit of the Cerebra metadata solution is that it enables a true enterprise search capability as well as a coherent governance and security strategy that doesn't require proprietary Cerebra or Oracle software to operate. Additional business-line benefits include:

- 👉 **Smarter Access to Information**
 - 👉 For ERP applications and workers
 - 👉 For Business Intelligence tools and analysts
 - 👉 For SOA Business Applications (SOBA) and workers
 - 👉 For Data Warehouses and architects
 - 👉 For Enterprise Search
- 👉 **Improved Quality**
 - 👉 Common, consistent definitions of information assets
 - 👉 Applied policies and business rules throughout an information lifecycle
 - 👉 Promoting reuse of IT assets across organizational boundaries
- 👉 **Streamlined Application/Integration Development**
 - 👉 Single source of accurate meta-information about distributed content
 - 👉 Method of allowing diverse organizations to reconcile semantics
 - 👉 Identification of redundant processes and information
- 👉 **Reduced Maintenance and Support**
 - 👉 Quickly identify impact of change
 - 👉 Automate the re-classification of large quantities of content
 - 👉 Model-driven behavior reduces typical coding time for development
- 👉 **Information Stewardship**
 - 👉 Track content ownership and change provenance policies
 - 👉 Understand root source of composite content, preserving security allowances

Conclusion

In summary, the Cerebra approach is not your typical metadata solution. It is designed so that it will directly benefit both business and technical users. Cerebra is thoroughly committed to an approach that ensures that customers' data and metadata never get locked into proprietary systems, and will have a lifetime well beyond any particular vendor's management software. We expect our customers to hold us to that promise.