

Semantic Integration in Enterprise Information Management

By Muralidhar Prabhakaran & Carey Chou

Creating structurally integrated and semantically rich information should be the focus of next generation EIM systems

INTRODUCTION

With the rapid growth in the modern day enterprise comes explosion of information, diversity in formats and the sources where they are located.

Present day Enterprise Information Management (EIM) Systems manage this disparity through what is known as Syntactic Integration. This is done through techniques like Enterprise Application Integration (EAI) and Enterprise Information Integration (EII). While this is a fundamental building block towards the "Single Version of truth" paradigm, it still fails to achieve the so called "Semantic Integration."

Semantic Integration can be realized when every single Information Asset has "one meaning" or "one context" to it. To achieve this state, the Enterprise Information Management Systems and the underlying architectures need to be enhanced.

This paper attempts at outlining the enhancements and discusses the technology options available.

CURRENT STATE OF ENTERPRISE INFORMATION ARCHITECTURE

As organizations grow, new systems and applications are added and processes are refined, modified or improved.

Enterprise Information Architecture (EIA) is constantly challenged by organizational expansion and changes due to:

- Information Complexity
- Need for Compliance
- Competitive Agility.

INFORMATION COMPLEXITY

Information complexity is created due to a number of factors. Addition of new applications and systems means that there are increasing number of interaction points between applications and consequent integration and exchange of data and information.

The significant factors that impact information management in this situation are the location and the format of data.

Addition of new applications in a flat enterprise means that data can be geographically disparate. Also, the enterprise is flooded with all possible formats of data, including Cobol copy books, relational databases, Extensible Markup language (XML) documents, Portable Document Format (PDF) and Word files.

NEED FOR COMPLIANCE

The need for compliance is one of the hot topics of discussions in the modern day enterprise. Compliance requirements like Sarbanes Oxley Act (SOX) and Basel-2 are compelling executives to ensure that the right information is available at the right time at the right level of granularity.

For executives to have the right confidence in the information being delivered, EIM initiatives in the enterprise need to ensure that the right people, processes and technology are available to deliver the single version of truth.

With the addition of newer systems and applications, EIMs are under pressure to ensure that the quality of data being delivered is consistently maintained.

COMPETITIVE AGILITY

Competitive agility is the capability of the information system to enable its organization to stay ahead of its competitors.

For continual success, businesses need real time information about activities within the organization and how and why of issues for better decision making. This puts an enormous strain on information systems to be responsive and accurate. To ensure accuracy and responsiveness, EIM needs to have robust information architecture in place.

CRITICAL SUCCESS FACTORS FOR EIM

According to Gartner[1], "Enterprise Information

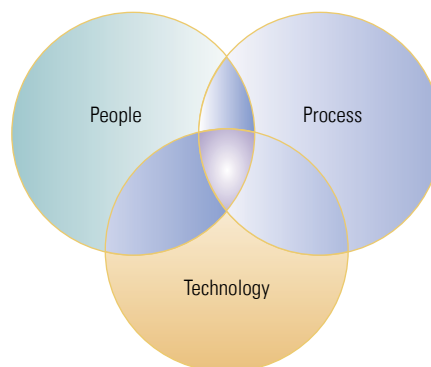


Figure 1: Next generation EIM strikes a balance between People, Processes and Technology

Source: Infosys Research

Management is an organizational commitment to structure, secure and improve the accuracy and integrity of information assets, to solve semantic inconsistencies across all boundaries, and support the technical, operational and business objectives within the organization's enterprise architecture strategy."

For a successful EIM initiative, organizations need to have alignment across three dimensions: People, Processes and Technology.

PEOPLE

The dimension of people is focused around the organizational structure and the governance policies to ensure that the right strategy and the appropriate executive support are secured for Enterprise Information Management initiatives. Given the extensive literature on this topic, this paper does not attempt to detail out issues related to organizational structure.

PROCESSES

Organizations need to take appropriate initiatives to ensure the accuracy and integrity of information assets. Towards this, each

organization needs to undertake the following key initiatives:

- Master Data Management
- Metadata
- Enterprise Wide Data Warehouse
- Service Oriented Architecture

Master Data Management (MDM):

Each organization has a set of key information assets that need to have consistent representation and process. Typical examples include customer, product or location data.

Today most of these assets are spread across various sources and geographies and most importantly there are several processes that either create or modify them.

As a part of the EIM umbrella, each organization needs to undertake a MDM initiative to standardize and reconcile all of their master data.

Metadata Strategy:

Also known as the data about data, Metadata is a key component for a successfully integrated enterprise. Despite this, organizations do not have dedicated programs to define their metadata strategies.

Usually departments try to define their own metadata approaches and quite often fall for a vendor defined solution that talk about metadata interchange among tools.

Defining a sound metadata strategy for business, process and technical metadata under the EIM umbrella is the first baby step towards a syntactically and semantically integrated enterprise.

Enterprise Wide Data Warehouse:

By developing the data warehouse, organizations focus on defining Key Performance Indicators (KPIs) and metrics that will enable them to

accurately measure performance and also serve as the backbone for compliance requirements. One of the other significant benefits of this is the inherent ability to improve the quality of data within the enterprise.

Successful data warehousing projects usually undertake elaborate steps to ensure the best quality data gets into the warehouses.

With a properly planned closed loop feedback, one can check the quality of data that is fed back into the operational systems. As a result, the quality of data improves across the organization.

Service Oriented Architecture:

Organizations need to seriously think about a Service Oriented Architecture (SOA) for a complete and successful EIM.

Although SOA is the buzzword of the day and there is hype surrounding it, a well-planned, thorough implementation is certain to provide the scalability, modularity and sustainability that Enterprise Information Architecture needs.

WHAT IS MISSING IN THE PRESENT ENTERPRISE INFORMATION ARCHITECTURE?

From the preceding arguments, it is evident that one of the core requirements to ensure a successful Enterprise Information Architecture is a sound integration strategy.

EIMs of today integrate data and information only at a syntactic level.

What this means is that most organizations define an enterprise wide information model that attempts to normalize the definition of information components. This is then mapped to various data and application sources. Further, the data from these sources are integrated structurally to adhere to the

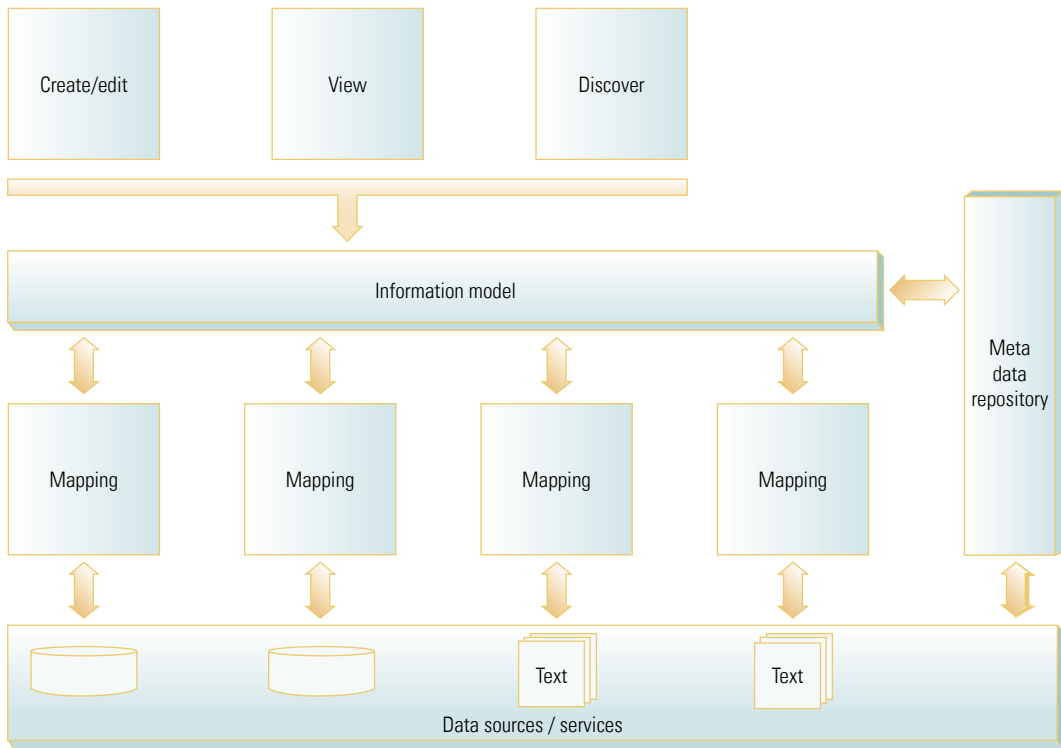


Figure 2: Typical architecture of an ontology based information model.

Source: Infosys Research

information model. This gives a consistent format of information within the enterprise.

However, what is amiss is the true meaning of the information components. Without the true meaning, it presents a major challenge to correlate the meaning of information models across domains.

Semantic integration is achieved when a piece of information asset has an unambiguous definition and can be expressed in a formal and explicit way, linking human understanding with machine understanding [2].

To achieve true semantic integration, Information Architectures of today need to be enhanced to accommodate a semantic integration layer.

HOW TO ACHIEVE SEMANTIC INTEGRATION?

An Ontology based approach seems the most suitable one for achieving semantic integration.

Ontology can be defined as a data model that represents a domain and is used to reason about the objects in the domain and the relationship between them.

How do we leverage Ontology to generate a semantic information layer? The answer lies in the fact that we can use a Ontology based language (e.g. OWL) to define an information model of a domain.

With the definition of the information model complete, we can then map the model to the data sources or services thereby creating a link between the 'human' and the 'computer' worlds.

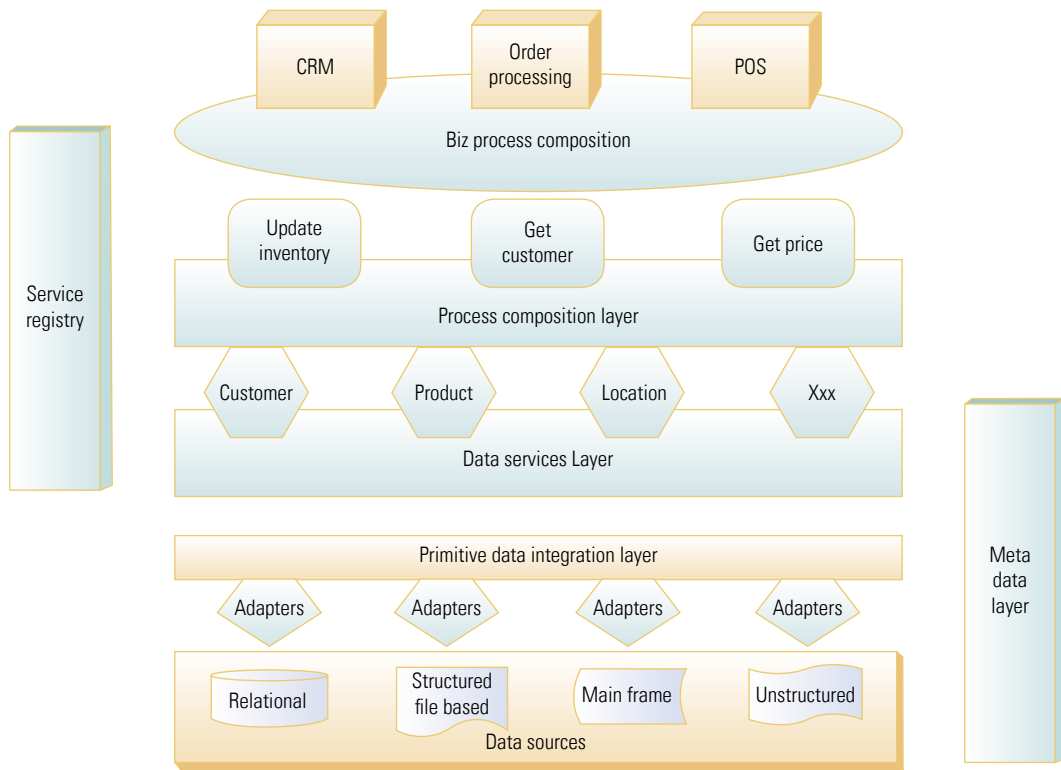


Figure 3: Typical EIM Layout in an Enterprise

Source: Infosys Research

However, the challenge in this approach is the availability of a true information model for a specific domain. To address this, an automatic meta data discovery service combined with a manual mapping task to create, update and maintain the ontology is advocated.

Figure 2 indicates a logical architecture to maintain the lifecycle of an ontology based information model. The lowest layer contains all the sources of information, being data sources or a data service that presents information in a normalized manner.

The information model layer is where the Information model is captured. An organization can create the information model with a domain specific ontology although this can be a uphill challenge.

Alternatively there is a discovery service that will use existing metadata repositories to automate the extraction of the bare bone semantic model using techniques like vocabulary mapping, metadata gleaning [3] and inference rules, after which manual intervention from domain experts within the organization can create a semantically rich information model. The model can be kept current through the edit service. And finally there is view/query service to view the structure and meaning of the model.

THE TYPICAL ENTERPRISE INFORMATION ARCHITECTURE

A typical Enterprise Information Management layout in today's enterprise has several layers. For

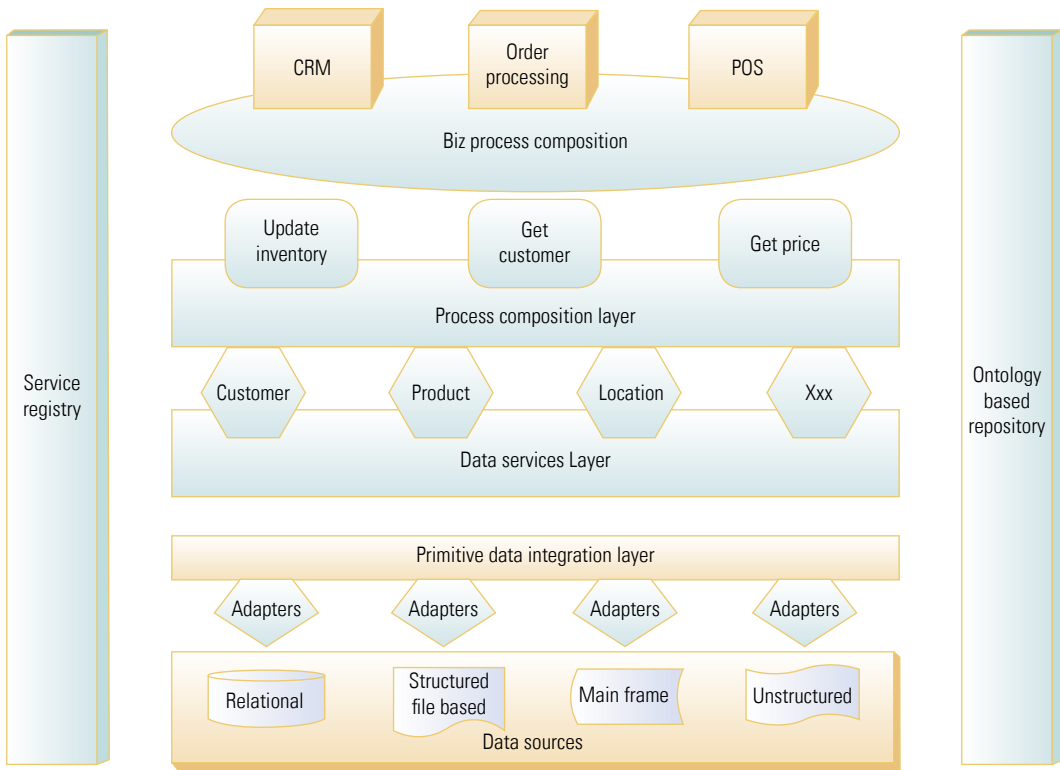


Figure 4: Enhanced Enterprise Information Architecture **Source:** Infosys Research

the sake of simplicity and ease of representation, details around process orchestration etc are omitted from Figure 3.

The basic layout is as follows:

LAYER 1:

This is the data sources layer. This tier represents the various sources of information like databases, XML files, mainframe and unstructured information.

LAYER 2:

This is the data integration layer where techniques like EII, ETL and EAI provide basic integration techniques for data. This layer primarily comprises of adapters to various sources.

LAYER 3:

This is the data services layer which essentially leverages the integration layer and the metadata repository to create data oriented services. Some samples of master data elements are shown as potential services in the figure.

LAYER 4:

The process composition layer at which basic processes start taking shape. For example a Get Customer or Get Price could be one of these services.

LAYER 5:

Final layer is the business process composition layer at which typical business processes are

exposed. There is a service repository that is available at the data layer and above which can be leveraged to define services and utilized at run time. Realistically however, in the current environment, the service repository is leveraged at the process composition layer where basic processes are exposed as web services.

One of the main drawbacks of this framework is that the integration at the data layer happens only at a syntactic level. To enhance this layer to be a semantic layer, the following changes are proposed to the Enterprise Information Architecture.

INTRODUCING SEMANTIC INTEGRATION WITHIN THE EXISTING ARCHITECTURE

The significant difference in the enhanced information architecture is the introduction of the semantic integration layer. This is the layer that utilizes the ontology based information model and translates or maps the model to the underlying data and information sources.

The semantic integration layer uses reasoning techniques to resolve potential semantic conflicts when acquiring metadata from source systems. The output of this layer is semantically rich and coherent information that can be consumed by the layers above to create rich services.

Some of the promising technologies in this space include Cerebra and Unicorn. These technologies lean heavily on Web Ontology language (OWL), Resource Description Framework (RDF) and other standards to provide scalable, maintainable, reusable and semantically rich enterprise.

Cerebra has a suite of products to create the information model and also provide a runtime environment to create, modify and utilize the semantic information.

Unicorn too provides a workbench to

model the information and a runtime engine to leverage it. Unicorn has been now acquired by IBM and the plans are to integrate it into the IBM Websphere metadata server.

Profium offers a suite of products to automate the process to extract semantics from various source formats so to reduce the manual effort of creating the bare bone semantics from disparate source information components.

CONCLUSION


By creating an actual semantically aware information architecture and combining it with enabling service oriented architecture, the enterprise gets ready for realizing the 'information as a service' paradigm.

With the enhanced EIM set up, each service that is defined at higher layer can consume services at the lower layer which is semantically aware.

With this, each service becomes rich with information and thereby provides the long due 'human understanding' element in the services provided by Information Technology.

REFERENCES

1. Enterprise Information management- Getting value from Information Assets, Gartner Business Intelligence Summit, 2006.
2. Bruijn Jos De, Semantic Information Integration Inside And Across Organizational Boundaries, DERI Technical Report, 2004
3. Gleaning Resource Descriptions from Dialects of Languages (GRDDL), <<http://www.w3.org/2004/01/rdxh/spec>>, [Accessed July 1, 2006]
4. <www.en.wikipedia.org>, [Accessed July 1, 2006]

5. <www.cerebra.com.>,[Accessed July 1, 2006]
 6. Unicorn,<www.unicorn.com>,[Accessed July1, 2006]
 7. <http://www.profiu.com/technology/technology.html>,[Accessed July1,2006]
 8. Wikipedia,<www.wikipedia.org.>,[Accessed July1, 2006] 
-

Authors in this issue

CAREY CHOU

Carey Chou is a Senior Technical Architect in the Retail, CPG and Distribution business unit, Infosys. He has several years of experience in agile enterprise integration solutions and high performance event driven architecture. He can be contacted at carey_chou@infosys.com

MURALIDHAR PRABHAKARAN

Muralidhar Prabhakaran is a Principal Architect and Head of the Data Management Group at the Retail, CPG and Distribution business unit, Infosys. His current interests include Unstructured Data Processing, Data Mining and generation next information integration. He can be contacted at muralidhar_p@infosys.com

For information on obtaining additional copies, reprinting or translating articles, and all other correspondence, please contact:

Telephone : 91-80-41173878

Email: SetlabsBriefings@infosys.com

© SETLabs 2006, Infosys Technologies Limited.

Infosys acknowledges the proprietary rights of the trademarks and product names of the other companies mentioned in this issue of SETLabs Briefings. The information provided in this document is intended for the sole use of the recipient and for educational purposes only. Infosys makes no express or implied warranties relating to the information contained in this document or to any derived results obtained by the recipient from the use of the information in the document. Infosys further does not guarantee the sequence, timeliness, accuracy or completeness of the information and will not be liable in any way to the recipient for any delays, inaccuracies, errors in, or omissions of, any of the information or in the transmission thereof, or for any damages arising there from. Opinions and forecasts constitute our judgment at the time of release and are subject to change without notice. This document does not contain information provided to us in confidence by our clients.

Infosys[®]

POWERED BY INTELLECT
DRIVEN BY VALUES