

Knowledge Management in Heterogeneous Data Warehouse Environments

Larry Kerschberg

Co-Director, E-Center for E-Business,
Department of Information and Software Engineering, George Mason University,
MSN 4A4, 4400 University Drive, Fairfax, VA 22030-4444, USA
kersch@gmu.edu; <http://eceb.gmu.edu/>

Abstract: This paper addresses issues related to Knowledge Management in the context of heterogeneous data warehouse environments. The traditional notion of data warehouse is evolving into a federated warehouse augmented by a knowledge repository, together with a set of processes and services to support enterprise knowledge creation, refinement, indexing, dissemination and evolution.

1 Introduction

Today's organizations are creating data and information at an unprecedented pace. Much of that data comes from organizational business transactions. Traditionally, that data has resided on corporate servers and has represented operational on-line transaction processing (OLTP) data. The goal of a *data warehouse* is to integrate applications at the *data level*. The data is extracted from operational systems, cleansed, transformed, and placed into the data warehouse or data mart according to a schema, such as the star or snowflake schema [1]. Although the notion of creating an *integrated* data warehouse is appealing conceptually, it may be infeasible operationally. Trends indicate that *federated* data warehouse architectures are more practical, from the political, operational, and technical points-of-view [2, 3].

Moreover, as organizations move their operations to the Internet and establish partnerships, via portals and extranets, with both their customers and suppliers, the "data" for the *e-enterprise* becomes distributed among many parties. This presents both challenges and opportunities in building enhanced data warehouses that reflect the information holdings of the e-enterprise.

Our notion of the data warehouse must be extended to include not only operational transaction-oriented data, but also data created by *knowledge workers* within the enterprise. We can then include technical reports, correspondences, presentations, audio, video, maps and other *heterogeneous* data types, as well as unstructured data.

Increasingly, we view the Internet and the World Wide Web [4] as data sources that complement e-enterprise information holdings. The data collected from the Web must be incorporated into the data warehouse for business decision-making.

Data warehouse holdings can be used for *business intelligence*, based upon knowledge created by means of data mining and knowledge discovery, which are

major themes of this conference. However, in order to acquire, catalog, organize, index, store and distribute the enterprise's information holdings, we need to address issues related to the management of knowledge resources, termed *knowledge management*. In this paper, we present an architecture for the management of enterprise knowledge assets in the context of data warehouses.

The paper is organized as follows. Section 2 presents an analysis of the evolution data warehouse architectures in response to evolving enterprise requirements. Section 3 presents an architecture for knowledge management, and discusses the knowledge-oriented services provided to support heterogeneous data warehouses. Section 4 presents our conclusions.

2 Data Warehouse Architectures

Data warehouse architectures have evolved in response to our evolving data and information requirements. Initially, the data warehouse was used to extract transactional data from operational systems to perform on-line analytical processing (OLAP) [5]. One of the problems with that centralized approach is that data in the warehouse is not synchronized with data residing in the underlying data sources. This has led to research on the view materialization problem [6, 7].

Although the goal of data warehousing had been to create a *centralized and unified view* of enterprise data holdings, this goal has not been fully realized. Many factors contributed to this, such a problem of semantic heterogeneity, terminology conflicts, etc. However, one of the overriding factors has been the need for organizations to assert ownership over the data. Organizations wish to *own* their data, wish to assume responsibility for the *curation*, or *stewardship* of their data, and wish to *share* their data according to well-defined sharing agreements. Thus, rather than spend large amounts of money on a centralized data warehouse, enterprises are implementing smaller data marts, and integrating them through federated architectures.

2.1 Federated Data Warehouses

The trend away from the centralized data warehouse leads to the notion of a *federated* data warehouse, whereby independent smaller warehouses within the corporation publish selected data in *data marts*, which are then integrated. One interesting case study is that of Prudential Insurance Company [8], which decided to build a federated data warehouse by combining data from a central meta-data repository with existing stand-alone Line-of-Business (LOB) warehouses. This approach turned out to be easier and less time- and resource-consuming. Before the federated warehouse, marketing personnel could not determine how many customers had both Prudential life and property insurance, in order to sell them securities. The data resided in distinct LOB data warehouses, implemented using differing commercial products and query languages. However, such queries are common in business decision-making, and the inability to answer such questions leads to lost revenues.

The Prudential approach involves the creation of an information hub containing data of interest to the entire corporation, while data of interest to LOBs remains in the local data warehouses. The hub can then be used to create specialized data marts for decision-makers.

The federated approach is appealing in that it common data can reside in a metadata repository, while rapidly changing data can reside in data marts or in the underlying data sources. We have had extensive experience with the federated approach [2, 9-11] and these concepts were proposed for the NASA EOSDIS Independent Architecture Study and they are now being implemented. We now turn our attention to another phenomenon that is imposing new requirements on data warehouses, that of e-business.

2.2 E-Enterprise Warehouses

The e-enterprise is based on inter-enterprise partnerships among customers and suppliers. These partnerships are predicated on the sharing of data, information and knowledge through interoperable business processes, data sharing protocols, and open standards such as XML [12].

E-enterprise strategic partnerships entail data that currently is distributed among specialized software-vendor-specific applications for Customer Relationship Management (CRM), Content Management (CM) for catalog integration and data meta-tagging, Enterprise Application Integration (EAI), Human Resources (HR), and back-end fulfillment systems such as Enterprise Resource Planning (ERP).

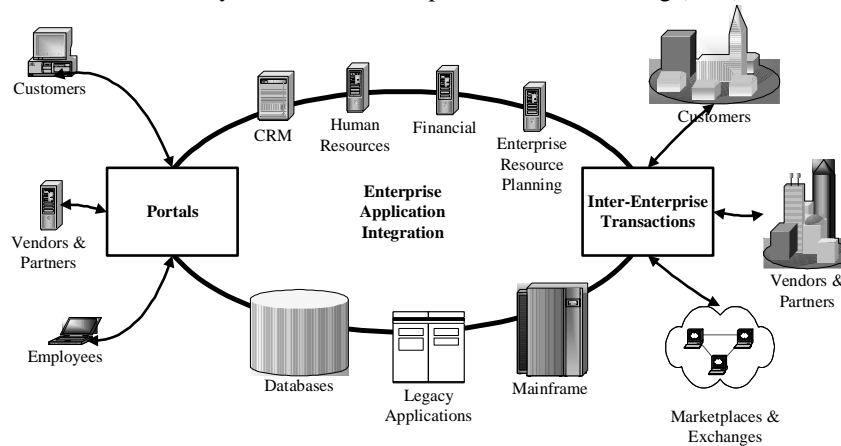


Fig. 1. E-Enterprise Heterogeneous Distributed Data Sources

The dynamic nature of the inter-enterprise relationships and the distributed heterogeneous data contained in proprietary software systems requires new approaches to data warehousing. The federated approach allows each application system to manage its data while sharing portions of its data with the warehouse. There will have to be specific agreements with customers, vendors and partners on how

much of their data will be made available to the warehouse, the protocols and standards to be used to access and share data, the data security attributes and distribution permissions of shared data, and data quality standards [13, 14] to which these partners must adhere.

Another driver is the advent of the Semantic Web [15] with its associated web services [16], which will enable the creation of dynamically configurable e-enterprises. Data warehouse concepts and tools should evolve to include mechanisms to access the databases of these web services so as to obtain e-enterprise-specific data regarding those services, e.g., performance metrics, status information, meta-data, etc., which would then be stored in the federated data warehouse. We envision intelligent agents [17, 18] interacting with web service providers to gather relevant information for the data warehouse.

The data warehouse will thus become a *knowledge repository* consisting of the data from traditional data warehouses, domain knowledge relevant to enterprise decision making, workflow patterns, enterprise metadata, and enterprise ontologies.

Associated with the knowledge repository will be process knowledge for data cleansing, data quality assurance, semantic meta-tagging of data, security tagging, data stewardship (curation) and knowledge evolution. These processes are described in more detail in the next section.

3 Knowledge Management Architecture for Heterogeneous Data Warehouse Environments

The evolution of data warehouses into knowledge repositories requires a knowledge management architecture within which to acquire data from heterogeneous information sources and services, to prepare data for

The knowledge management architecture we propose supports the following:

- Access to both internal and external information sources;
- Repositories that contain explicit knowledge;
- Processes and tool support to acquire, refine, index, store, retrieve, disseminate and present knowledge;
- Mechanisms for cooperative knowledge sharing among knowledge workers;
- Organizational structures and incentives to enable and foster a knowledge sharing and learning organization;
- People who play knowledge roles within the organization, including knowledge facilitators, knowledge curators, and knowledge engineers; as well as
- Information technology support for the entire architecture.

Figure 2 below shows an architecture denoting the Knowledge Presentation, Knowledge Management, and Data Sources Layers. At the top layer, knowledge workers may communicate, collaborate and share knowledge. They are provided information by means of the Knowledge Portal, which can be tailored to the profile of each knowledge worker.

The Knowledge Management Layer depicts the Knowledge Repository and the processes that are used to acquire, refine, store, retrieve, distribute and present knowledge. These processes are used to create knowledge for the repository.

The Data Sources Layer consists of the organization's internal data sources including documents, electronic messages, web site repository, media repository of video, audio and imagery, and the domain repository consisting of the domain model, ontology, etc. Also depicted are the external sources of data, including web services that can be used to augment the internal holdings.

3.1 The Knowledge Management Process Model

The process model associated with knowledge management consists of well-defined activities which: 1) help to ensure the quality of the data and information used by knowledge workers, 2) assist in the refinement of data and information into knowledge, 3) allow the efficient storage and retrieval of metadata and knowledge, 4) promote the timely dissemination and distribution of knowledge, and 5) support the tailored presentation of knowledge. These activities are presented in the following subsections, and we review some of the major sub-activities.

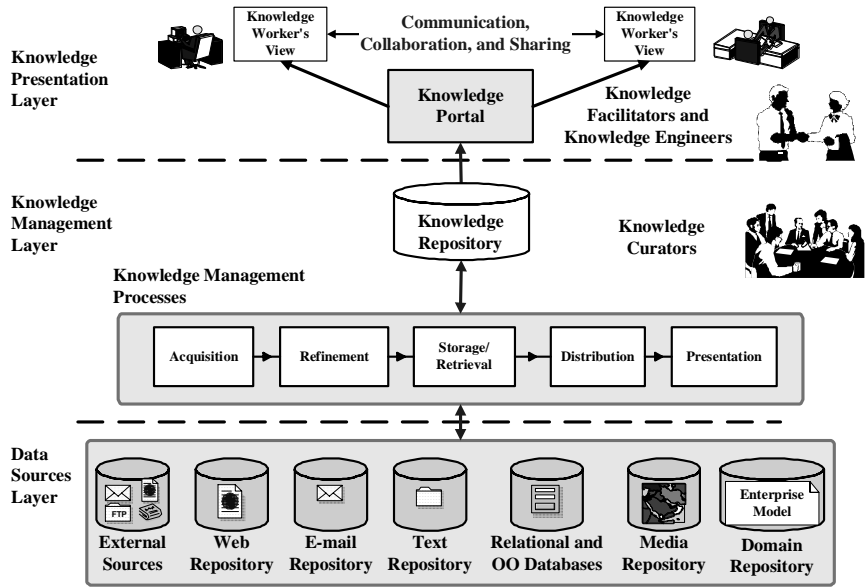


Fig. 2. Three-layer Knowledge Management Architecture

Knowledge Acquisition

During knowledge acquisition the Knowledge Engineering captures knowledge from domain experts through interviews, case histories, and other techniques. This knowledge can be represented as rules and heuristics for expert systems, or as cases for use in a case-based reasoning system.

Knowledge Refinement

Another important source of knowledge may be found in corporate repositories such as document databases, formatted transaction data, electronic messages, etc. Part of the knowledge acquisition process is to cross reference items of interest from information contained in multiple repositories under multiple heterogeneous representations. During knowledge refinement this information is classified and indexed, and metadata is created in terms of domain concepts, relationships and events. In addition, the domain context and domain usage constraints are specified. Data pedigree information is also added to the metadata descriptors, for example, intellectual property rights, data quality, source reliability, etc. Data mining and data analysis techniques can be applied to discover patterns in the data, to detect outliers, and to evolve the metadata associated with object descriptors.

Storage and Retrieval

The refined data, metadata and knowledge are indexed and stored for fast retrieval using multiple criteria, for example, by concept, by keyword, by author, by event type, and by location. In the case where discussion groups are supported, these should be indexed by thread and additional summary knowledge may be added and annotated. Access controls and security policies should be put in place to protect the knowledge base and the intellectual property it contains.

Distribution

Knowledge can be distributed in many ways, as for example, a corporate knowledge portal. Here knowledge workers may find relevant information sources for their tasks. Electronic messaging may also be used to distribute knowledge in the form of attachments of documents, presentations, etc. Another approach is to have active subscription services whereby agents inform users of relevant information in e-mail messages with hyperlinks to knowledge in the repository.

Presentation

The Knowledge Portal may handle knowledge presentation, and the interface may be tailored to the needs and preferences of each individual user. The portal should support user collaboration so as to combine tacit knowledge with explicit knowledge for problem solving.

3.3 Knowledge Management System Architecture

Figure 3 shows a conceptual model of a Knowledge Management System to support the KM processes and services. A selected number of services, which may be internal to the enterprise or outsourced through partnerships, subscriptions or web services, are now discussed.

The Knowledge Presentation and Creation Layer

The services provided at this layer enable knowledge workers to obtain personalized information via portals, to perform specialized search for information, to collaborate

in the creation of new knowledge, and to transform *tacit knowledge* into *explicit knowledge* [19] via discussion groups. Our work on WebSifter [20, 21] indicates that personalized search preferences together with user-specified, ontology-directed search specification and results evaluation can enhance the precision of documents returned by search engines. Thus search services are an important component of knowledge management.

The knowledge creation services allow knowledge workers to create value added knowledge by annotating existing knowledge, providing metatags, and aggregating heterogeneous documents into named collections for future use.

Knowledge Management Layer

This layer provides middleware services associated with knowledge indexing and information integration services (IIS). Data warehouse services are listed among the IIS, together with federation, agent, security and mediation services. Services such as data mining, metadata tagging, ontology & taxonomy, curation and workflow services are used to support the KM Process Model described in Section 3.1.

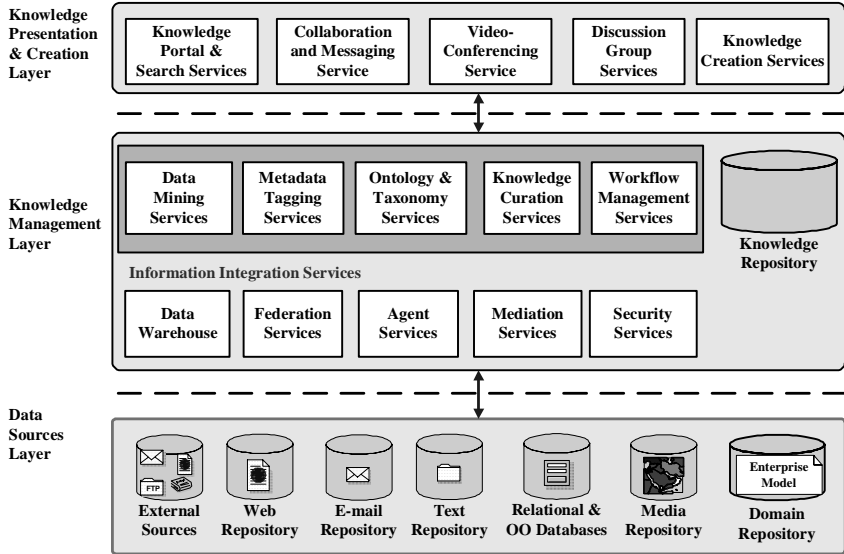


Fig. 3. Conceptual Model for a Knowledge Management System

Data Mining Services. These services include vendor tools for deducing rules from numeric data, as well as concept mining from text. This knowledge can be used to enhance the Knowledge Repository and to provide refined knowledge to decision-makers.

Metatagging Services. Appropriate indexing of knowledge assets is crucial as collections grow. XML [12] and Resource Description Framework (RDF) [22] are emerging as open standards for tagging and metadata descriptions. The Digital

Library community has proposed the Dublin Core Metadata Initiative for tagging books.

Ontology & Taxonomy Services. The organization of enterprise knowledge is an area of ongoing research. Clearly, the construction of domain-specific ontologies is of utmost importance to providing consistent and reliable terminology across the enterprise. Hierarchical taxonomies are an important classification tool. Our research in this area includes the Intelligent Thesaurus [2, 10] and we have used user-specified taxonomies to guide the WebSifter meta-search engine [20, 21]. The intelligent thesaurus is an active data/knowledge dictionary capable of supporting multiple ontologies to allow users to formulate and reformulate requests for information. The intelligent thesaurus is similar to the thesaurus found in a library; it assists analysts in identifying similar, broader or narrower terms related to a particular term, thereby increasing the likelihood of obtaining the desired information from the information sources. In addition, active rules and heuristics may be associated with object types as well as their attributes and functions. Content management in inter-enterprise environments can make use of ontologies, particularly in the area of catalog integration [23, 24].

Agent Services. As the data warehouse evolves into a knowledge repository and as the e-enterprise forms strategic partnerships with customers, partners and suppliers, we envision intelligent agents assisting in KM tasks such as: 1) monitoring e-enterprise business processes for performance information, 2) consulting authoritative external ontologies to obtain proper terminology for metatagging, 3) collecting meta-data regarding objects that flow through enterprise processes, and 4) communicating with inter-enterprise processes and partners to coordinate information interoperability.

Mediation Services. Mediation refers to a broad class of services associated with the Intelligent Integration of Information (I³) [25]. Mediation services facilitate the extraction, matching, and integration of data from heterogeneous multi-media data sources such as maps, books, presentations, discussion threads, news reports, e-mail, etc.

One example is the mediation of temporal data of differing granularity. This is of particular importance in the context of multidimensional databases and data warehousing applications, where historical data is integrated and analyzed for patterns and interesting properties. A *temporal mediator* [26] consists of three components: 1) a repository of *windowing functions* and *conversion functions*, 2) a time unit thesaurus, and 3) a query interpreter. There are two types of windowing functions: the first associates time points to sets of object instances, and the other associates object instances to sets of time points. A conversion function transforms information in terms of one time unit into terms of some other time unit. The time unit thesaurus stores the knowledge about time units (e.g., names of time units and relationships among them). The time-unit thesaurus stores concepts such as the seasons, fiscal year definitions, and calendars, and translates these time units into others.

Users pose queries using the windowing functions and desired time units using a temporal relational algebra. To answer such a user query, the query interpreter first

employs the windowing functions together with the time unit thesaurus to access the temporal data from the underlying databases and then uses the time unit thesaurus to select suitable conversion functions, which convert the responses to the desired time units.

4. Conclusions

Knowledge-based support for decision-making is becoming a key element of an enterprise's business strategy. Traditional data warehouses will evolve into knowledge management environments that handle not only operational transaction-oriented data, but also semi-structured, heterogeneous information culled from external sources and integrated into decision-oriented knowledge for enterprise decision-makers. This paper proposes a knowledge management process model together with a collection of services that can be used to manage and encourage knowledge creation within the enterprise. The key to the successful implementation of enterprise knowledge management is top management support for the creation of a knowledge sharing and learning organization.

Acknowledgements

The research was sponsored, in part, by an infrastructure grant from the Virginia Center for Information Technology, and by the E-Center for E-Business industry sponsors. The author thanks Benjamin S. Kerschberg for helpful discussions.

References

- [1] R. Kimball, *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*: John Wiley & Sons, Inc., 1996.
- [2] L. Kerschberg and D. Weishar, "Conceptual Models and Architectures for Advanced Information Systems," *Applied Intelligence*, vol. 13, pp. 149-164, 2000.
- [3] J. M. Firestone, "DKMS Brief No. Nine: Enterprise Integration, Data Federation, and DKMS: A Commentary," Executive Information Systems, 1999.
- [4] T. Berners-Lee, R. Cailliau, A. Loutonen, H. F. Nielsen, and A. Secret, "The World-Wide Web," *Communications of the ACM*, vol. 37, pp. 76—82, 1994.
- [5] W. H. Inmon and R. D. Hackathorn, *Using the Data Warehouse*. New York: John Wiley & Sons, Inc., 1994.
- [6] R. Alonso, D. Barbara, and H. Garcia-Molina, "Data Caching Issues in an Information Retrieval System," *ACM Transactions on Database Systems*, vol. 15, 1990.
- [7] L. Seligman and L. Kerschberg, "A Mediator for Approximate Consistency: Supporting 'Good Enough' Materialized Views," *Journal of Intelligent Information Systems*, vol. 8, pp. 203 - 225, 1997.
- [8] J. Moad, "Extra Helping of Data," in *PC Week Online*, 1998.

- [9] D. A. Menascé, H. Gomaa, and L. Kerschberg, "A Performance-Oriented Design Methodology for Large-Scale Distributed Data Intensive Information Systems," presented at First IEEE International Conference on Engineering of Complex Computer Systems (Best Paper Award), Florida, 1995.
- [10] L. Kerschberg, H. Gomaa, D. A. Menascé, and J. P. Yoon, "Data and Information Architectures for Large-Scale Distributed Data Intensive Information Systems," presented at Proc. of the Eighth IEEE International Conference on Scientific and Statistical Database Management, Stockholm, Sweden, 1996.
- [11] H. Gomaa, D. Menascé, and L. Kerschberg, "A Software Architectural Design Method for Large-Scale Distributed Information Systems," *Journal of Distributed Systems Engineering*, 1996.
- [12] W3C, "Extensible Markup Language (XML); <http://www.w3.org/XML/>," 2001.
- [13] A. Motro and I. Rakov, "Estimating the Quality of Databases," in *Proceedings of FQAS 98 Third International Conference on Flexible Query Answering Systems, Lecture Notes in Artificial Intelligence*, vol. 1495, T. Andreasen, H. Christiansen, and H. L. Larsen, Eds. Berlin: Springer-Verlag, 1998, pp. 298-307.
- [14] A. Motro and P. Smets, "Uncertainty Management in Information Systems: from Needs to Solutions." Norwall, MA: Kluwer Academic Publishers, 1996, pp. 480.
- [15] J. Hendler, "Agents and the Semantic Web," in *IEEE Intelligent Systems*, 2001, pp. 30-37.
- [16] S. A. McIlraith, T. C. Son, and H. Zeng, "Semantic Web Services," in *IEEE Intelligent Systems*, 2001, pp. 46-53.
- [17] L. Kerschberg, "Knowledge Rovers: Cooperative Intelligent Agent Support for Enterprise Information Architectures," in *Cooperative Information Agents*, vol. 1202, *Lecture Notes in Artificial Intelligence*, P. Kandzia and M. Klusch, Eds. Berlin: Springer-Verlag, 1997, pp. 79-100.
- [18] L. Kerschberg, "The Role of Intelligent Agents in Advanced Information Systems," in *Advances in Databases*, vol. 1271, *Lecture Notes in Computer Science*, C. Small, P. Douglas, R. Johnson, P. King, and N. Martin, Eds. London: Springer-Verlag, 1997, pp. 1-22.
- [19] I. Nonaka and H. Takeuchi, *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*: Oxford University Press, 1995.
- [20] A. Scime and L. Kerschberg, "WebSifter: An Ontological Web-Mining Agent for E-Business," presented at IFIP 2.6 Working Conference on Data Semantics (DS-9), Hong Kong, China, 2001.
- [21] A. Scime and L. Kerschberg, "WebSifter: An Ontology-Based Personalizable Search Agent for the Web," presented at International Conference on Digital Libraries: Research and Practice, Kyoto Japan, 2000.
- [22] W3C, "Semantic Web Activity: Resource Description Framework (RDF); <http://www.w3.org/RDF/>," 2001.
- [23] B. Omelayenko and D. Fensel, "An Analysis of Integration Problems of XML-Based Catalogs fo B2B Electronic Commerce," presented at IFIP 2.6 Working Conference on Data Semantics (DS-9), Hong Kong, China, 2001.
- [24] D. Fensel, *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Berlin: Springer-Verlag, 2001.
- [25] G. Wiederhold, "Mediators in the Architecture of Future Information Systems," *IEEE Computer*, vol. 25, 1992.
- [26] X. S. Wang, C. Bettini, A. Brodsky, and S. Jajodia, "Logical Design for Temporal Databases with Multiple Temporal Types," *ACM Transactions on Database Systems*, 1997.