CONTENT AUTOMATA: CHALLENGES AND STRATEGIES

Chang-Yang Lin
School of Business
Eastern Kentucky University
521 Lancaster Avenue, Richmond KY, 40475
Phone: (859) 622-1087
Fax: (859) 622-8071

ABSTRACT

This paper explores unstructured data integration with the objective of planning an open infrastructure for integrating business content across heterogeneous data sources. Itreviews the unstructured data integration process, examines the challenges or requirements associated with content integration, and presents a high-level deployment strategy for content automata. This strategy,aimed at "aggregated" solutions for content projects, comprises three directions: (1) creating and maintaining relevant enterprise metadata; (2) integrating content through open standards; and (3) instituting effective content governance to enforce metadata management and content integration. This strategy will not only provide practitioners with insightson planning content automata but will also stimulate academics into doing prospective research.

Keyword: Access Virtual Integration, Foundation Integration, Metadata, Taxonomy

INTRODUCTION

Businesses are full of content—commonly appearing in emails, financial statements and reports, notes from customer support records, market research, web pages, medical images, call center audio recordings, and video presentations. Many business decisions are based on information and knowledge derived from these content sources. According to anAssociation for Information and Image Management survey, ninety-nine percent of respondents stated that content is involved in core business processes (AIIM, 2008).Nevertheless, most contentis locked in inaccessible data stores. Thus, content cannot be used to generate relevant business intelligence nor can it be used to facilitate true business analytics.

Most content is unstructured data. Despite the significant development devoted to automation techniques and technologies, only a small amount of content is already integrated and classified. The difficulty of creating and maintaining a unified semantic layer for unstructured data has been cited as one of the main reasons why organizations are not engaging in content automata (Jhingran, etc., 2002). Automating content so that it can be accessed, analyzed, and shared is becoming a critical factor for organizations in optimizing their business processes to stay competitive. The Aberdeen Group reported three main benefits that drive best-in-class companies in integrating and automating their content: improved employee productivity, reduced risks, and better customer insight (Brink, 2009). Regulatory compliance has recently become an additional driver due to the requirements of the Health Insurance Portability and Accountability Act, the Sarbanes–Oxley Act, and e-Discovery.

This paper explores unstructured data integration with the objective of planning an open infrastructure for integrating business content across heterogeneous data sources.Sections 2 and 3 review integration approaches involved in unstructured content and then examine the challenges of this integration. The specific requirements of employing integration techniques and technologies such as taxonomy, ontology, and enterprise content management (ECM) systems are also discussed. Section 4then employs an open infrastructure to present a high-level deployment strategy for content automata. Finally, Section 5provides a summary.

REVIEW OF CONTENT INTEGRATION

Two fundamental approaches of integrating data are consolidation and federation.Consolidation involves the capturing of data from multiple, disparate sources and integrating it into a single aggregated persistent data warehouse. During the integration process, the “Extract, Transform, Load” (ETL) technique is used to cleanse and standardize data.

Federation involves creating a unified virtual data view. This view does not contain data itself; instead, it builds a referenced metadata file to connect actual data. Enterprise Information Integration (EII) and Enterprise Application Integration (EAI) technologies are employed to implement federation. While EII focuses on data and querying it, EAI places an emphasis on specific applications allowing them to interoperate each other through their integrated schemas (Halevy, 2005). The integration efforts have been largely focused on structured data due to its well-defined semantic schema.

Unstructured Data Integration

By first tagging contextual information to unstructured data, consolidation and federation can be extended to content sources. Two integration approaches involving unstructured data are called foundation integration and access virtual integration (Inmon and Nesavich, 2008).

Foundation integration combines disparate unstructured data sources in their entirety, integrates them into structured content, and creates a single persistent content warehouse in a XML, relational, or object-oriented format. The content warehouse becomes a foundation to support queries from the requesting applications. An example of foundation integration technology is an ECM system, which stores its content into a content storeand addsa semantic layer on top of thiscontent store.

Access virtual integration is where content is accessed, gathered, and used to form a result based on an index that allows for a virtual mapping of content. This virtual integration does not require an explicit migration of content. Instead, it depends on specific techniques such as metadata, taxonomy, and classification to communicate with content sources.

These approaches can also be extended to the integration of unstructured data with structured data. In this integration, a mapping between content taxonomy and database schema must be defined. Enterprise data mashup automates the extraction of Web data and can structure content and relate that to enterprise structured data.

Content Integration Products

Many vendors have extended their products with a semantic option to integrate unstructured data. For example, PowerCenter from Informatica expands its ETL capabilities to include binary documents, flat files, and messages; WebSphere Information Integrator—an EII product from IBM allows queries to access a federated view of unstructured data; and Oracle Multimedia enables Oracle Database to store, manage, and retrieve images, audio, and video data in an integrated mode with enterprise data. The EII and ETL vendors are adding support for XML-based data to enablebusiness content interchange.

CHALLENGES OF CONTENT INTEGRATION

This section examines the challenges of content integration from variousstudies (Paganelli, etc., 2004; Halevy, 2005; Inmon and Nesavich, 2008; Baum, etc., 2013) including those by leading IT research firms (Unitas, 2002; Delphi Group, 2004). These challenges or requirements are summarized in Table 1.

Table 1: Challenges of Content Integration

	Challenges or Requirements	Goals
Metadata, taxonomy	▪ An undefinedcontent structure ▪ “Point” taxonomies	▪ A metadata ▪ A coherent taxonomy
Taxonomy vs. schema	▪ The difficulties of mapping content taxonomy and database schema	▪ An integrated view of taxonomy and schema
Volume, type, location	▪ The need to manage huge volumes of content in any format across heterogeneous data sources ▪ The need to make content being accurate, consistent, and auditable	▪ An open infrastructure to facilitate the development of aggregated solutions ▪ Data quality
Solutions	▪ “Point” solutions exist in silos	▪ Aggregated solutions
Content technologies	▪ Content tools continue to be nascent and difficult to use	▪ Mature content tools ▪ Different training

Metadata, Taxonomy, and Ontology

By its nature, unstructured content does not have an identified structure, making it difficult to directly extract information for integration. Adding semantic information into unstructured content is a key requirement toward content integration. Paganelli, etc., (2004)proposed a three-layered data approach to represent the context of use in organizations (i.e. who, where, how, under which role a document is accessed).

Effective content integration will also require taxonomies and ontologies. Taxonomies can be regarded as a classification scheme that is used to organize content objects into a hierarchical structure where content objects are placed. Taxonomiesoperate as a directory, providing a navigational path through a content hierarchy. To resolve the problems of semantic heterogeneity, ontologies are developed to provide formal descriptions of concepts and their relationships in a specific domain (Delphi Group, 2007).

Building taxonomies and ontologies remains a challenge to organizations because of the complexity and expert knowledge involved. Business professionals must have deep business backgroundsand be capable of using the taxonomy software. To date, learning the use of the taxonomy software has a steep learning curve.

Content Taxonomy vs. Database Schema

Very little amount of content is currently accessible by relational systems. The evidence implies the difficulties of defining the mappings between content taxonomy and database schema. Mapping taxonomies and schemas requires thoughtful planning because their structures are inherently different. While database schemas address table structures, content taxonomies are a flexible hierarchy. A variety of data types and data formats have also created further complexities to the mappings.

Velocity, Volume, Type, and Location

The exponential growth of big data, which is mostly unstructured data, adds to the challenges for content integration. First, big data imposes three basic requirements on data integration: the need to process huge volumes of data at high velocity in any format across heterogeneous data sources; the need to correlate big data with other enterprise data; and the need to integrate big data technologies (e.g., NoSQL) with relational technologies to streamline operations (Baum,etc., 2013). Further, making big data relevant requires it being accurate, consistent, and auditable.

Siloed Content Solutions

"Point" content solutions exist in silos in many organizations. These solutions are mostly aimed at short-term purposes and have their own requirements for structuring, processing, storing, and retrieving content. Their taxonomies are often a point scheme due to a lack of standards. Future content projects may create more silos and seek more point solutions by bringing in additional and inconsistent technologies.

Evolving Content Technologies

Content technologies continue to evolve, addressing such features as metadata, classification, taxonomy, ontology, mappings, integration techniques, ECM systems, and XML-based components. These represent relatively new technologies that require further and different training for professionals.

DEPLOYMENT STRATEGY FOR CONTENT AUTOMATA

The challenges—the sheer volume of unstructured data, its access velocity, its non-identified nature, and its complex taxonomy—have a clear message: integrating unstructured data into a single enterprise-wide persistent content warehouse is technically, economically, and managerially infeasible. Rather, the strategy aimed at “aggregated” solutions should be employed for a subset of business content. An open infrastructure enabling sharing, interoperability, and scalability must be established.

Under an open infrastructure, the deployment strategy comprises three directions: (1) creating an enterprise metadata; (2) integrating content through open standards; and (3) instituting effective content governance to enforce the above two directions. The following provides a high-level overview of these three directions.

Creating Enterprise Metadata

A single enterprise metadata that provides a unified view of content is an important enabler for content integration. The first step of creating an enterprise metadata is to add properties and contextual information into content objects. The properties may contain information such as date created, responsible person, synopsis, and key words. For new content objects, a policy must be in place to enforce the entry of metadata at their transactions; for existing content objects, a discovery and acquisition tool may be utilized to automate the process. In this step, content objects are digitized into their standard formats.The second step is to define the taxonomy structure depending on the context and to classify content objects into a hierarchy. The third step is to build ontologies. Ontological data sets, often containing many data items and relationships between them, can be modeled using W3C’s Resource Description Framework (RDF).

During the creation process, knowledge developers may employ software tools to perform methods of automatic categorization. The software tools should support semantic modeling standards such as RDF-schema (RDFs) and Web Ontology Language (OWL). To support interoperability, the software tools should be capable of representing and storing metadata including rules, classification schemes, taxonomies, and ontologies in a universal data format like XML.

Integrating Content through Open Standards

The development framework for content projects must be openness, platform independence, and consistent use of standards. The platform provides relational and XML content stores as well as access to a federation of content servers. The platform should support service-oriented architecture and a full range of operating systems.

Under the development framework, aggregated solutions can be developed using approaches like foundation integration, access virtual integration, or a hybrid of the two. These aggregated solutions can be integrated into the enterprise solution through open standards. The following describes content development with an emphasis on the use of common standards at different tiers.

At the back-end tier, content is migrated into a content store in XML. This content store may be deployed with an ECM system. The aggregated ECM solution enables the sharing of content and enterprise data. At the middle-level tier, the aggregated EAI solution provides a virtual content view by integrating aggregated taxonomies into enterprise taxonomy. Further, thisaggregated solution enables a combination of content taxonomy and database schema into a virtual data-content view. Both taxonomy and schema are represented in XML, which is used as Web Services for transferring data and metadata between data sources and application servers.

Instituting Effective Content Governance

Two aspects of content automata described above suggest an important role for content governance: the ability to create and maintain enterprise metadata; and the ability of the various content applications to access, analyze, publish, and distribute content. These two aspects are united to establish key guidelines and policies for content governance. The examples of guidelines and policies are listed under the “Instituting Effective Content Governance” columnas a part of the content deployment strategy in Table 2.

CONCLUSION

This paper suggests a strategy in which “aggregated” content solutions are developed. An open infrastructure facilitating the development of aggregated solutions must be built. Under this open infrastructure, the deployment strategy comprises three directions: (1) creating and maintaining relevant enterprise metadata including coherent taxonomies; (2) integrating content through open standards; and (3) instituting content governance to enforce metadata management and content integration projects. Table 2 provides a summary of this high-level deployment strategy.

Table 2: Summary of Content Deployment Strategy

	Instituting Effective Content Governance
Creating and maintaining relevant enterprise metadata	▪ Enforce the metadata entry process for content objects ▪ Represent metadata, taxonomies, and ontologies in XML ▪ Select semantic modeling tools for taxonomies and ontologies based on W3C’s RDF and OWL standards ▪ Develop “aggregated” taxonomies that can be integrated into an enterprise taxonomy ▪ Define ontologies for specific domain areas
Integrating content through open standards	▪ Use techniques and technologies that support organizational and industrial standards ▪ Techniques, technologies, and standards are built into the methodology of content development ▪ Define an aggregated virtual view between content metadata and database schema ▪ Employ Web services in XML ▪ Digitize content objects into their standard formats ▪ Use XML as storage format for content stores ▪ Select ECM products that support XML storage and provide connection to enterprise systems

REFERENCES

Jhingran, A., Mattos,N., and Pirahesh, H. (2002). Information Integration: A Research Agenda. *IBM System Journal*, 41(4), 2002.

AIIM. (2008). Enterprise Content Management and Content Management Solution.Retrieved from http://www.emc.com/collateral/analyst-reports/aiim-ecm-cms.pdf.

Halevy, A. (2005). Enterprise Information Integration: Successes, Challenges and Controversies: Introductory Remarks. *SIGACM-SIGMOD.*

Inmon, B., and Nesavich, A. (2008).*Tapping into unstructured data: Integrating unstructured data and structural analytics into business intelligence.*NJ: Prentice Hall.

Brink, D.(2009, June). *Securing Unstructured Data: How Best-in-Class Companies Manage to Serve and Protect.* Aberdeen Group.

Baum, D., Radzik, I, Hansen,D., and Adelberg, B.(2013, January).*Five New Data Integration Requirements and How to meet them with Oracle Data Integration*.Retrieved from http://www.oracle.com/us/products/middleware/data-integration/5-new-di-reqs-wp-1898900.pdf.

Delphi Group. (2004, June). Information Intelligence: Content Classification and the Enterprise Taxonomy Practice. Retrieved from http://www.delphigroup.com/whitepapers/pdf/20040601-taxonomy-WP.pdf.

Paganelli, F.,Khaled, O., Pettenati, M., Pirri,F., and Giuli, D.(2004) Designing a Metadata Model for Unstructured Document Management in Organizations. In *Innovations through Information Technology*, M. Khosrow-Pour, M. (Ed.).Idea Group.

Unitas. (2002, January). A Single View: Integrating Structured and Unstructured Data within the Enterprise. Retrieved from http://lsdis.cs.uga.edu/GlobalInfoSys/Structured-and-Unstructured-for-EIPs.pdf.

Bitzer, S., and Schumann, M. (2009). Mashups: An Approach to Overcoming Business/IT Gap in Service-Oriented Architectures. In Nelson, M., Shaw, M.,and Strader, T. (Eds.). *Value Creation in E-Business Management*, pp 284-295. Springer-Verlag Berlin Heidelberg.