A Business Case for Enterprise Content Integration using Ontology-based Content Analytics

Edward Curry [1], Bill McDaniel[1], Dmitry Shingarev[1], Milena C. Caires[1], Mark Leyden[1], Sean O'Riain[1], Karl Flannery[2], Sabrina Kirrane[2], Christopher Green[2], Brendan Walsh[2], and Liam O'Morain[1]

[1] Digital Enterprise Research Institute, National University of Ireland, Galway, IDA Business Park, Lower Dangan, Galway

[2] Storm Technology Ltd, IDA Business Park, Lower Dangan, Galway

firstname.lastname@deri.org, {kflannery, skirrane, cgreen, bwalsh}@storm.ie

Abstract. Content integration is a key challenge within an organizations Enterprise Content Management (ECM) strategy. In this paper we present the challenges associated with the integration of structured and unstructured information sources within a ECM. Content analytics is a viable approach to the integration of structured and unstructured sources. This paper provides an overview of a semantically powered integration approach to discover relationships between the structured and unstructured content. This is achieved using information association using ontology-based entity detection and disambiguation. The commercial potential of the technology is discussed in terms of its business value proposition and the market positing of potential products and services.

Keywords: Enterprise Content Management (ECM), Enterprise Information Management, Content integration, Content analytics.

1 Introduction

Enterprise Content Management (ECM) is a set of critical technologies that helps an organization capture, store, preserve, and deliver important content and documents related to organizational processes [AIIM '08]. Typically, an organization's ECM strategy or process is delivered by the combination of a number of different services including document lifecycle management, digital asset management, web content management, collaboration support, workflow/business process management, etc. When developing an ECM strategy an organization can choose to acquire a full ECM platform or to combine their current information assets in an ECM solution. Given the significant investment (in terms of cost, effort, and training) an organization can have in its legacy systems the latter approach is often the preferred route. Within such solutions Enterprise Application Integration (EAI), Enterprise Services Buses

(ESB), and Service Oriented Architecture (SOA) play an important role in their implementation and use [AIIM '08].

Enterprise Information Management (EIM), a term coined by Gartner, will play a central role in any ECM strategy. Effective and flexible EIM techniques enable the ECM process to solve unique business challenges, such as content sensitive information access, in a simple and straight forward manner. Information access is an important responsibility of EIM in relation to both structured and unstructured enterprise information sources. The flexible integration of these sources to support changing business processes is a key challenge within EIM. A promising approach for integration utilizes content analysis to identify relationships between the structured and unstructured content. By extending an ECM strategy to employ advanced content analysis services it can meet the requirement of flexible content integration for these sources.

In this paper we discuss the commercial potential for an integration service, using ontology-based content analytics, for structured and unstructured content that could be employed within an ECM strategy.

2 Business Value Proposition

Within an ECM strategy it is often necessary to associate related structured and unstructured information in order to meet the business requirements of the organization. This requirement can manifest in a number of ways:

- Unifying access to customer correspondence, e-mails, technical product documents, transcripts of voice conversations and other information for a customer service or help desk department needs to handle customer queries more efficiently.
- Associating unstructured job applications (the job seeker's CV) to the relevant job vacancy within a database.
- Linking unstructured product documentation (user manuals, marketing material, technical specifications etc) from various repositories to a structured product catalog.
- Merging multiple content systems as the result of a merger or acquisition to provide single point access for the multiple content stores of the combined organization.

Figure 1 illustrates a typical deployment where structured and unstructured sources need to be integrated. Within these deployments the following characteristics are common:

- Information is stored in multiple heterogeneous systems.
- Large volume of unstructured data that is constantly growing.
- Structured information frequently changes.
- The gathering of information reduces employee productivity.

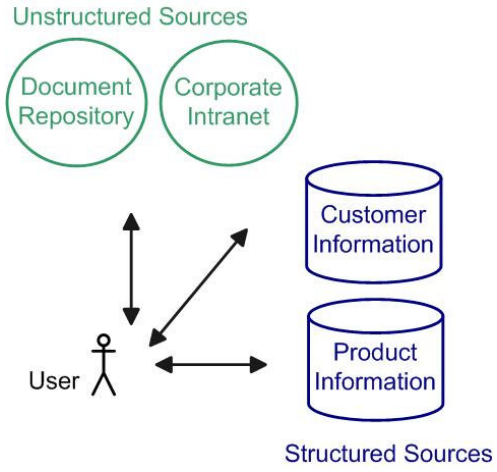

Fig. 1. A sample ECM scenario requiring the integration of structured and unstructured information sources.

Content integration may be achieved in this deployment using the following methods:

- Manual Integration – A human operator associates the content manually.
- Meta-data Integration - Meta-data, if it exists, that describes the associations can be used to integrate the content.
- Content Analysis Integration – Associations between the sources are created by examining both sources to discover relationships.

The first two options require the intervention of a human to either create the links directly or to create the meta-data from which links can be derived. These approaches create static links that are inflexible to changes within the business environment, require significant effort for large data sources, have limited capacity to cope with information growth, and come with a large associated maintenance cost.

With this in mind, the business value propositions of content analysis integration are:

- **Reduced Costs** – Discovering relationships via content analysis removes the labour cost of the manually approaches.
- **Improved Information Agility** – Content analysis integration is more agile in dealing with new content and changing content in a timely and cost effective manner.
- **Improved User Experience** – Content analysis services are more effective at handling large data sources and can provide greater coverage of the data sources. This leads to a higher quality service offering that can directly affect the user experience of the consuming application (i.e. within a product catalogue this can improve the sales experience).

- **Reduced Impact on Employee Productivity –** Streamlining the process of information gathering into a single point of access reduces the impact on employee activity.

3 Market Positioning

Content analytic-based integration offering discussed in this paper is part of the larger ECM market. According to Gartner, in 2007 the worldwide software revenue for the ECM market is worth approximately $2.9 billion. Gartner predicts that total software revenue from the worldwide ECM market will grow at a compound annual rate of 12.9% through 2011 [Shegda, Bell, Chin and Gilbert '07]. Because the capabilities of an ECM system can be unique to a particular organization their costs can vary. According to The Yankee Group the average price of an ECM for "midsize companies or divisions of large organizations spend $315,000 to $880,000 on selecting, implementing and maintaining such systems. Large organizations spend upwards of $1.7 million on enterprise-wide deployments." ['04] ECM suites can be priced as high as $10,000 per seat.

A key growth area within the ECM market is in information management. "Content technologies are steadily gaining more capabilities to integrate with, or handle some aspects of, structured data, as well as document-centric data. Gartner's vision for the evolutionary path of these technologies is called enterprise information management (EIM)" [Shegda, Bell, Chin and Gilbert '07]. As both database and ECM vendors IBM, Microsoft and Oracle have the potential to be leaders within this field, currently none of these companies offer a coherent vision for EIM.

Our technology is positioned to address this opportunity in analytics adoption. As a light weight scalable infrastructure it will allow the processing of huge information volumes with robust content-analytical functionality that can provide more informed decision making capability affecting everything from lower-level operational support to executive-level strategic planning.

Initially, the technology was positioned as a product to reside within a corporate intranet providing a private internal content-integration service. This standalone content integration service can be use for integration within ECM strategy and we are currently entering an industrial pilot of this service with a major Multi-National Company (MNC). The pilot is focusing on linking unstructured product documentation (user manuals, marketing material, technical specifications etc) from various repositories to a structured product catalog. However, the technology can also be positioned in the following ways:

- **Hosted Solution**: - In a similar manner to the Software as a Service model, the technology could be packaged as a hosted "Content as a Service" using a "pay as you use" pricing model, offering significant costs saving to client companies.
- **"Revenue Sharing"** with established ECM vendors: The major technology components could be licensed under a revenue sharing model to other third party vendors that may wish to avail of the technology benefits within their specific ECM suites. The revenue-sharing approach would exploit the established market-position and customer based of the ECM vendors. The technology would be of

interest to large document management vendors (EMC, Open Text, Xerox, etc) and database vendors (Oracle, IBM, Sun) looking to exploit the EIM potential of their offerings.

4 Technical Solution

Our solution to the content integration problem, illustrated in Fig 2, uses semantic techniques to discover relationships between the structured and unstructured content. This is achieved using information association using ontology-based entity detection and disambiguation.

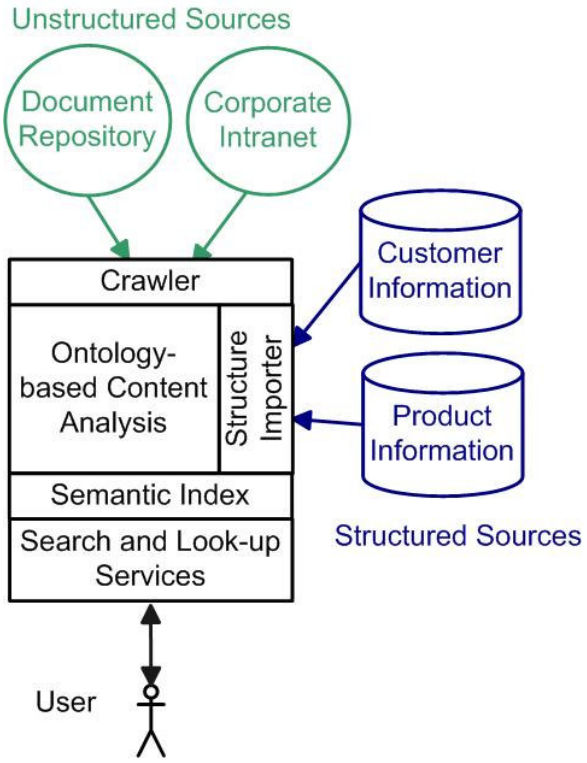

Fig. 2. A content-based integration service for integrating structured and unstructured information sources within a sample ECM scenario.

Within the scenario deployment, loosely-based on our industrial pilot, the structured sources are semantically described using ontology-based domain modeling. There is an initial cost associated with the definition of the ontology, and the creation of a structure (instance) importer. However, the cost of ontology definition can be reduced by reusing, where possible, the schema of the structured information sources. When

in place the maintenance cost of the ontology will be proportional to schema changes within the structured information sources.

Once the structured information is imported into the service the unstructured information sources are then crawled. The unstructured content is analyzed using the ontology of the structured information. The content-analysis employs techniques such as Natural Language Processing (NLP) and statistical analysis to discover relationships between the structured and unstructured information sources. Results from the content-analysis are stored in the semantic index. The semantic index can now be used to provide information management support to applications and users. Relevant product documentation can be linked to the product catalogue; document searches across the repositories can be enhanced with relevant context (product/customer details) from the structured information sources.

Advantages of content-based integration include increased consistency with more comprehensive document lists. The solution also reduces maintenance costs associated with the manual approaches while allowing repositories to be indexed more frequently (weekly/monthly).

5 Conclusion

Content integration is a key challenge within a organizations Enterprise Content Management strategy. Content analytics is a viable approach to the integration of structured and unstructured information sources. In this paper we present a semantically powered approach for integrating structured and unstructured content sources by discover relationships between the structured and unstructured content. This is achieved using information association using ontology-based entity detection and disambiguation.

The business value proposition of the technology has benefits in the areas of reduced costs, improved information agility, improved user experience, and reduced impact on the productivity of the employee. The market positing of potential products and services include a standalone product (entering an industrial pilot), a "Content as a Service" hosted solution, and the licensing of the technology to third party vendors, under a revenue sharing model, for inclusion within their ECM suites.

Acknowledgments. This work is supported by the Lion project supported by Science Foundation Ireland under Grant No. SFI/02/CE1/I131.

References

1. Understanding the TCO of a Hosted vs. Premises- Based ECM Solution. The Yankee Group (2004)
2. AIIM: Building an ECM Strategy - An AIIM White Paper Alternatives and Decision Points. Association for Information and Image Management (2008)
3. Shegda, K.M., Bell, T., Chin, K. and Gilbert, M.R.: Magic Quadrant for Enterprise Content Management. Gartner, Inc (2007)