Efficient Information Access in Enterprise Content Management Systems using Semantic Technologies

Raj Kumaran and Lixin Tao
Seidenberg School of CSIS, Pace University, White Plains, New York

Abstract — This research attempts to meet some of the business challenges faced in today's enterprise in information access by combining the Enterprise Content Management and Semantic web technologies. Searching for information or information access is often a frustrating and a challenging task within an enterprise. Previous research clearly highlights this as an area of improvement. Loss of employee productivity due to inefficient information access is an area that begs further research and enhancement. Inaccessible or disunited content hampers timely decisions and innovation. In this research Drupal content management system is extended using semantic technologies to pave way for efficient information access. Some tangible benefits derived from this implementation are demonstrated. Ontology, a key component of semantic technologies, is used as backbone to accomplish these benefits. By categorizing, classifying and linking contents based on ontology with enough forethought, better information access is gained.

Keywords— CMS, Content Management System, ECM; enterprise content management; enterprise search; ontology; semantic technology; Drupal, taxonomy.

I. INTRODUCTION

The Association for Information and Image Management (AIIM), an organization that provides independent research, education and certification programs to information professionals, defines Enterprise Content Management as below:

“Enterprise Content Management (ECM) is the strategies, methods and tools used to capture, manage, store, preserve, and deliver content and documents related to organizational processes. ECM tools and strategies allow the management of an organization's unstructured information, wherever that information exists.”

Document Management, Imaging, Records Management, Workflow, Web Content Management and collaboration all fall under the big umbrella of ECM. Neither a single technology nor a tool can meet these needs. ECM systems systematically capture and manage all the unstructured information in an enterprise to support key organizational processes. The information managed exists in many different digital forms such as text documents, video and audio files, e-mails, spreadsheets, slide shows and other proprietary files.

It is imperative to understand what the term “content” really means and how to distinguish it from data and information. Data is a small snippet of computer information. Computers are built to process data. Databases are invented to store vast quantities of these snippets. Data lacks human meaning. On the other hand content is also information but it has human meaning and context. Raw information that passes casually is not content. However when the information is given a usable form and accessibility to intended audience, it is called content. Usually the information becomes content by adding a data layer around it. This data is called metadata. Metadata typically describes the information and provides a context for it; in doing so it makes the information usable by computers and beneficial for people. For an example, the title of this paper is a piece of data. When we gather all the titles to create a table of contents, incorporate in the beginning of the “Proceedings of Michael L. Gargano 12th annual Student-Faculty Research day” readers use it to access this paper, it is a piece of content.

Functionality provided by ECM:

• Optimizes access to information
• Enhances collaboration and communication
• Supports business agility
• Meets regulatory compliance and corporate governance
• Ensures document integrity, data quality and security

II. CHALLENGES IN ENTERPRISE CONTENT MANAGEMENT

Content is growing within an enterprise at an alarming rate. As per one study the enterprise content doubles every 9 to 12 months [1]. User generated contents via blogs, forums and wikis play a major role in this prolific growth. However, the enterprises don’t have a well-planned organization to maintain their content. Enterprises have their content in multiple
disparate repositories spanning across file systems to database. This gives rise to an unstandardized methodology to manage content.

This issue presents major challenges during information access. Research shows 20% of the time is being lost in searching within the enterprise. That is a loss of 1 day in 5 days week. Searching against the various disparate systems of different types ranging from content management systems to files systems to database is in itself a herculean task.

In order to avoid these content silos, the content are centrally managed via Enterprise Content Management (ECM) systems. In ECM content is generally marked or tagged with keywords for ease of search. However if the contents are not tagged properly in ECM, the information access issue still remains unresolved.

The reason why the content is not properly tagged is multifold:

- Content tagging is voluntary and discretionary.
- Subjective and inconsistent tagging: unstandardized, non-hierarchical, random words are used for tagging.
- Heterogeneous tagging as they originate from different people.
- Tags fall short in capturing exact meaning and context due to language ambiguity.
- Tagging requires extra effort.

This leads to mismanaged information that is hard to be accessed and acted upon.

In addition, the traditional key word search presents its own limitations with the quality of the results. The two fundamental issues with key words search are as below:

- A same term can have more than one meaning, and
- More terms can have the same meaning. [5]

Semantic technology addresses many of the issues described above and paves way for an effective enterprise content management. Especially within an enterprise the content can be governed by a common ontology. When the contents are categorized based on ontology, the semantics get added to each piece of content in the form of metadata. The metadata added thus by, can be leveraged at the time of search, and navigation to bring meaningful experience to users.

This ontology based tagging has many added advantages. As the terms in ontology are arranged in a hierarchy from most generic to the most specific, the search can be broadened or narrowed depends on the need. The contents can be linked based on the relationship between the terms. Content can be selectively syndicated with respect to certain terms.

The key challenges from business perspective are:

A. Information access (i.e findability/search)

Employee productivity is severely impacted due to lack of an efficient information access system. The user needs to initiate new searches each time he wants to navigate from one piece of content to another related content. ECM can anticipate this need and connect the logically related content items. This is more important when the connection between those content items is elusive or subtle for human eye.

Use case

A user visits a company website and looks for different types of life insurance, but he isn’t necessarily sure what types of products are offered in this particular company. He searches with a key word “life”. The semantic powered ECM is smart enough to anticipate user’s need and present the results as well as options to narrow the search by presenting the links to different types of life insurance as shown in the left navigation bar of the following diagram.

![Figure 1 – Search Results with Navigation](image)

B. Document Integration

When information is sought, often the results presented are not in any logical order. Considerable effort is required to glean and synthesize the results to compile an actionable document. ECMs can address this common problem by integrating the content fragments in predetermined order and present a document that is readily actionable.

Use case

A user X looking for information about a product searches and locates different types of that products information. She/he is burden with the task to synthesize the information in appropriate order before utilizing it. Below is a screen shot how our implementation of ECM gracefully handles this scenario.
C. Consolidation of content after company mergers

Mergers and acquisitions present a slew of challenges. The cultural gap between the two companies needs to be bridged. There are processes, systems and terminologies that need to be merged. One of the main challenges stems from their ontology (the terms that describe the domain knowledge and the relationships). ECM can reconcile the discrepancies in the ontology and enables a seamless integration of ontology.

Use case

Company A acquires Company B. Both companies have their own set of content. However their ontologies used to describe their domains do not have an exact match. Let’s assume both companies provide automobile insurance. A user from company A originally looking for key word “auto” in company B doesn’t find it, as this concept is widely known as “vehicle” within company B. The language ambiguity of calling a same concept with different words leads to this issue.

The ECM system can reconcile these differences behind the scenes and provide a seamless experience to the user whether they search a term used in company A or company B as long as the meaning of the terms are one and the same. Below is the search output of term “vehicle”. Under the hoods implementation ensures vehicle and Auto are synonymous and provide a desired user experience by surfacing results that are tagged either with “vehicle” or “auto” or any other term of similar meaning.

III. ENTERPRISE CONTENT MANAGEMENT PLATFORMS

There are various ECM systems ranging from proprietary viz., IBM FileNet, Documentum, Adobe, Open Text Vignette, SDL Tridion to open source CMS (Content Management System) such as Drupal, Joomla etc. Content is authored and managed in these systems.
Drupal is a leading open source content management system and used by many organizations and individuals to build engaging, content-rich websites [7]. It is both an ECM system as well as a website creation tool. The power and flexibility of Drupal is widely acknowledged. This research uses Drupal as a proof of concept environment to carry out the experiments.

Especially, taxonomy module in Drupal is an important part of the information architecture. It is incredibly powerful and one of Drupal’s greatest assets as content management system [7].

Ontologies can be flattened into taxonomies and imported into Drupal. This enables the content to be linked and searched using semantic concepts. This is a key distinguisher in organizing the content and leveraging the value of it later.

IV. EFFICIENCY GAINS

A. Linking content

When a user performs a key word search and selects one of the presented results, additional related topics need to be shown. These additional results are linked to the original item.

1) Simple linking

Based on the tagging used in a content, present additional topics

![Figure 5 – Simple Linking](image)

2) Deep linking based on class hierarchy

![Figure 6 – Linking based on class sub-class relationship](image)

3) Linking based on related concepts

![Figure 7 – Linking of subtle relationships](image)

Benefits

- Match the search results to the intent of the user
- Navigate the user to sub topics
- Navigate the user to related topics

B. Automated Document Integration

In an enterprise, the information about a particular project, product or platform is maintained by multiple departments. Each department is responsible for certain part of that information. For example, application development team has the list of application functionalities, infrastructure team has the information about the servers, and software license etc, operational support team has knowledge about incident and problem management. When a user looks information about this platform in an enterprise content management system, they are often presented with results that is at best fragmented and unsorted and at worst irrelevant and inundating. It will be beneficial if the retrieved information can be processed to provide a holistic view about this platform. All underlying
details such as filtering the appropriate documents, properly ordering them, consolidating them can be transparent to the user. This would be a value added and readily actionable document.

**Figure 8 – Automatic Document Integration**

1) **Benefits**

- Most up to date
- Filtered by relevancy
- Ordered logically
- Presented in an actionable form

**C. Ease of Content Integration**

Further, today’s business demands include frequent mergers and acquisitions. When content repositories from two different enterprises are merged challenges are inevitable. Enterprises may use different terms to denote a same concept. For example, an auto insurance enterprise may use the term “auto” whereas another could use the word “vehicle” to refer the same concept in their ontology. Traditional key word based search would not be able to understand the semantics behind these terms and intelligently extract documents containing either word and compound the results.

In contrast, semantic search would treat these two terms equally and surface the articles containing both in search results when only one term is used for search. A user searching for “auto” would not only get the documents that contain the term “auto”, but also the ones that contain the term “vehicle”. This is achieved by the underpinnings of Semantic web technology.

**Figure 9 – Ontology based mapping of key terms**

During search users will be provided an opportunity to refine their search based on the relevant fragment of the ontology.

Drupal content management system will be used to create and tag the content appropriately. The tagging will be governed by ontology of terms specific to the enterprise’s business.

1) Need for multiple Ontology support
2) Ontology reconciliation

a) Ontology de-duplication
b) Ontology Mapping

**V. METHODOLOGY**

Enterprises can mitigate these issues to some extend by mandating tagging in CMS. The CMS can be further enhanced by tagging with taxonomy of key words. This approach offers great benefits by providing accuracy, consistency, and standardization. This solution can be further enhanced by creatively combining the underpinnings of semantic technology within CMS.

Many features of semantic technology can empower the CMS and eventually result in a set of robust applications in which the CMS is used as a backend to retrieve the content.

First, semantic search addresses many of the issues faced in traditional search. Especially within an enterprise the content can be governed by a common ontology. The terms taxonomy and ontology are frequently confused and often used interchangeably. However there exists a marked difference between them [8]. Taxonomy describes a class – sub-class relationship, whereas ontology describes the whole domain by encapsulating many relationships of the terms.

When the contents are categorized based on ontology at the time of creation in a CMS, the semantics get added to each piece of content in the form of metadata. The semantic wrapper added thus by, can be leveraged at the time of display to bring meaningful outcome to users.
This ontology based tagging has many added advantages. As the terms in ontology are arranged in a logical hierarchy from most generic to the most specific, the information access can be broadened or narrowed depending on the need by traversing up or down in the ontology hierarchy.

This research uses Drupal content management system and Apache Solr search engine. Domain ontology is created using Protégé. The ontology is then flattened to taxonomy and imported into the Drupal implementation. Many additional contributed modules are added to the core installation to support the research ideas. Content is created in the Drupal CMS and later accessed to demonstrate the power of semantic technology. As an empirical study, the semantic powered Drupal system we developed is compared with a standard Drupal system in terms of information access.

![Figure 10 - POC environment product stack](image)

VI. MAJOR CONTRIBUTIONS

The major contribution of this research is identification of areas of improvement in today’s ECM solutions and display of a prototype how the ECM can be enhanced using semantic technology to add value to the enterprise.

This prototype is developed using Drupal Content Management System to demonstrate the benefits derived. Following are some of these benefits:

- Enhanced Information Retrieval that has better
  - Accuracy,
  - Consistency,
  - Standardization
- Ease of Content Integration during mergers and acquisitions.
- Faceted search for ease of navigation
- Consolidated document generation.

VII. CONCLUSION

As a primary source of origination of content, the current ECM systems can go a long way in meeting the challenges faced in an enterprise. Content is growing at an alarming rate. It is often referred as “content is king”, as content captures the intellectual property of an enterprise. However research and creative solutions are inevitable to fully exploit the treasure buried in the content. This research attempts to portray some ways how the CMS can take advantage of the semantic technology to mitigate the issues faced in today’s industry. It specifically addresses how information access within an enterprise can be improved and harnessed as means to increase productivity within an enterprise and position itself for competitive advantage.

ACKNOWLEDGMENT

My sincere gratitude to Dr. Tao for his patient guidance and kind support during my research efforts. Thanks to my colleague Srinivasan Shanmugam for constantly reminding me the task ahead of us and motivating me to make progress.

REFERENCES


35th IEEE Annual Computer Software and Applications Conference


[13] Chris Mavergames, Silver Oliver, Lorne Becker, "Systematic Reviews as an interface to the web of (trial) data: Using PICO as an ontology for knowledge synthesis in evidence-based healthcare research".

[14] Konstantinos Christidis and Gregoris Mentzas, Social Networking and Collaboration, Supercharging the Enterprise 2.0, "National Technical University of Athens Dimitris Apostolou, University of Piraeus


C1.7