



>> Mphasis Application Note 001 - January 2008

[mphasis.com](http://mphasis.com)

## Content Metadata Generation Using Enterprise Search

### Chirag Gandhi

Content metadata provides a huge value by enabling enhanced manageability and findability. However, enforcing authors to provide this information reduces the operational efficiency leading to the suboptimal usage of the content management solution. This negatively impacts the ROI and user feedback. One of Mphasis' clients was faced with this problem. On deeper analysis we discovered that we could provide a cost effective solution by reusing an Enterprise Search Engine to generate metadata. Plugging this into a content editorial workflow provided the client with a balance between automation and control.

### Problem Statement

Enterprises frequently need to implement a content management solution to handle the product related documentation being published for their clients. In such cases the content management solution needs to address two main goals:

- a. Improve the efficiency of the back and middle office by providing streamlined workflows
- b. Improve the findability for the customers by allowing the system to control the metadata being provided by the authors

A frequent challenge faced by clients implementing a content management solution is the level of enforcement to place on the metadata. Too little metadata leads to suboptimal findability, while too much metadata reduces the operational efficiency. In many situations the team tasked to manage the content might not be the authors of the content – in such cases, expecting this team to provide contextual metadata like keywords would mean that each document would have to be read and understood. Such activities would lead to further reduction in operational efficiencies.

Commercial products offering automated categorization are fairly mature, providing a combination of rule based and learning based solutions. The accuracy of these tools is around 60% though in controlled circumstances they can be tuned to achieve nearly 80% accuracy. Introduction of the tools increases the licensing cost and adds to the maintenance effort. Additionally, these tools require specialized skills which would mean additional investment by most clients. The overall effect is an increase in the adoption barrier for the client.

## Solution Description

Content metadata is broadly classified as:

- a. Asset Metadata – who, where and when is the content being published (e.g. Author, Publish Date, etc.)
- b. Use Metadata – how can it be used (e.g. Access Control Lists, Time of Access, Expiry Date)
- c. Subject Metadata – what does the content contain and why (e.g. Key words, Description, Subjects, etc.)
- d. Relation Metadata – is the content related to other content assets (e.g. Related Documents)

Asset and Use metadata are important

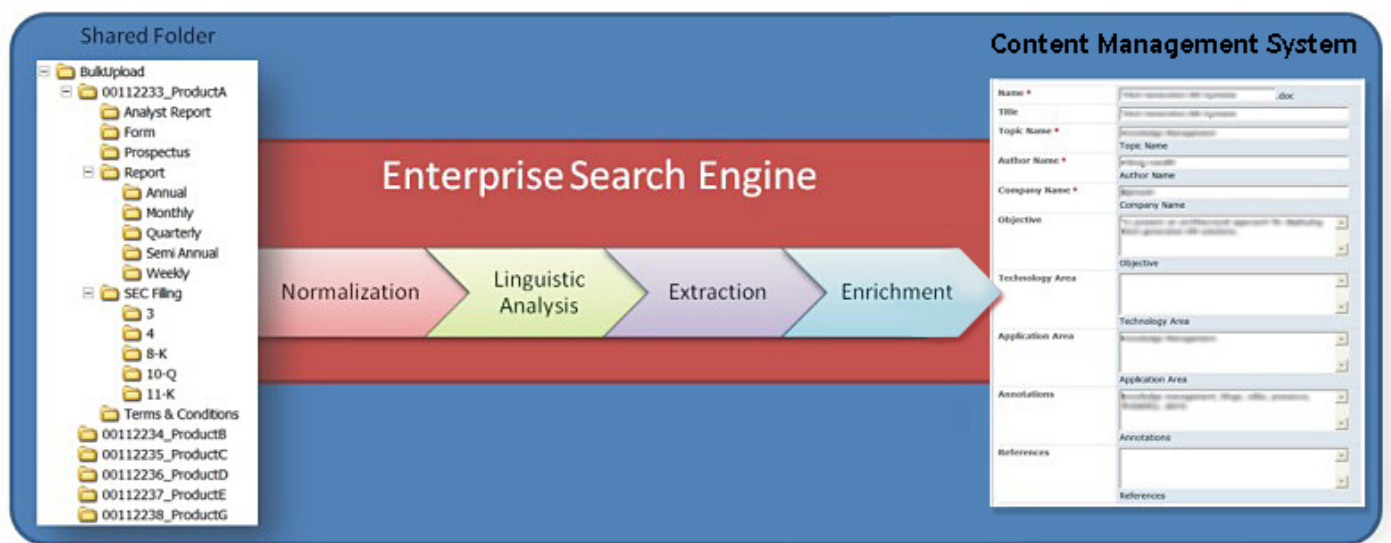
algorithms to perform the required analysis on the content.

Concept Search is a feature wherein the search engine provides the means for users to search for concepts rather than just keywords. To achieve this, the engine performs linguistic analysis of the document discovering the key phrases and cepts using various statistical models. The algorithms consider the grammar and structure and make use of thesauri.

We designed a solution which would utilize the capabilities of an Enterprise Search product. Operationally, this requires a shared folder to be created

processor needs to be configured to perform the necessary format (both physical and logical) normalization. Once the document has been normalized, the document processor can then perform linguistic analysis. During the linguistic analysis the subject metadata is generated. Once the linguistic processing is complete, the document is ready for the extractors. These extractors will extract the remaining metadata from the file properties, access control lists and in some cases rule based content analysis. Further enrichment of the metadata is done by using content matching capabilities of the enterprise search engine to generate the relation metadata.

Figure 1: Logical Solution Design



for managing the content repositories, while Subject and Relation metadata add tremendous value to the findability of the content. While Asset and Use metadata capture can be automated, the Subject and Relation metadata requires introspection and analysis of the content. Enterprise Search products capable of “Concept Search” possess the necessary

where the content can be “uploaded”. The uploaded content needs to be placed in an organized manner – the organization being achieved by a hierarchy of folders.

Once a “document” is placed into the correct folder; it is pushed to the Enterprise Search document processing system. The Enterprise Search document

The metadata laden payload is then exported from the enterprise search engine and the appropriate fields in the content management system are populated. Once in the content management system we expect an editorial workflow to allow the middle office to approve the metadata prior to the content being published.

## Case Study

One of the world’s largest retail fund houses’ Japan Office commissioned MphasiS to build the next generation customer facing portal. During the initial study it was determined that findability would be an extremely important criterion for success. In order to achieve the improved findability the “keywords” and “related documents” features were envisioned. FAST ESP was selected as the enterprise search engine and Interwoven’s TeamSite was the content management system of choice.

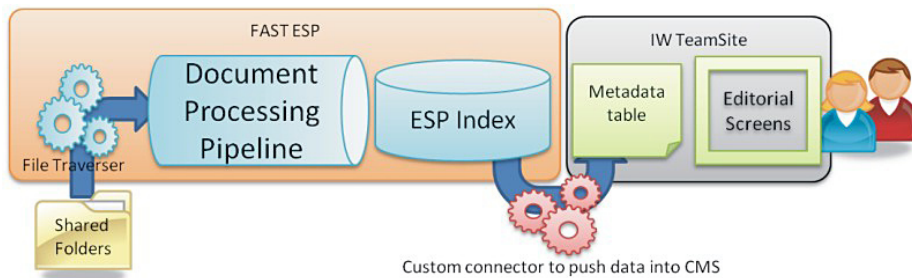
the metadata; this was piloted in a timeframe of two weeks.

A shared folder was created to act as the root of the “bulk upload” repository. FAST’s File Traverser was used to push the documents to the FAST Document Processing Pipeline. All the documents were provided by third parties; hence it was not possible to depend on the file properties to extract metadata. Author and title information had to be implemented using custom logic in the document processing pipeline. All the normalizations were performed using out-of-box

## Conclusions

One of the primary reasons for the failure of organizations to manage their content effectively has been the dependency on rich metadata. The metadata collection requirements led to strict rules for the authors, which in turn led to restricted usability of the solution. Additionally, enhancement of a solution would mean expensive manual rework on the existing content. While specialized tools exist to alleviate some of the tedium of these tasks, the tools themselves need specialized skills to be managed efficiently. Enterprise Search Engines are becoming a part of the core infrastructure for any organization today. Some of these engines possess the analytical algorithms to process linguistic information and provide a search framework based on concepts. We strongly believe that such search engines, in addition to providing Findability solutions, can be repurposed to provide an assistive approach to addressing the metadata enrichment concerns of the enterprise.

Figure 2: High Level Design for Client Implementation



The team handling the content already had a herculean task. With every new fund over 100 documents needed to be published. The client projected the increase in new fund launch frequency from two funds a month to two funds a week. Most existing funds were already providing updates to at least one “report” every week. The projected increase in operation spend was starting to look exponential.

Initially Interwoven’s MetaTagger had been considered, but was rejected as it required skills which the organization did not possess. Subsequently, MphasiS proposed the use of the enterprise search engine for the automated enrichment of

components of the FAST document processor. The keywords were generated using FAST’s Keyword generator. A custom component was created to use FAST’s “find similar” feature to populate the list of related documents.

A custom component was created to query the FAST Index, extract the search result, transform it and load it into the custom metadata table in Interwoven. Interwoven workflows were created to provide the required editorial functionality.

The solution has been a success and has been demonstrated to other regions for a wider adoption.

## About Mphasis

Mphasis supports G2000 companies around the world in the improvement of their business processes. Our unique strength lies in the intersection of our Information Technology (IT) and Business Process Outsourcing (BPO) capabilities. While our forte is in the BFSI (Banking, Financial Services, Insurance) and Technology industries, our focus extends to the Healthcare, Mobile, Logistics, Telecom, Life Sciences, Consumer Electronics and Utilities industries. The convergence of technologies such as web services, workflow software and business performance monitoring along with business intelligence and customer focus drive all our services delivery offerings.

For more information on a Pilot Project option, write to [vtr@mphasis.com](mailto:vtr@mphasis.com)

---

## Contact us

### USA

Mphasis  
460 Park Avenue South  
Suite # 1101, New York  
NY 10016, U.S.A.  
Tel: +1 212 686 6655  
Fax: +1 212 686 2422

### UK

Mphasis  
Capital Tower  
91 Waterloo Road  
London, SE1 8RT, UK  
Tel: +44 0 203 1700 954  
Fax: +44 0 203 1700 950

### INDIA

Mphasis  
The Millenia  
Tower A & B, 1 & 2  
Murphy Road, Ulsoor  
Bangalore 560 008, India  
Tel: +91 80 2556 7500  
Fax: +91 80 2556 7515

Mphasis  
Edinburgh House  
43-51 Windsor Road  
Slough SL1 2EE, UK  
Tel: +44 0 1753 217700  
Fax: +44 0 1753 217701