Semantic Content Management with Apache Stanbol

Ali Anil SINACI and Suat GONUL

SRDC Software Research & Development
and Consultancy Ltd.,
ODTU Teknokent Silikon Blok No:14, 06800 Ankara, Turkey
{anil,suat}@srdc.com.tr

Abstract. Most of the CMS platforms lack the management of semantic information about the content although a lot of research has been carried out. The IKS project has introduced a reference architecture for Semantic Content Management Systems (SCMS). The objective is to merge the latest advancements in semantic web technologies with the needs of legacy CMS platforms. Apache Stanbol is a part of this SCMS reference implementation.

Keywords: Apache Stanbol, Interactive Knowledge Stack, IKS Project, Semantic Content Management Systems

1 Introduction

Interactive Knowledge Stack (IKS) [1] is an FP7 research project targeting to Content Management System (CMS) providers in Europe so that current CMS frameworks gain semantic capabilities. Most of the CMS technology platforms do not address *semantic* information about the content, hence lack the intelligence [2]. Therefore, today's implementations cannot provide the interaction with the content at the users's knowledge level. The objective of IKS project is to bring semantic capabilities to current CMS frameworks. IKS puts forward the "Semantic CMS Technology Stack" which merges the advances in semantic web infrastructure with the needs of European CMS industry through coherent architectures which fit into existing technology landscapes.

Apache Stanbol [3] has been created within the Apache Software Foundation to meet the requirements addressed by the IKS project for *Semantic* Content Management Systems. Apache Stanbol is an open source modular software stack and reusable set of components for semantic content management. Each component provides independent and integrated services to be used by the CMS vendors/developers. The components are implemented as OSGi [4] components based on Apache Felix [5]. Moreover, all components can be accessed via RESTful service calls.

2 Research Background and Application Context of the Demonstration

CMSs needs to deal with huge amount of unstructured data. In recent years, many studies have focused on automatic extraction of *knowledge* from unstructured content. Inline with this, several advancements and algorithms have been developed in the field of Information Extraction (IE) and Information Retrieval (IR). IKS project has focused on the integration of these latest techonological foundations and proposed a reference architecture for Semantic Content Management Systems (SCMS) [6]. A SCMS is a CMS with the capability of extracting and managing semantic metadata of the content items through different components for specific tasks according to the layered architecture.

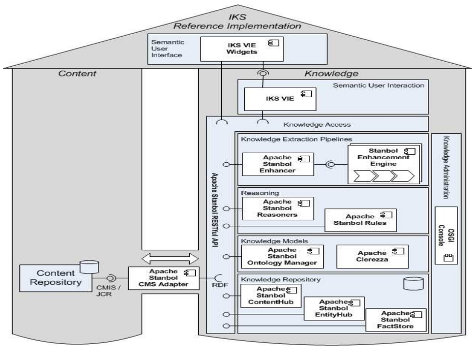

Fig. 1. SCMS Reference Implementation: Apache Stanbol [7]

Apache Stanbol is the reference implementation for the Knowledge Access part of the reference architecture if a SCMS. The relation can be visualized as in Fig. 1. Several Stanbol components are implemented to serve in different layers of a SCMS.

In the application context of the demonstration, a legacy CMS (such as Apache Jackrabbit [8]) makes use of several componenets of Apache Stanbol (e.g. CMS Adapter, Enhancer) to semantify existing unstructured content and then manages the knowledge through other components (e.g. Contenthub, Entityhub). This enables intelligent content management, categorization, semantic search and powerful faceted search mechanisms, hence turns a CMS to a SCMS.

3 Key technologies and Relation to Pre-existing Work

Components of Apache Stanbol implements latest advancements in semantic systems area and provides an integrated, comprehensive use for the users (CMS vendors/developers). The OSGi model which is adopted by Apache Stanbol supports elegant separation of different components required by the Knowledge column (Fig. 1). In addition, each component exposes its interfaces in terms of REST API.

Stanbol Enhancer implements Knowledge Extraction Pipelines through the Enhancement Engines. Enhancing the unstructured content stems from recent approaches [9] such as named-entity recognition, clustering and classification algorithms. Each Enhancement Engine processes the unstructured content (in addition to the results of other engines) and adds semantic information to the metadata of the content. Natural Language Processing [10] based engines extract valuable knowledge such as person, location and organization entities from the unstructured content. Extracted knowledge is represented in a triple-graph provided by Apache Clerezza [11] and persisted through Stanbol Contenthub.

Stanbol Entityhub is used to retrieve semantic information about the entities available through Linked Data sources such as DBpedia [12]. Independent domain ontologies can also be registered to Entityhub to create a new source for entities.

Stanbol Contenthub provides services to manage the knowledge on top of the content items. The Contenthub makes use of Apache Solr [13] as its backend to store the knowledge. Indexing through Apache Solr is performed through a number of configuration files (Solr cores). To give an example for simple semantic search, in the default index (default Solr core) of Contenthub, if a submitted document includes the keyword "Istanbul", then the country information "Turkey" and the regional information "Marmara" are indexed along with this document as its knowledge. This leads to much more accurate search results over the knowledge of Contenthub.

LDPath [14] is a valuable outcome of Linked Media Framework (LMF) Project [15]. LDPath is a simple path based query language over RDF (similar to Xpath or SPARQL Property Paths) which is particularly designed for querying the Linked Data Cloud by following RDF links between resources. To be able to support domain specific indexing, LDPath has been integrated into the document submission and search processes of Contenthub. After LDPath integration, the semantic indexing performed by Contenthub is realized in a much smarter way so that adopters from CMS industry can enhance their legacy content with these semantic storage and search capabilities according to their specific needs.

Stanbol CMS Adapter acts as a bridge between existing unstructured content of legacy CMSs and Apahce Stanbol. CMS Adapter enables connecting content repositories through standard interfaces like JCR [16] and CMIS [17]. Furthermore it provides RESTful services to enable content repositories to submit their content models. CMS Adapter enables extracting already available semantics in the CMS into an ontology and store it in the knowledge base of Stanbol through Contenthub.

4 Demonstration and Benefits to the Audience

The demonstration presents the use of several Apache Stanbol components. A JCR based CMS connects to Stanbol through CMS Adapter and performs several semantic operations such as entity extraction from free text through Enhancement Engines, management of such entities from Linked Data cloud, management of ontologies in different formats (RDF, OWL) and management of the extracted knowledge.

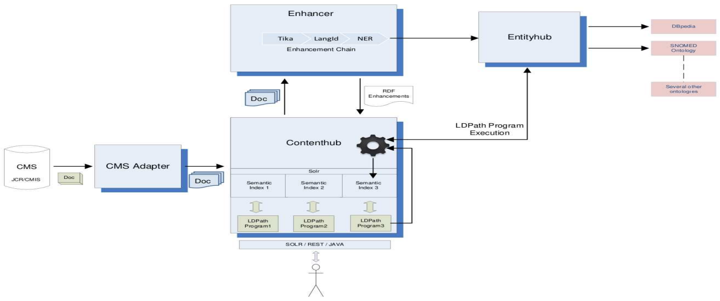

Fig. 2. Interaction between Apache Stanbol Components

Since the adoption of semantic advancements are poor in the CMS arena, the demonstration provides a proof of what can be done with semantic technologies in real world. A use-case can be summarized as follows while the interaction between several Stanbol components can be followed from Figure 2:

- A CMS (such as Apache Jackrabbit) gives the necessary information to Stanbol so that CMS Adapter can connect and retrieve the content items from the underlying JCR repository.
- CMS Adapter analyzes the underlying data model of the CMS and generates an RDF based ontology, and submits to Stanbol knowledge base.
- CMS Adapter submits all content items inside the CMS to Stanbol Contenthub.
- Contenthub submits the text-based content to Stanbol Enhancer, retrieves the enhancement results.
- Enhancer makes use of Entityhub while identifying the entities such as people (e.g. Dennis Ritchie), locations (e.g. Tokyo) and organization (e.g. European Commission). Stanbol comes with DBpedia as the default entity source. Any other ontology can also be easily registered to Stanbol system.
- Contenthub manages several *semantic* indexes on Apache Solr by means of LDPath programs. The knowledge to be indexed in the knowledge repository is extracted through the execution of LDPath.
- Contenthub provides *semantic search* on the content items. For example, if a keyword is related with an entity in a document, the document can be found even if the content does not include the keyword.
- Contenthub provides faceted search to refine search results.
- Contenthub provides a “tokenization” service for the queries. The entities are extracted from the query string and the search is directed in a more intelligent way with these tokens.

– Contenthub makes use of Wordnet, domain ontologies and referenced sites within Stanbol to suggest new query keywords to the user.

The audience learns about latest semantic technologies and more importantly be aware of their implementations. Demonstrating the capabilities of Stanbol will create such as realization that joining to the Apache Stanbol community and contributing to the implementation of latest semantic advancements is a good chance for the audience.

IKS project has an *Early Adopters* [18] programme. CMS vendors can get involved in this programme and turn their legacy CMS into a SCMS with the help of Apache Stanbol. The Early Adopters Programme provides grants to help developers evaluate and validate their software. The demonstration is a hands-on proof in this respect also; hence a CMS developer realizes the ease of integration with Apache Stanbol.

References

1. Interactive Knowledge Stack (IKS), `http://www.iks-project.eu`
2. Gokce B. Laleci, Gunes Aluc, Asuman Dogac, Anil Sinaci, Ozgur Kilic and Fulya Tuncer A Semantic Backend System to Support Content Management Systems Knowledge-Based Systems Journal, Vol. 23, pp.832-843, Dec. 2010.
3. Apache Stanbol, `http://incubator.apache.org/stanbol/`
4. OSGi Alliance. OSGi Service Platform - Core Service Specification Version 4.3, 2011, `http://www.osgi.org/Release4/HomePage`
5. Apache Felix, `http://felix.apache.org`
6. Fabian Christ, Benjamin Nagel: A Reference Architecture for Semantic Content Management Systems. In M. Nttgens, O. Thomas, B. Weber (eds.): Proceeding of the Enterprise Modelling and Information Systems Architectures Workshop 2011 (EMISA11), Hamburg (Germany). GI, LNI, vol. P-190, pp. 135-148 (2011)
7. Fabian Christ, Ali Anil Sinaciand Suat Gonul. Development of IKS Reference Architecture for Semantic Content Management Systems. Deliverable, 2012, `http://iks-project.googlecode.com/svn/doc/D5.0-Final`
8. Apache Jackrabbit, `http://jackrabbit.apache.org`
9. Sunita Sarawagi. Information Extraction. Foundations and Trends in Databases, pages 261-377,2008.
10. Apache OpenNLP, `http://incubator.apache.org/opennlp/`
11. Apache Clerezza, `http://incubator.apache.org/clerezza/`
12. DBPedia, `http://dbpedia.org/`
13. Apache Solr, `http://lucene.apache.org/solr/`
14. LDPath, `http://code.google.com/p/ldpath/`
15. Linked Media Framework (LMF), `http://kiwi-project.eu/`
16. Content Repository for Java techonology API, Java Specification Request 170, `http://jcp.org/en/jsr/detail?id=170`
17. OASIS Content Management Interoperability Services, `www.oasis-open.org/committees/cmis/`
18. IKS Project Early Adopter Programme, `http://www.iks-project.eu/projects/early-adopter-programme`