

# Semantic Enhancement: The Key to Massive and Heterogeneous Data Pools

Violeta Damjanovic<sup>1</sup>, Thomas Kurz<sup>1</sup>, Rupert Westenthaler<sup>1</sup>, Wernher Behrendt<sup>1</sup>, Andreas Gruber<sup>1</sup>, Sebastian Schaffert<sup>1</sup>

<sup>1</sup>Salzburg Research, Austria

E-mail: {violeta.damjanovic, thomas.kurz, rupert.westenthaler, wernher.behrendt, andreas.gruber, sebastian.schaffert} @salzburgresearch.at

**Abstract** – This paper surveys various semantic enhancement approaches and techniques in order to answer on demands of today's massive and heterogeneous, Web-based Content Management Systems (CMS). Furthermore, this paper transfers results and shares our experiences in field gained through the development of the Interactive Knowledge Stack (IKS) (the EU FP7 IKS project), more specifically - development of its subproject called the Apache Stanbol. Recently, we decided to make a step further and to provide integration of the IKS technology stack with our newly implemented Linked Media Framework (LMF) that stores and retrieves content and metadata for media resources and resource fragments in a unified way. Hence, this paper, discusses the initial integration points between the Apache Stanbol and the LMF, as well as the benefits they could gain from each other.

## 1 Introduction

A modern Content Management System (CMS) replaces in-house developed CMS for intranet sites and integrates firmly within the Web and document-oriented environment. Documents are central to Knowledge Management (KM), but documents created by using semantic enhancement approaches and techniques bring the advantages of semantic search and interoperability. These benefits require long term systems that provide semantic (search) engine optimisation for managing unstructured information, interoperability that lead to better information sharing, access control and collaborative working, automation to help maintain KM, and more. Our motivation in this paper is to discuss various semantic enhancement approaches and techniques in order to enrich today's massive, heterogeneous, Web-based CMS. In addition, our motivation is to share own experiences in field gained through the development of the Apache Stanbol<sup>1</sup> (the FP7 IKS<sup>2</sup> project) and the Linked Media Framework (LMF): Recently, the IKS Stanbol semantic enhancer has been incubated as an Apache project (Nov. 2010). The LMF is a framework that bridges the gap between the document Web and the Semantic Web (Web of Data). It is an open source development to demonstrate extended Linking Open Data (LOD) principles for

storage, linking and retrieval. Our recent plan is to merge the Apache Stanbol and the LMF with the aim to provide better content enrichment of heterogeneous and massive scale of data by analysing textual and media content.

Paper organization: Section 2 surveys various semantic enhancement approaches. Section 3 discusses both the Apache Stanbol and the LMF. Section 4 identifies ways in which they can be integrated. Section 5 provides conclusion remarks.

## 2 Background: Semantic Enhancement Approaches and Techniques

Starting in 1999 as »a dream for the Web [in which computers] become capable of analysing all the data on the Web – the content, links, and transactions between people and computers« [1], the Semantic Web is still a significant technology to deal with the heterogeneous and massive scale of data, and to enhance knowledge representation, acquiring, and utilizing. It is an initiative of the World Wide Web Consortium (W3C) inspired by the vision of its founder, T.B. Lee, of having a more flexible, integrated, automatic and self-adapting Web that provide a richer and more interactive experience for the end-users [2]. The W3C has developed a set of standards and tools to support this vision: for example, two major working groups around these technologies are the RDF (Resource Description Framework)<sup>3</sup> and the OWL (Web Ontology Language)<sup>4</sup>. RDF is a framework for representing information on the Web. OWL is designed for use by applications that need to process the content of information instead of just presenting information to humans. While combined with ontology, semantic data (RDFS and OWL data) can be reasoned about and queried by using logic rules. This section surveys related research in semantic techniques and approaches to enhancement of various content and data pools. In order to bootstrap the Semantic Web by publishing and interconnecting datasets using RDF, the LOD community emerged [3]. The LOD cloud is in its everyday expansion: according to LOD statistics<sup>5</sup>, at the moment of writing this paper it counts 25.2 billion triples. Hence, the following survey equally investigates LOD browsers and search engines.

<sup>1</sup> Apache Stanbol: <http://incubator.apache.org/stanbol/>

<sup>2</sup> IKS project <http://www.iks-project.eu/>

<sup>3</sup> RDF: [http://www.w3.org/2011/rdf-wg/wiki/Main\\_Page](http://www.w3.org/2011/rdf-wg/wiki/Main_Page)

<sup>4</sup> OWL: [http://www.w3.org/2007/OWL/wiki/OWL\\_Working\\_Group](http://www.w3.org/2007/OWL/wiki/OWL_Working_Group)

<sup>5</sup> <http://www4.wiwiw.fu-berlin.de/lodcloud/>

## 2.1 Semantic search and browsing

The five distinct research directions in semantic search include the following [4]: (i) augmenting traditional keyword search with semantic techniques, (ii) basic concept location (e.g. multi-facet search, semantic auto-completion, search behaviour research), (iii) complex constraint queries for creating query patterns as intuitively as possible, (iv) problem solving, and (v) connecting path discovery. The most prominent LOD browsers and search engines are surveyed in [5]: **Sindice**<sup>6</sup> is a scalable index of the Semantic Web. It crawls the Web for RDF documents and Microformat, and indexes resulting URIs, Inverse Functional Properties (IFPs) and keywords. **Sig.ma**<sup>7</sup> is a semantic information mashup enabled by Sindice that works as semantic browser in which user starts from any entity and browses to the resulting page. **Falcons**<sup>8</sup> is a keyword-based search engine for the semantic URIs that provides different query types for object, concept and document search. **SWSE**<sup>9</sup> is a search engine for RDF data. It is similar to fulltext search, but information retrieval capabilities of SWSE are much more powerful. **hakia**<sup>10</sup> is an Internet search engine that uses QDEXing technology, an alternative new infrastructure to indexing that is based on *SemanticRank* algorithm, a solution mix from the disciplines of semantics, fuzzy logic, computational linguistics, and mathematics.

## 2.2 Semantic mediation: merging and mapping

Mediators were first introduced as parts of distributed information systems in the late eighties [6]. Since then, many application areas have adapted mediators as the approach for overcoming heterogeneity of applications and data sources. Therefore, there are several architectural approaches for *ontology-based semantic mediation*, depending on the type and number of ontologies involved and on a reconciliation approach applied (e.g. *merging* or *mapping*). **Merging** unifies two or more ontologies with overlapping parts into a single ontology that includes all information from the sources. **Mapping** builds the mapping statements that define relationships between concepts of ontologies and rules that specify transformations between two ontologies. An *any-to-any merging model* is shown in [7][8]. OntoMerge [8] introduces bridging axioms between two ontologies, which are not required to be based on the same generalized terminology. The Artemis project [9] shows *any-to-any mapping model* by devising mapping rules among two ontologies. It introduced the OWLmt tool with the purpose to establish the mappings between OWL ontologies and rules execution. The Harmonize project [10] defined a reference Harmonization Ontology and RDF interchange format that is based on the MAFRA tool [11].

<sup>6</sup> Sindice: <http://sindice.com/main/about>

<sup>7</sup> Sig.ma: <http://sig.ma/>

<sup>8</sup> Falcons: <http://www.w3.org/2001/sw/wiki/Falcons>

<sup>9</sup> SWSE: <http://swse.org/>

<sup>10</sup> hakia: <http://www.hakia.com/>

## 2.3 Semantic annotation

Semantic (Web) annotation goes beyond familiar textual annotations of the documents [12]. Semantic annotation formally identifies concepts and relations between concepts in documents, and is intended for use by machines. At the moment of writing this paper, there are two general frameworks for semantic annotation: the W3C Annotea [13] and the CREAM [14]. The Annotea, with its emphasis on collaboration, has influenced the development of a number of systems with good user interfaces that are well suited to distribute knowledge sharing. The CREAM, with its greater emphasis on the deep Web and the annotation of legacy resources, has pushed the development of annotation systems more aimed towards corporate KM.

The review of existing annotation systems given in [12] indicates that research is very active and there are many systems which provide some of the requirements, but that fully integrated environments are still some way off. Technical challenges to development include supporting multi-media document formats, addressing issues of trust, provenance and access rights, resolving the problems of storage.

## 2.4 Semantic analytics and knowledge discovery

Semantic analytics is a process of analysing, searching, and presenting information by using explicit semantic relationships between known entities [15]. The current generation of RDF query languages (e.g. RQL<sup>11</sup>, SquishQL, RDQL<sup>12</sup>) do not support path variables as first class entities and cannot be used for querying for path relationships [16]. In addition, most RDF query systems do not provide adequate querying paradigms to support querying for complex relationships that is highly desirable in many domains. Recently, processing of queries on LOD has gained significant attention. Here, we differ between federated and centralized approaches to processing of queries on LOD. Ladwig and Tran in [17] provided a systematic overview of LOD query processing techniques. Hartig et al. in [18] proposed a method for dealing with the dynamic aspects of LOD query processing that is called *link traversal-based query execution*. An important characteristic that distinguishes this approach from others, such as e.g. *query federation*, is retrieval and query-local processing of LOD from the Web. Since LOD are graph data, we cannot directly reuse existing approaches from relational databases or XML technologies. The existing approaches to integrated or federated query processing over LOD are still in their infancy, and there is not one best solution that can be applied in all cases. Thus, for specific requirements, one need to decide which criteria and constraints are most important and which solution is best suited [19].

<sup>11</sup> RQL: <http://139.91.183.30:9090/RDF/RQL/>

<sup>12</sup> RDQL: <http://www.w3.org/Submission/RDQL/>

### 3 Experiences: Semantic Enhancement via Apache Stanbol and LMF

The objective of the IKS project is to bring semantic capabilities to current CMSs. Thus, IKS puts forward so-called “Semantic CMS Technology Stack” which merges the advances in Semantic Web infrastructure and services with real-world CMS industry needs. The IKS Reference Implementation (RI) (IKS-RI) answers on demands for storing and processing semantic metadata, supporting knowledge extraction, representation and discovery. The IKS-RI is shown in Figure 1 [20]. It is designed in such a way that any existing CMS with the CMS Server architecture can be extended to become semantic CMS.

The major part of the IKS-RI is implemented as its subproject called Apache Stanbol that is built upon the OSGi component model. Apache Stanbol defines a sub-framework called *Stanbol Enhancer* to implement so called the *IKS Knowledge Extraction Pipelines*. The most important knowledge that can be extracted with Apache Stanbol is *identification of certain types of entities (e.g. persons and locations) within the content*. The extracted knowledge is stored via *Stanbol ContentHub* that allows for its retrieval for a given content. *Stanbol EntityHub* is used to retrieve semantic information about entities that are available via accessible LOD sources. *Stanbol EntityHub* also cache the information in an Apache Solr database and allows for them to be published within Apache Stanbol.

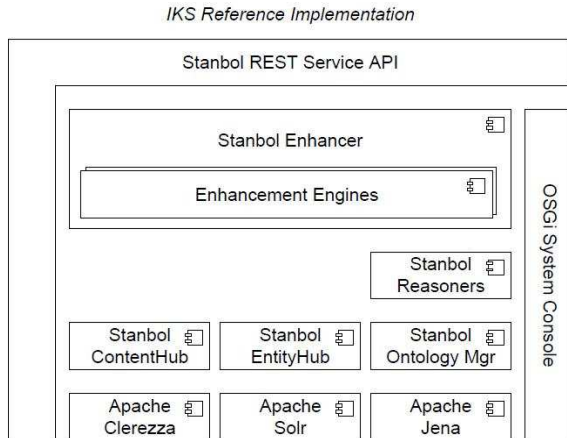


Figure 1. The IKS-RI (showing the three layers of a semantic CMS: (ii) *Semantic Lifting*, (iii) *Knowledge Representation & Reasoning*, (iv) *Persistence*. The first layer (i) *Presentation and Interaction* is out of the scope of this paper)

#### 3.1 The LMF approach

At the same time, we implement the LMF<sup>13</sup> that stores and retrieves content and metadata for media resources and resource fragments in a unified way. LMF implements the LOD extensions proposed in [21], as well as integration of concepts from Linked Media, Media Management, and Enterprise Knowledge Management, in a way that supports semantic annotation of media content, metadata storage, indexing

and search. The LMF conceptual model exposes the following functionalities of the LMF: RESTful Resource Management, Extended Content Negotiation, Semantic Enhancement via LOD, and Semantic Indexing. The LMF is grounded on Service-Oriented Architecture (SOA) that is implemented via CDI/Weld<sup>14</sup>, which is the RI of the Java standard for dependency injection and contextual lifecycle management. A communication between the LMF and the outside world is happened via RESTful web services. The LMF includes beside others the following modules [22] (Figure 2):

**LMF Core:** it implements the LOD Server with the proposed LOD extensions that enable handling resources via the *ResourceWebService*. In addition to triple management, the LMF Core provides transaction persistence management and versioning.

**LMF Search:** it offers semantic search over resources that is based on an Apache Solr database. It includes extendable rule-sets for RDF, Simple Knowledge Organization System (SKOS), Dublin Core (DC) and GeoNames. The LMF REST API is implemented as frontend that can serve to a number of existing commercial systems in a way that they can expose their content and metadata by following LOD principles.

**LMF Sparql:** it provides a SPARQL endpoint for querying the data that are contained within the LMF. SPARQL is standardized query language for RDF graphs. The LMF implements a Sesame-Sails repository. It currently offers only querying but as soon as Sesame allows SPARQL update, the LMF will provide the same.

In addition, the LMF includes an (optional) rule-based reasoner<sup>15</sup> called sKRWL [23] that is highly customizable and allows for user-defined rules to be evaluated over triples in the LMF Triple Store. The rules can be uploaded and stored in the LMF via an easy-to-use web service. The evaluation strategy is an incremental forward-chaining reasoning with truth maintenance and is reasonably efficient even for big data sets. The truth maintenance can be used to provide explanations for inferred triples as well as for efficient updating of the reasoning information.

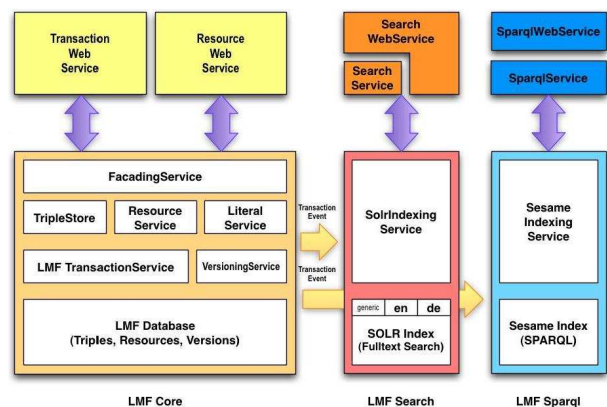


Figure 2. The LMF Basic Services

<sup>13</sup> LMF source code: <http://code.google.com/p/kiwi>

<sup>14</sup> CDI/Weld: <http://seamframework.org/Weld>

<sup>15</sup> LMF Reasoner: <http://code.google.com/p/kiwi/wiki/Reasoning>

## 4 Further Integration

The LMF allows for extending its core system with additional services and functions. One of such extensions is the *LMF Enhancer* component that analyses unstructured content and searches for additional named entities (e.g. persons, locations, organizations) within textual and media content, and links them with external information sources. It provides an interlinking pipeline that is based on the Apache UIMA<sup>16</sup> (Unstructured Information Management Architecture), which allows the usage of powerful but complex and strictly licenced analysis engines based on Natural Language Processing (NLP). One opportunity to integrate the LMF and the Apache Stanbol can be seen as a replacement of the current *LMF Enhancer* by the *Stanbol Enhancer*. There are also attempts to merge these two frameworks. Based on the fact that the Apache Stanbol and the LMF currently cover complementary areas, one can clearly predict the benefits of combining these two projects. For example, the LMF offers (i) an almost ready implementation to the *Stanbol ContentHub*, (ii) rule-based reasoner that allows to process Datalog-style rules over RDF triples, (iii) full LOD capabilities (server as well as client), (iv) semantic search, (v) Linked Data Caching mechanism, which is based on the existing Web Caching approaches. Finally, as it is planned for the *Stanbol Enhancer* to be extended to multimedia interlinking, a bundling of these two approaches will ensure better results. The technical discussions on their integration can be tracked on the Stanbol mailing list<sup>17</sup>.

## 5 Conclusion

Recently, the CMS community has been showing increasing interest in the adoption of Semantic Web technologies to serve real-world CMS needs. In addition to well-known examples such as Drupal<sup>18</sup> and Fedora Commons<sup>19</sup>, the IKS project has been gathering an expanding open source community for semantic CMS. Currently, IKS has two noteworthy subprojects: (i) Apache Stanbol (incubating project), and (ii) Vienna IKS Editables (VIE)<sup>20</sup> dealing with the issue of interaction with semantic content. We are working on integrating Stanbol, LMF and VIE to form a technology stack for semantic CMS that can also deal with LOD.

## Acknowledgement

The research presented here is based upon work supported by both the IKS project (FP7 231527) and the OP4L project (SEEERANETPLUS-115).

## References

- [1] T.B. Lee & M. Fischetti: Weaving the Web. HarperSan Francisco. ISBN 978-0-06-251587-2.

<sup>16</sup> Apache UIMA: <http://uima.apache.org/>

<sup>17</sup> Stanbol mailing list: <http://incubator.apache.org/stanbol/>

<sup>18</sup> Drupal: <http://drupal.org/>

<sup>19</sup> Fedora-commons: <http://fedora-commons.org/>

<sup>20</sup> VIE: <http://www.iks-project.eu/projects/vienna-iks-editables>

- [2] T.B. Lee, J. Hendler, O. Lassila: The Semantic Web. Scientific American. 1999
- [3] C. Bizer et al.: Linked Data-The Story So Far. Int. Journal on Semantic Web and IS (IJSWIS), Vol.5(3), 2009
- [4] E. Mäkelä: Survey of Semantic Search Research, Proc. of the Seminar on Knowledge Management on the Semantic Web, Department of CS, University of Helsinki, 2005.
- [5] T. Kurz, T. Burger, and R. Sting: R3- A Related Resource Recommender. Proc. of the APRESW 2010, 2010.
- [6] M. Vujasinovic et al.: A Semantic-Mediation Architecture for Interoperable Supply-Chain Applications. Int. Journal of Computer Integrated Manufacturing, Vol. 22(6), 2008.
- [7] N. Anicic et al.: An Architecture for Semantic Enterprise Application Integration Standards, Proc. Of Inter. Conf. Interoperability of Enterprise Software and Applications, Springer-Verlag, 25-34, 2006.
- [8] OntoMerge: Ontology Translation by Merging Ontologies.
- [9] V. Bicer et al.: Artemis Message Exchange Framework: Semantic Interoperability of Exchanged Messages in the Healthcare Domain, SIGMOD Record 34(3), 71-76, 2005
- [10] M. dell'Erbaa et al.: HARMONISE: A Solution for Data Interoperability, Proc. of IFIP I3E 2002, 114-127, 2002
- [11] A. Maedche et al.: MAFRA— A Mapping FRamework for Distributed Ontologies, Proc. of EKAW 2002, LNCS 2473, Springer- Verlag, 235–250, 2002
- [12] V. Uren et al.: Semantic annotation for knowledge management: requirements and a survey of the state of the art. *Journal of Web Semantics*, 4(1), pp. 14–28, 2006.
- [13] J. Kahan et al.: Annotea: an open RDF infrastructure for shared web annotations. Proc. of the 10th Inter. World Wide Web Conference (WWW 2001), Hong Kong, 2001.
- [14] S. Handschuh, S. Staab, R. Studer: Leveraging metadata creation for the Semantic Web with CREAM. Proc. of the Annual German Conf. on AI, 2003
- [15] M. Perry et al.: Geospatial and Temporal Semantic Analytics. In *Encyclopaedia of Geoinformatics*, H. Karimia (Ed.), Idea Group, 2008
- [16] RDF Query Survey. Online available at: <http://bit.ly/coL5qw> (accessed August 2011)
- [17] G. Ladwig, T. Tran: Linked Data Query Processing Strategies. In Proc. of ISWC2010, 453-469, 2010
- [18] A. Hartig, C. Bizer, J. Freytag: Executing SPARQL Queries Over the Web of LOD. Proc. of the ISWC 2009
- [19] P. Haase, T. Mathäss, M. Ziller: An Evaluation of Approaches to Federated Query Processing over Linked Data. Proc. of the I-SEMANTICS 2010, Austria, 2010
- [20] F. Christ, B. Nagel: A Reference Architecture for Semantic Content Management Systems. Proc. of EMISA 2011, Sept 2011 (to be published in October 2011)
- [21] T. Kurz, S. Schaffert, T. Bürger: *LMF – A Framework for Linked Media*. Workshop on Multimedia on the Web (MMWeb 2011) at the iSemantics 2011 Conf., Austria. (to be published in September 2011)
- [22] S. Schaffert, T. Kurz: Linked Media: Extending Linked Data for Updates and Arbitrary Media Formats Using the REST Principles. (submitted to the ISWC2011)
- [23] J. Kotowski, F. Bry: A Perfect Match for Reasoning, Explanation, and Reason Maintenance: OWL 2 RL and Semantic Wikis. Proc. of the 5<sup>th</sup> SemWiki2010, 2010.