Unleash the power of Semantic Web in the Enterprise

Georg Schmidt
brox IT-Solutions GmbH
An der Breiten Wiese 9
30625 Hannover
+49 (5 11) 33 65 28 - 77
g.schmidt@brox.de

Igor Novakovic
empolis GmbH
An der Autobahn
33311 Gütersloh
+49 (5241) 80 - 7 43 86
Igor.Novakovic@empolis.com

ABSTRACT

An important requirement to the enterprise IT is the ability to manage information with high flexibility. Semantic web research and resulting technologies are therefore getting more and more vital within business processes.

One question is how to get the research work - done at universities or within corporations - into the enterprise easily. One possible answer to this question is the availability of an open source information processing framework, which meets the requirements of an enterprise. This framework should be mature and flexible enough to design any application.

To move towards such a flexible architecture, which is able to process vast amounts of information in an enterprise, a joint development by BROX [1] and Empolis [2], has been started on Eclipse [3].

Project overview

D.3.3 [**C++, Java**]: Information processing architecture.

General Terms

Management, Measurement, Documentation, Performance, Design, Economics, Reliability, Security, Standardization,

Keywords

Eclipse, Enterprise, Framework, Information Annotation, Information Extraction, Management, Open Source, Operations, OSGi, OSS, Scalability , Semantic Web, SMILA

1. INTRODUCTION

Business processes in the enterprise are getting more and more information driven, meaning that organizations are spending more IT budget for creating information and making them accessible to employees. Examples for the evolution in these areas can be observed in centralized purchasing organizations or research portals.

To get the most benefit out of unstructured information, additional metadata (like in the semantic web) must be created to make information vigorously usable in business processes. Information from Enterprise Resource Planning (ERP) systems has to be annotated, for example to offers that have been created by suppliers. To support processes like resourcing machine readable information in CAD drawings has to be annotated (e.g. for parts). Offers and CAD drawings are often kept in unstructured data sinks.

All this information is strongly distributed throughout an enterprise. Central purchasing organization may have more than 30 IT systems containing structured and unstructured information. Information in these systems is usually - contrary to hyper linking in the internet - strongly disconnected and has no relationship to other information, so called information silos.

To create relationships between this information or to annotate metadata a processing infrastructure is required. This processing infrastructure is currently developed as open source software on Eclipse [4].

The target of the joint development is to generate an infrastructure that allows information driven application to be easily developed in the enterprise. This includes the flexible possibility to make research of universities available in enterprise organizations, keeping quality and maturity required by an enterprise in mind.

2. COMPARISON

Information processing infrastructures are a must have for every search or information management vendor. Additionally, there are initiatives existing on the open source sector, which are aiming into the same direction.

These information processing structures are usually strongly vendor dependent creating extraordinary maintenance costs in the enterprise, because e.g. connectors to Documentum are re-implemented for every information processing technology, having different configuration options and chances for errors while delivering similar features.

Product development often focuses on technology related features e.g. search features while core maintenance and operation requirements are usually not that focused in the product implementations (e.g. encryption, management tools …). The origin of this gap is the missing connection between technology vendors and operation and maintenance departments.

Commercial vendors of such technology are e.g. Autonomy [7], Empolis [2], Google [8], Microsoft (FAST) [9] … having each an

own processing infrastructure for unstructured information management. Even those well known companies sometimes do not meet all operational requirements that have been set by operation and management departments.

Open source projects such as Solr [10] or Open Pipeline [11] are also preparing such an environment, but usually not on enterprise maturity levels.

3. PROJECT GOALS

The SMILA project [4] is trying to create a flexible, basic infrastructure for information management, which is easy to maintain in an enterprise environment. This infrastructure will be flexible in terms of deployment and component usage, e.g. by allowing for different deployment scenarios or just by providing for the availability of subcomponents, which can be used by any application.

Our development is focused on the requirements of large enterprises, including operation and management. The framework will be able to process vast amounts of data in a distributed environment.

This information architecture will allow information annotation or semantic web use cases. Data being generated by an application can be used by "intelligent" web services or agents within an enterprise.

The project is not developing information extraction components, but it will probably create components, that will integrate existing technologies like GATE [13] or LingPipe [14].

The core goal is to make features and components of a semantic web available to the enterprise, which allows the usage of the results of scientific research in an enterprise environment.

Further research will be made easier by making a full featured information annotation infrastructure available. This infrastructure allows for focusing on the research. As additional advantage, the results of research projects may easier be used in the open market, because when creating a spin-off, the need for creating an own information processing framework no longer exists.

Since the beginning of 2008 a group of 13 developers is working on the project. The Eclipse project proposal has lead to the creation review [5] which resulted in the creation of SMILA as official Eclipse project. We soon will open up our work to the open source community under the Eclipse Public Licence (EPL)

4. ADVANTAGES

A common open source infrastructure will lower the market entry barrier for technology companies or university spin-offs.

This infrastructure will allow universities or information management vendors to reduce their maintenance cost for developing a "commodity" information management infrastructure that is able to process vast amounts of information.

The infrastructure will be open source software and therefore will also be available to universities. University spin-offs therefore will have a much lower entry barrier because they don´t have to invest in building a "commodity" infrastructure or build up the know-how to develop it.

By using component technologies like OSGi [12], a flexible application architecture is created while allowing the easy re-usage of components.

5. CONCLUSION

The availability of a common open source infrastructure for information management will decrease the barrier to create semantic web enabled applications. This barrier reduction is achieved by adding several benefits, like scalability, component models, a maturity level of an enterprise application and many more features to a common framework, which is available to companies and the research community.

The community at large (e.g. researchers or companies) can easily make their technology available or focus on research while using an environment, which is able to address high volume issues. The processing of information will therefore be made much easier and results can be tested on real world data or scenarios.

Generally available extensions, such as crawlers, will help the community at large in validation of their technology.

6. ACKNOWLEDGMENTS

Our thanks go to ACM SIGCHI for allowing us to modify templates they had developed.

7. REFERENCES

[1] brox IT-Solutions GmbH, An der Breiten Wiese 9, 30625 Hannover, DOI= http://www.brox.de/

[2] Empolis GmbH, An der Autobahn, 33311 Gütersloh, DOI= http://www.empolis.com/

[3] Eclipse Foundation Inc., 102 Centrepointe Drive, Ottawa, Ontario, Canada, K2G 6B1, DOI= http://www.eclipse.org/

[4] SMILA proposal on the internet. DOI= http://www.eclipse.org/proposals/eilf/ .

[5] SMILA creation review on the internet. DOI= http://www.eclipse.org/proposals/eilf/SMILA_Creation_Review.pdf

[6] Temporary project communication WIKI until complete project creation on eclipse.org.user: eilf-gast, password:jh08hh15ar DOI= http://bugs.brox.de/confluence/display/ECS/Home

[7] Autonomy, Inc., One Market Plaza, Spear Tower, Suite 1900. San Francisco, CA. 94105, USA, DOI= http://www.autonomy.com/

[8] Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA, DOI = http://www.google.com/enterprise/

[9] Microsoft DOI= http://www.microsoft.com/enterprisesearch/

[10] Solr, DOI= http://lucene.apache.org/solr/

[11] Open Pipeline, DOI= http://openpipeline.org/

[12] OSGi – Open Service Gateway initiative DOI= http://www.osgi.org/, DOI= http://en.wikipedia.org/wiki/OSGi

[13] GATE – General Architecture for Text Engineering DOI= http://www.gate.ac.uk/

[14] LingPipe DOI= http://alias-i.com/lingpipe/

Unleash the power of Semantic Web in the Enterprise (Live Demo)

ABSTRACT

A live demonstration will show the SMILA project in action. The presentation contains an overview about SMILA´s architecture, a description of the design and its goals, a live demonstration containing an exemplary indexing and search process, where information is being annotated during the indexing process and a discussion about extensibility of the framework.

1. INTRODUCTION

The SMILA project is an open source information processing infrastructure designed for the enterprise. SMILA allows annotation and extraction of information and is easily embeddable into custom applications.

2. ARCHITECTURE

(Power Point/Discussion)

Architecture and components of SMILA are explained. Furthermore there will be a discussion about scalability and software installation options within an enterprise.

3. DESIGN GOALS

(Discussion)

Several design decisions have been made over time within the SMILA project. Some exemplary decisions (e.g. OSGi, storage, how to build a distribution etc.) are selected and explained.

4. SAMPLE USAGE

(Live Demonstration)

A downloadable sample implementation of SMILA shows information extraction and annotation in combination with Lucene as search engine.

4.1 Data Extraction

Sample crawlers for data extraction on file system or web sites are shown in conjunction with the advanced incremental indexing option.

A brief overview about possible extension points for crawlers and agents is given.

4.2 Information Annotation

Information annotation is demonstrated using self-implemented pipelets within a BPEL designer. These annotations are added to the respective information, which is being processed.

4.3 Indexing

The definition of a Lucene index is demonstrated with various indexing options.

4.4 Operation and Management

Operation and management is demonstrated via the java JMX console. Operational features that are needed within an enterprise (e.g. backup, monitoring …) are explained.

4.5 Search

A fully functional Lucene index will be created during the demonstration. Various search options are being demonstrated in this part of the demonstrations.

5. EXTENSIBILITY

(Discussion)

SMILA´s extensibility model is demonstrated including links to documentation and how-to's. Exemplified extension options, like crawlers and pipelets are demonstrated in detail.

The way of extending the architecture using own components is discussed.