Real world reasons to consider a new OSS Platform for Information Logistics

SeMantic Information Logistics Architecture (SMILAs)

9th European Conference on Case-Based Reasoning
September 1st – 4th, 2008, Trier, Germany
Overview

- Background
- Decision for Eclipse & CSD
- Requirements for a state-of-the-art information annotation infrastructure
- Framework architecture
- Large enterprises as a technology consumer
- Use cases
- Benefits of a framework from different perspectives
Background / Status quo

- Management of unstructured information is critical
- Information distribution within the enterprise is complex
- Use cases are driven by meta data
- A standardized common framework for unstructured information processing is not available
- Problems with proprietary solutions
Typical information distribution scenario

> 6 terabyte of structured information

> 24 terabyte of business critical unstructured information
Sample use cases in the Web

Applications are based on strongly structured data

Some meta data must be created from unstructured information

Sample use cases in an enterprise
- Purchasing
- Engineering
- Portals
- Service and maintenance portals
- ...
Background – Disadvantages of proprietary solutions

- Hard to implement, to maintain and to extend
- Reinventing the wheel all the time
- Slow innovation
- Long development cycles
- Support of standards?
- Flexibility?
Decision for Eclipse & Consortium based Software Development (CSD)

> Decision to launch at Eclipse
  - Global developer community in place
  - Proven track record to serve as a platform for commercial and non-commercial software
  - (share cost of infrastructure and monetize personal investment)

> Why work in a consortium, why go open source?
  - Reduce investment risk for all
  - Mount a credible initiative that can become a standard
  - Build inroads for the rollout of your semantic applications
SMILA Mission and Goal

> Mission

To create a common data logistics infrastructure for next generation semantic information management systems.

> Goal

To create concepts/key components and sample implementations of the information logistics framework.
What are we doing at SMILA?

> Build a common information logistics infrastructure to serve as a platform for key technologies:

- Text and data mining
- Information modeling (Ontologies, Taxonomies, Topic Mapping etc.)
- Visualization / Navigation
- Document translation
- Concept extraction
- Case based reasoning
- Metadata management
- Security and encryption
- Data compression and scalability
Who is involved already? Where do we stand?

- Project initiation by Empolis and brox  
  January 2008
- Eclipse incubation (12 FT developers)  
  June 2008
- DFKI joining Eclipse to work on SMILA  
  July 2008
- Successful SMILA presentation at SAP  
  July 2008
- Official Theseus platform  
  July 2008
- First SMILA release (indexing process)  
  August 2008
- Presentation to CBR Community  
  September 2008
- Platform for next Computer Cooking Contest...
Requirements for an up to date information annotation infrastructure

- Support of well known standards
- Building blocks with a sound component model (OSGi)
- Support for central application management
  - Management tools
  - Configuration management
  - Update handling and management
- Security
- Cross language capabilities (e.g. Java to C++)
- Availability of different distributions for different use cases
  - Grid
  - Cluster
Requirements for an up to date information annotation infrastructure

- Deployment flexibility
  - Ease of deployment
  - Deployment on cheap hardware
- No information should be lost
- Scalability
- Robustness
- Community and partner friendly
- Well documented
- Availability of support
- Enterprise level maturity
Architecture overview

Key Ideas
• Crawlers/Agents push data into Connectivity / Entry Point
• Connectivity Module filters, converts versions, extracts binaries etc. and pushes into queue
• Message-driven queue stores data and guarantees delivery
• 1...n servers respond to messages, process data and write back to queue
• Potentially multiple instances of servers for load balancing and increased throughput
  • Open issue: synchronization of persistence
• 1...n processes inside server arranged via BPEL (~pipelines, ~strategies)
• Search yet to be defined separately but the objective is to separate the processes of (a) filling the index and (b) using the index for search (unlike in e:IAS)
Core Technologies

- OSGi/SCA as component model
- Message queue
- BPEL
- XML
- Storage (XML, distributed file system, ...)
- Search technologies as well as KM technologies
  - Samples: Lucene, IBM, Fast, Google, ...
  - Information extraction
    - GATE
    - Document Converter (e.g. Apache POI, Stellent, ...)
  - Extreme diversity of technology companies (> 2000)
How to extend/use SMILA

Pipelets – Components in workflow system that modify/annotate information
- Well documented
- Samples (XML transformation, document converters, indexing, ...)
- Visual designer could soon be used for orchestration
- Possible implementations (UIMA, GATE, “your components”, ...)

Crawlers – Components for extracting information from data sources
- Well documented
- Samples (Database integration, Documentum, web crawler, Sharepoint, ...)

Easy integration in your software due to used component model
- Please contact the community for support
- Process not yet documented
Large enterprises / corporations (applications and application maintenance)

> Applications
  - Responsibilities (e.g. Operation, Development, ...)
  - Application owners (departments / specialist divisions)
  - Optimal application functionality
  - Implementation
  - Technology know-how
  - Sustainability (standardization departments...)

> Application maintenance
  - Simple maintenance
  - Learning curve
Large enterprises / corporations (status quo)

- Different technologies (e.g. > 70 search technologies within large enterprise corporations)

- Costs
  - Implementation costs
  - Up to five times higher maintenance costs

- Investment protection
  - Standardization departments

- Limited communication
  - Application and network zones
  - Firewalls
  - Protocols
Large Enterprises / Corporations
(Introduction of a new Technology)

> Technology penetration of a large corporation using a search technology as a sample
  - 5% accomplished → New technology strategy
  - Are investments lost?

> Learning curve
  - How could employees be educated?
  - How could knowledge be transported to the new technology?

> How could this issue be eased?
  - Standardization → Framework
  - Technology vendors
  - Large enterprise corporation
  - We should think of the current state as a „Pre-JDBC/ODBC“ era!
Use cases

Features

- Flexible application scenarios
- One framework different deployments

Applications

- Search
- Doublet cleansing / avoidance
- Database Offload
- Intelligent Web Services
- Process support
- Creation of meta data
- ...

Information Excellence
Benefits of a framework for researchers

> Focusing on research
  - Creation of an own information processing infrastructure could be avoided
  - Ability to use real world data

> Easy creation of university spin-offs.
  - Availability of a enterprise ready commercial framework
  - Availability of support

> Ability to publish one’s work
  - No legal drawbacks for publishing
  - Ability to create software downloads from projects

> Availability of a community to discuss questions
Benefits of a framework from different perspectives

➢ Customers
  – Infrastructure standardization
  – Additional functionality created by the ecosystem

➢ Technology vendors
  – Get rid of one of the largest cost drivers
  – Additional functionality created by the ecosystem

➢ Research
  – Faster innovation cycles
  – Short Time-to-market
The SMILA project

> Incubation at Eclipse running

> Project overview
  – Currently 12 developers
  – Concepts available at SMILA-Wiki
  – First prototype with horizontal walkthrough

> We are preparing a downloadable version at Eclipse
  – If you are interested please contact our team

> SMILA is a technological base for the SME contest of € 90 Mio funded “Theseus” project

> Commercial support available
Contact

- August Georg Schmidt, brox IT-Solutions GmbH
gschmidt @ brox.de
- Igor Novakovic, empolis GmbH

Resources

- eclipse.org/smila
- wiki.eclipse.org/smila
- Newsgroup: eclipse.rt.smila
eclipse.org/smila