KW Partner: FU Berlin

1 Overview

Challenge
EAI (principally search) in an University setting (low budget, high heterogeneity). A high need for automation is present in this setting.

Solution
The RDF-based integration of IT systems including "homegrown" text-based systems. Dynamic decentralised semantic-enabled search.

Why a Semantic solution
The semantic interfaces to information allow the client greater re-use of existing information, are flexible, allowing easy expansion of the system and lower costs of administration for this reason. Automated support for all integrative processes is vital. The bottom line is that searches using semantic technology are more accurate and deliver better information.

Key Business Benefits
A new product in a high-growth sector.

Other business partners
A consultancy to handle teaching issues
A user partner

Keys components

Existing Software
EAI solutions
Data repositories
Peer to peer networking

Research and development
Autonomous systems/self-organisation
Semantic Grid
Semantic Web Services
Rights and security
Ontology mapping

Technology locks
Artificial Intelligence issues
Semantics in P2P networks/EAI

This business case deals with the provision of a peer-to-peer network for facilitating information exchange across an university setting in the field of science, overcoming the challenges of heterogeneity in enterprise applications and their data.

Universities, as large institutions consisting of various departments, faculties and administrative offices, store their data in a distributed heterogeneous manner yet for typical university-wide activities based on information pooling and sharing they require low cost solutions to a single entry point access to this data.

In this case, we consider the need in the scientific community within and across universities to share their research and publications, making results easily findable to other researchers. We take as a basis the paradigm of P2P (peer-to-peer) networking to enable a decentralised means to share documents across a scalable network of participating nodes, and of EAI (Enterprise Application Integration) to enable communication between heterogeneous, possibly legacy systems which host these documents.

In this context we use metadata as a basis for the efficient and accurate search and indexing of the documents on the network in order to ensure users find quickly and correctly the material they require.

Figure 1 – The e-science portal S2S

2 Current Practices and Technologies

2.1 Current business practises

At the communication level, Enterprise Application Integration (EAI) solutions are commonly used within enterprises to connect heterogeneous IT systems. Web Services are the most current evolution in EAI, using HTTP and XML as a standardized communication protocol and message structure respectively for integrating systems over the Internet. This integration is point-to-point and the messages exchanged are syntactically heterogeneous.

At the data level, data can be pooled in a repository, though this leads to latency problems (access to data in the central repository while it is being changed locally). This data pooling requires hard-coded data translation (from the source to target format) which grows exponentially as both users and providers expand in terms of formats used by their systems and proves to be unscalable.

A possible solution being examined by the eScience portal S2S[1] (see Figure 3.1) is the use of peer-to-peer networking to store documents in a distributed manner and with a standardized client to ensure consistent access to heterogeneous data on heterogeneous systems. It uses a text- and metadata-based search to allow users to locate the documents they seek on the network. The metadata is generated when documents are placed onto the network in a semi-automatic basis and using this metadata new or changed documents are identified each time the client logs into the network and a local search index is updated.

To provide for integration at the communication level, JXTA is used to provide for platform-independent peer-to-peer communication. At the data level, the heterogeneous documents are considered generically as objects on the network and their information for search functionality made accessible through a consistent metadata representation of each document.

2.2 System requirements Analysis

The approach of S2S still comes up against problems of P2P and EAI in terms of data and application heterogeneity. It is difficult to bend EAI solutions onto P2P, so that the peer-to-peer contact occurs only on the basis of a specific client that must be downloaded and installed on each machine. Semantics would be a possible solution to enabling the automatic integration of new peers into the network based on metadata about their communication needs. Likewise self-organisation of the network would be facilitated by a standardized means to express information about the network and its peers, such that changes can be propagated along the peers and each can react correctly to that change.

Likewise, messages between peers must be based on the same structure and syntax so that clients can understand one another. Mappings based on different domain vocabularies would allow peers to communicate which seek different languages. Extraction based on such vocabularies would improve the text-based analysis to provide metadata for the documents. Finally, metadata extraction could be further improved and automated so that data in legacy systems can be introduced to the network without great effort and new documents can be

---

[1] http://s2s.neofonie.de

indexed effectively. Rights and security need also to be taken into account on the network infrastructure.

2.3 Review of the current systems

S2S exists in a research field which contains other activities in P2P and EAI for example:
- Alvis (http://www.alvis.info)
- EduTella (http://edutella.jxta.org)
- NeuroGrid (http://www.neurogrid.net)
- JXTA (http://www.jxta.org)

Unlike such pure research projects, S2S seeks to apply these innovative ideas practically in the implementation of a new type of application, in which information vital to research is effectively made available and an overview of research efforts in a given area is made possible. There are other commercial implementations of P2P search (e.g. www.grub.org; www.jigle.com) that differ from the approach of S2S in that they exhibit some of the following features:
- Search is based on a very narrow metadata set, extracted e.g. through music interpretation or video
- The user is restricted to generating content with a specific tool
- The data must exist locally to the user or the user can not choose which data will be indexed
- Users can not form into (thematically related) groups

S2S on the other hand exhibits these features, which could be then extended and improved through semantics:
- Search can be based on full text or metadata fields
- User is free to generate his data in whatever format he prefers

User can select which data to be shared is stored locally and which is found on accessible FTP or Web servers