A semantic approach to access heterogeneous data sources: the SEWASIE Project*

S. Bergamaschi and M. Vincini

Dipartimento di Ingegneria dell'Informazione

Università degli Studi di Modena e Reggio Emilia

{bergamaschi.sonia, vincini.maurizio}@unimo.it

Intoduction

SEWASIE is implementing an advanced search engine that provides intelligent access to heterogeneous data sources on the web via semantic enrichment. This can be thought of as the basis of structured secure web-based communication. SEWASIE provides users with a search client that has an easy-to-use query interface, and which can extract the required information from the Internet and to show it in a useful and user-friendly format. From an architectural point of view, the prototype will provide a search engine client and indexing servers and ontologies.

There are many benefits to be had from such a system. There will be a reduction of transaction costs by efficient search and communication facilities. Within the business context, the system will support integrated searching and negotiating, which will promote the take-up of key technologies for SMEs and give them a competitive edge.

The Business Scenario

Throughout Europe, much of the industrial fabric is made of small and medium-sized enterprises (SMEs) in fields such as agriculture, manufacturing, commerce and services. For social and historical reasons, these tend to aggregate into sectorial clusters in various parts of respective countries. Today, this kind of economic organization is threatened byglobalisation.

One of the keys to sustainability and success is being able to access information. This could be a cheaper supplier, an innovative working method, a new market, potential clients, partners, sponsors, and so on. Current Internet search tools are inadequate because they not only are they difficult to use, the search results are often of little use with their pages and pages of hits.

Suppose an SME needs to find out about a topic - a product, a supplier, a fashion trend, a standard, etc. Suppose, for example, a search is made for 'fabric dyeing processes' for the purpose of finding out about the disposal of the dyeing waste material. A query to www.google.com for 'fabric dyeing' listed 540 hits at the time of writing, which related not only manufacturers of fabric dyeing equipment, but also the history of dyeing, the dyeing technology, and so on. Eventually a useful contact may be found, and the search can continue for relevant laws and standards concerning

* This paper is founded by IST 2001-34825 Project

waste disposal. But is it *law* or the *interpretation* of the law? What if the laws are of a different country where the practices and terminologies are different?

Users need an easy-to-use interface to the search system, one that is usable by non-expert users and with poor network connections. The main requirement is to get structured results obtained from an interpretation of vague queries followed by some filtering techniques based on designer rules and the acquired experiences. For instance, starting from a request simply constituted by the keyword “punch” the engine should answer with one or more documents containing information in an easy-to-accessible format such as sellers, prices, manufacturers, technical literature on its use, importer and so on. Of particular interest for products and manufactures is to know if there are goods at low price or if there are auctions, or other negotiation mechanisms, for making the purchase.

The SEWASIE Idea

SEWASIE (SEmantic Webs and AgentS in Integrated Economies) will design and implement an advanced search engine that provides access via a machine-processable semantics of data, which can form the basis of structured web-based communication. Tools and methods will be developed to create and maintain multilingual ontologies, with an inference layer grounded in W3C standards (XML, XML Schema, RDF(S)), that are the basis for the advanced search mechanisms; these will provide the terminology for the structured communication exchanges.

Search results will be personalised and visualised according to users’ preferences. The system will be an open and distributed architecture based on intelligent agents (brokers, mediators and wrappers), and will accommodate scalability and flexibility issues such as: the ability to fit in changing and growing environments; to interoperate with other systems while offering one central point of access to the user, etc.

Special Query Agents will support users when querying heterogeneous web information sources. The query is sent to a query agent that moves through the SEWASIE information nodes and retrieves the information requested by the user. Information nodes are independent components that semantically enrich existing data sources by linking the data to ontologies and other metadata. The system will also be capable of real-life business evaluation of the results, striving to develop a system and tools which not only solve the problem, but do so in a usable, marketable way.

Project objectives

SEWASIE has the following specific objectives:

- To develop an agent-based, secure, scalable and distributed system architecture for semantic search (ontology based) and for structured web-based communication (for electronic negotiation).
- To develop a general framework responsible for the implementation of the semantic enrichment processes leading to semantically-enriched virtual data stores that constitute the information nodes accessible by the users. The created ontology must have a multilingual interface, based on a logical layer and coded using widespread W3C standards.

- To develop a general framework for query management and information reconciliation taking into account the semantically enriched data stores. First, commonalities among queries have to be detected, then the relevant virtual data stores responsible for answering parts of the queries determined and the queries accordingly split. Finally, the sub-answers have to be combined in order to provide the user with an overall answer to the original query.
- To develop an information-brokering component that includes methods for collecting, contextualising and visualising semantically-rich data. To obtain these result, intelligent information filtering and knowledge guidance services have to be developed on the basis of semantic web technologies. Structured data has to be linked to semi- or unstructured data via ontologies. The collected data has to be visualised to show related documents and search result contexts for the purpose of financial control.
- To develop structured communication processes that enable the use of ontologies. The communication tool enables structured negotiation support for human negotiators engaging in business-to-business electronic commerce and employing intelligent software agents for some routine communication task.
- To develop end-user interfaces for both the semantic design and the query management. The first is a tool supporting the design, the management, and the storage of the semantic information associated to virtual data stores together with a conceptual modelling methodology associated to the devised data model. The latter is a tool for end-user query management and intelligent navigation exploiting the semantic information associated to virtual data stores and to the global virtual view.

Sewasie Architecture

The SEWASIE project designs and implements an advanced search engine enabling intelligent access to heterogeneous data sources on the web via semantic enrichment to provide the basis for structured web-based communication. In particular, a user should be able to access the SEWASIE system through a central point of access where (s)he are provided with tools for query composition, for personalising search results and other web data, for visualising results, and for communicating with other business partners about search results, e.g. in electronic negotiations.

The SEWASIE system aims to realise a virtual network, SEWASIE Virtual Network (SVN) (shown in Figure) whose nodes are SEWASIE Information Nodes (SINode).

SINodes are mediator-based systems, each including a Virtual Data Store, an Ontology Builder, and a Query Manager. A Virtual Data Store represents a virtual view of the overall information managed within any SINode and consists of the managed information sources, wrappers, and a metadata repository. The managed Information Sources are heterogeneous collections of structured, semi-structured, or unstructured data, e.g. relational databases, XML or HTML documents. A Wrapper implements common communication protocols and translates to and from local access languages. There is one wrapper linked to each information source. According to the metadata provided by the wrappers, the Ontology Builder performs semantic enrichment processes in order to create and maintain the current Ontology which is made up of the global virtual integrated view (in short GVV) of the managed sources and the mapping description between the GVV itself and the integrated sources. Ontologies are built on a logical layer based on existing W3C standard. The Metadata Repository holds the ontology and the knowledge required to establish semantic inter-relationships between the SINode itself and the neighbouring ones. A Query Manager provides the functionalities for solving a query within an SINode and constitutes the SINode interface to the network.

Users send queries to the SEWASIE system about data and metadata via the User Interface that provides one common point of access for all SEWASIE services and tools. The user interface transfers queries to the Query Agents that are intelligent information agents with the specific task of solving a query within the SVN. Starting from a specified SINode, the query agent then accesses other SINodes and thus collects partial answers. Brokering Agents are responsible for maintaining the knowledge related to the SEWASIE network and the user profiles. More precisely, brokering agents classify SINodes on the basis of user profiles. They filter and contextualise Query Agent answers, possibly linked to OLAP reports. They serve also as intelligent filters, which monitor Web sites of competitors or potential collaborators, and they are responsible for handling the acquisition of a new SINode and the consequent update of the SEWASIE network. Moreover, a Brokering Agent manages the semantic inter-relationships between the SINode and the neighbour ones. Neighbourhood can be evaluated by means of a similarity measure between the native node ontology and that of the visited one. Finally, the Communication Tool provides the means for structured web-based communication. It uses query results, contextualised information and ontologies from SINodes as the basis for the communicative content. The Communication Agent performs communication tasks in the early phase of electronic interactions. Given that most of the communication is performed by human communication partners, they interact with the tool via the communication interface. The user profile influences the structure of the communicative interactions.

Acknowledgements

This work is supported in part by the 5th Framework IST programme of the European Community through project SEWASIE within the Semantic Web Action Line. The SEWASIE consortium comprises in addition to the author' organization (Sonia Bergamaschi is the coordinator of the project), the Universities of Aachen RWTH (M. Jarke), Roma La Sapienza (M. Lenzerini, T. Catarci), Bolzano (E. Franconi), as well as IBM Italia, Thinking Networks AG and CNA as user organisation.

Project Name:	**SEWASIE**
Project No:	**IST-2001-34825**
Start Date:	***01/05/2002***
End Date:	***30/04/2005***
Key Action 3 Area:	**Information Access, Filtering, Analysis and Handling**
Project Type:	**Shared-cost RTD**
Project Co-ordinator:	**Università degli Studi di Modena e Reggio Emilia (IT)** Prof. Sonia Bergamaschi
Project Participants:	**CNA SERVIZI Modena s.c.a.r.l. (IT)** **Università degli Studi di Roma “La Sapienza” (IT)**

	Rheinisch Westfaelische Technische Hochschule Aachen (DE) **Libera Università di Bolzano (ITALY)** **Thinking Networks AG (DE)** **IBM Italia SPA (IT)** **Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung eingetragener Verein (DE)**
Project Website:	***www.sewasie.org***