30 September 2002
Siguenza, Spain

EKAW'02 workshop on
Ontologies for Multi-Agent Systems

SEWASIE:
a Semantic Search Engine

Sonia Bergamaschi
Università di Modena e Reggio Emilia

Outline

- What is SEWASIE?
  - Objectives
  - Expected results
  - Main Innovations
  - The high level Architecture
  - Components
  - A SINode
- A P2P paradigm for SEWASIE (for later discussion)
  - INTER SINode Network
  - Brokering Agent Network

SEWASIE

Information Society Technologies

- SEWASIE (Semantic Webs and AgentS in Integrated Economies) is a research project founded by EU on action line Semantic Web (May 2002/April 2005)
  http://www.sewasie.org
- The consortium details
  - Università degli Studi di Modena e Reggio Emilia (ITALY)
  - CNA SERVIZI Modena s.c.a.r.l. (ITALY)
  - Università degli Studi di Roma "La Sapienza" (ITALY)
  - Rheinisch Westfaelische Technische Hochschule Aachen (GERMANY)
  - Libera Università di Bolzano (ITALY)
  - Thinking Networks AG (GERMANY)
  - IBM Italia SPA (ITALY)
  - Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung eingetragener Verein (GERMANY)

Contact information:
Prof. Sonia Bergamaschi
DII – Università di Modena e Reggio Emilia
Tel: +39 059 2056132 Fax: +39 059 2066126
bergamaschi.sonia@unimo.it
http://www.dbgroup.unimo.it/Bergamaschi.html

SEWASIE Objectives

Information Society Technologies

Design and implementation of an advanced search engine

The SEWASIE project pursues the following aims:

- To develop an agent-based secure, scalable and distributed system architecture for semantic search (based on ontologies) and for structured web-based communication.
- To provide semantic enrichment processes for knowledge-based extraction of meta-information of heterogeneous data sources.
- To develop a general framework for query management and information reconciliation based on a semantically enriched data and trusted agent structure.
- To develop an information brokering component which includes methods for collecting, contextualising and visualising semantically rich data.
- To develop communication processes that enable the use of multilingual ontologies.
- To provide the end-user with efficient interfaces for formulating queries using a graphical representation and for intelligent navigation through the semantically enriched information space.

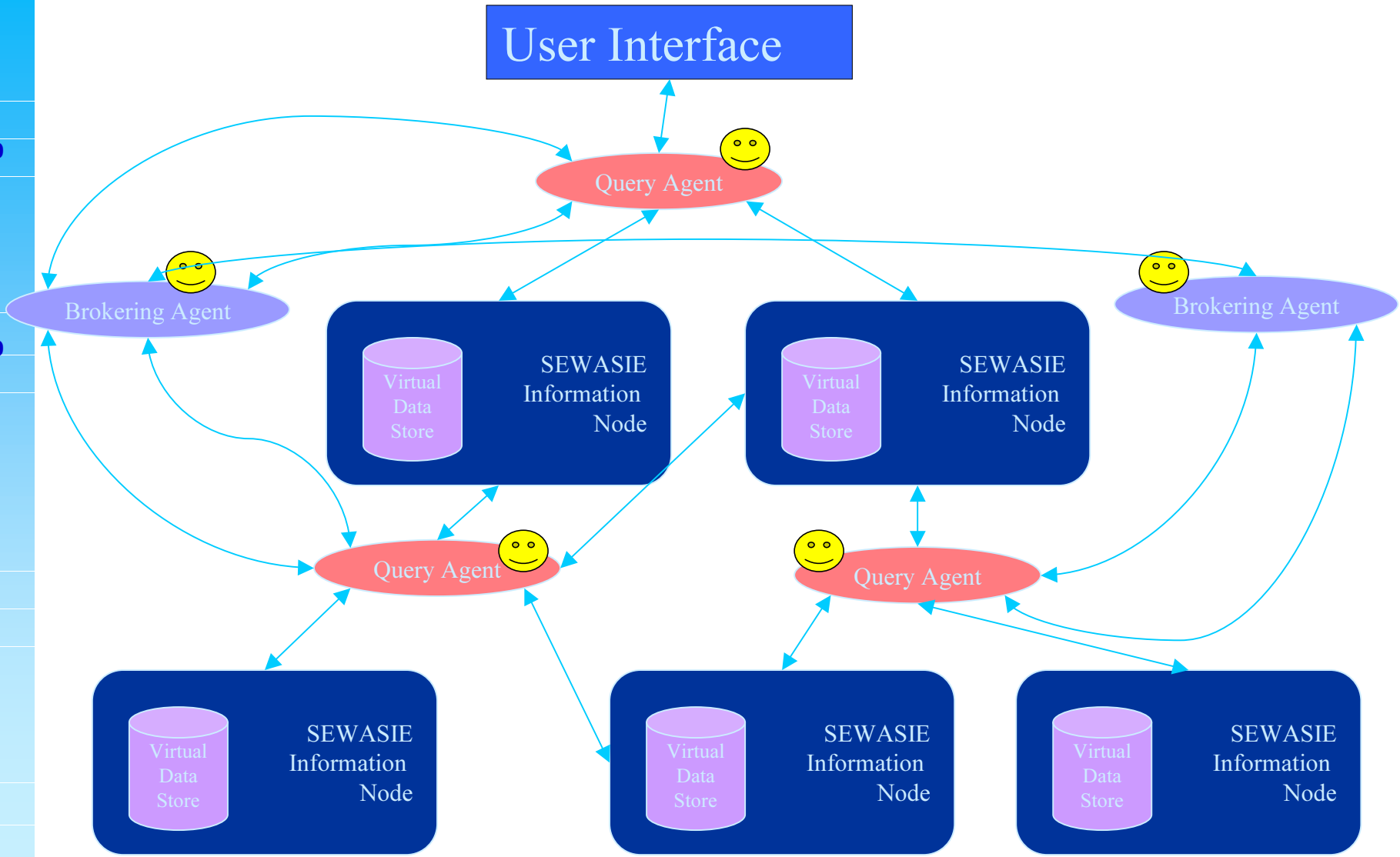
Expected Results

Information Society Technologies

The SEWASIE vision helps European enterprises to compete in a global market and to form strategic alliances at a European level by providing a sophisticated retrieval, brokering and communication service on basis of the semantic web technology.

- In particular, SEWASIE has to:
- Help European SMEs to find the right strategic information at the right time in a multinational environment;
- Provide advanced and novel services for monitoring and linking information in the context of risk management and competitor analysis;
- Provide ontology-based communication mechanisms for negotiation in multi-language environments;
- Ease the use of complex cross-language retrieval and data condensation tools by providing intuitive interfaces.

The very high level architecture



The very high level architecture

- Tools and methods has to be developed to create/maintain multilingual ontologies, with an inference layer grounded in W3C standards (XML, XML Schema, RDF(S)).
- Search results will be personalised and visualised according to users' preferences.
- From an architectural point of view, SEWASIE aims to provide an open and distributed architecture based on intelligent agents (brokers, mediators and wrappers) facing scalability and flexibility issues, i.e. the ability to fit in changing and growing environments and to interoperate with other systems, while offering one central point of access to the user.
- The main actors on stage are
  - The user interface
  - The query agent
  - The brokering agent
  - The information node (SINode)

The very high level architecture

Different Degrees of expansion and integrability of the system

- The case of a **STATIC WORLD**, where one universal ontology with a reference vocabulary is defined beforehand, and all sources have to fit in there somehow; such a set-up may be envisioned for smaller, strongly structured worlds
- The case of a **DYNAMIC WORLD**, where no universal ontology exists except as juxtaposition of all the existing ontological domains identified at any given time; this set-up appears to be typical of larger, open, partially structured, worlds with autonomous components
- Moreover, it should be noted that the integration policy may be different at the SINode level and at the global level. In fact, a more stringent ontology at the SINode and a loose juxtaposition at the global level may be a reasonable starting point.

Main Innovations (1)

Information Society Technologies

The SEWASIE project aims to develop 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.

multilingual ontologies and agents

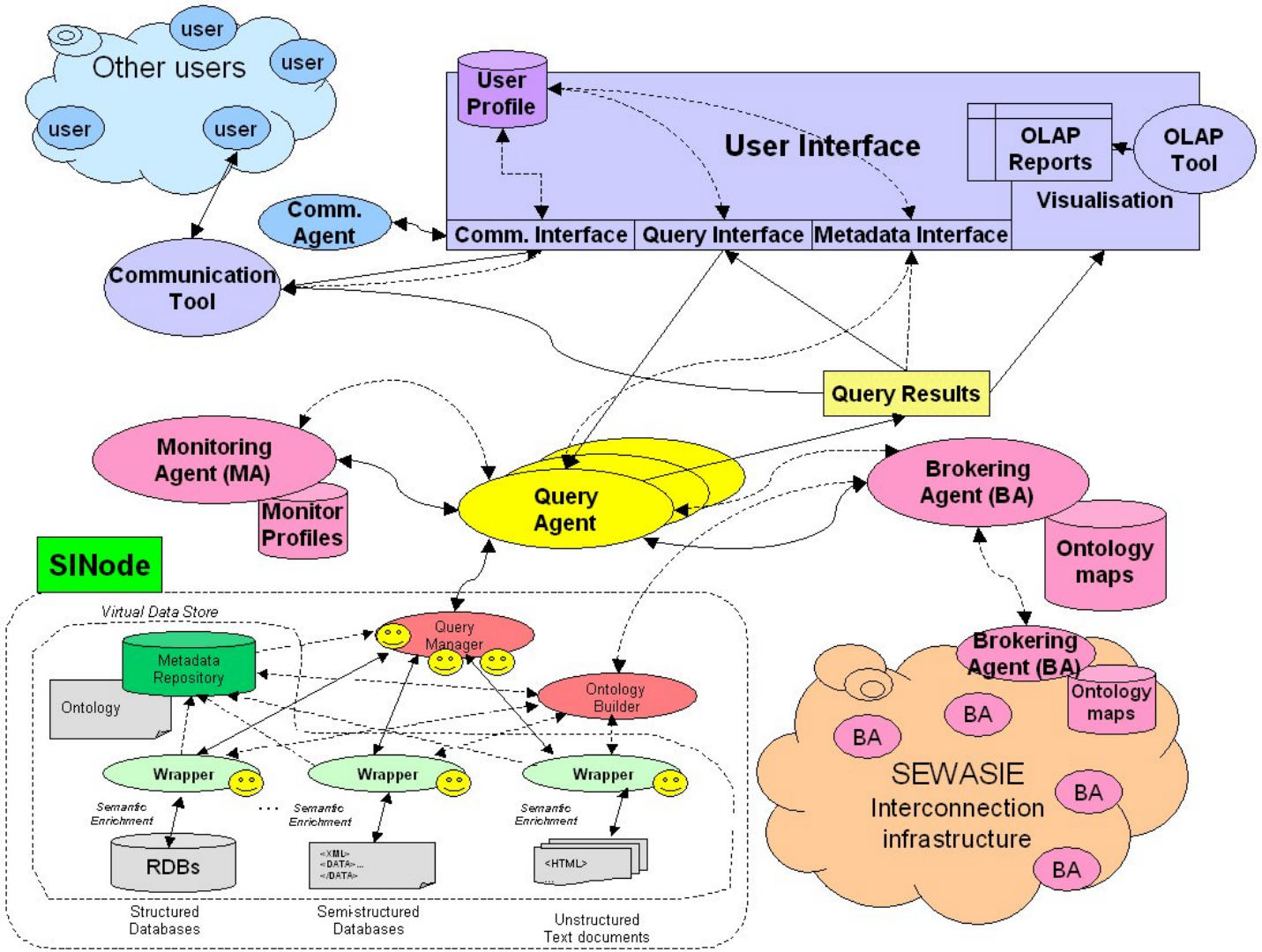
- The SEWASIE system will realise a virtual network, **SEWASIE Virtual Network** (**SVN**) whose nodes are **SEWASIE Information Nodes** (**SINode**).
  - ***SINodes*** are multi-database mediator-based systems, each including a Virtual Data Store, an Ontology Builder, and a Query Manager
  - The managed ***Information Sources*** are heterogeneous collections of structured, semi-structured, or unstructured data, e.g. relational databases, XML or HTML documents

Main Innovations (2)

- Ontologies are multilingual
- The Brokering Agent/Agents maintains the knowledge related to the SEWASIE Virtual Network and the user profiles.
- In the SEWASIE Virtual Network, the Brokering Agent classifies SINodes, it is responsible for handling the acquisition of a new SINode and for consequently updating of the SEWASIE Virtual Network.
- In query solving phase, starting from a specified SINode, the Query Agent accesses other SINodes and thus collects partial answers.
- To select SINodes useful to solve a query, a Query Agent interacts with a Brokering Agent.

Sewasie Architecture (Under development)

Information Society Technologies



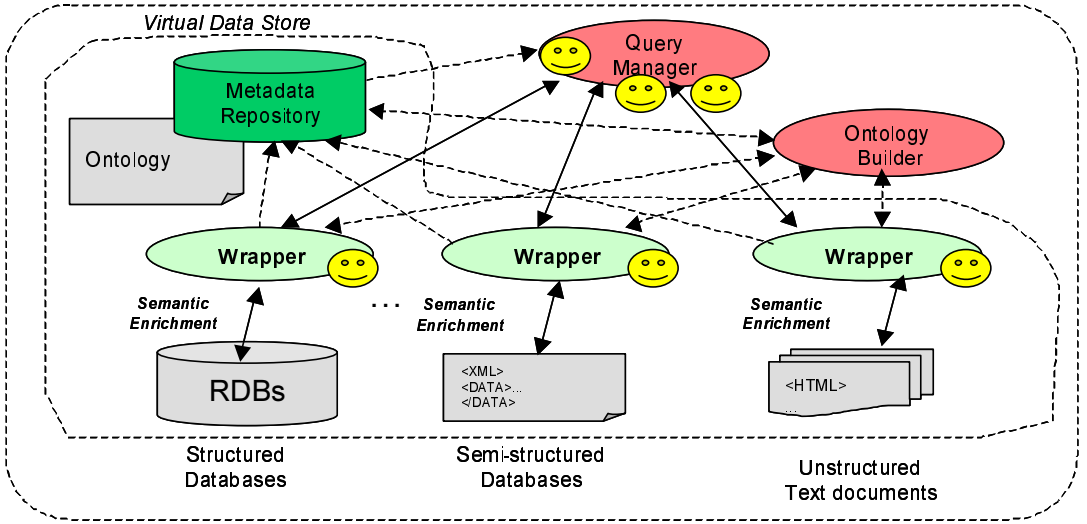
Components (actors and stages)

- The basic (generic) user query scenario we have in mind concerns a user at a workstation (or handling a handheld computer, or a cellular phone with network connection capabilities), looking for information on a topic. The user may then issue a request expressed in some "natural" language style to the network. The user interface translates the user request into a query, keeping into account the past history and present context of the user, and sends a probe out (the query agent, QA) scouting for answers.
- The QA connects into the network of SEWASIE brokering agents (BAs) and queries them for info on the matter of interest. A typical interaction between a QA and a BA may imply that the BA will provide directions to relevant SINodes and information on SINode contents, or reference the QA to other BAs. The QA will then move to such nodes and query them, or may move on to the other BAs to ask them for directions again.
- When the QA receives the SINode answers, it has to integrate them, possibly querying some BA again (data reconciliation).
- Another type of user query is longer-term network monitoring request. While the previous one is a short-term straight request which terminates with the return of answers or the decision that there aren't any, the monitoring request is an open-ended request for information available which rather looks for changes in the content of the network. In this case the QA will monitor a certain predefined view of the domain. The QA will return to the user interface any change that will be detected over time.

The SINode module

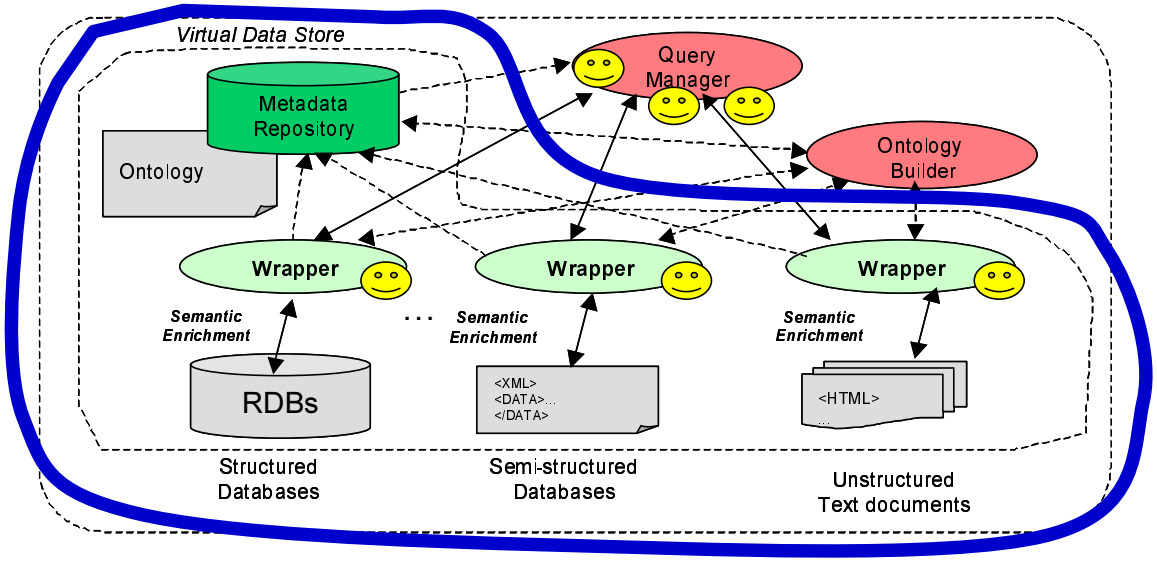
Information Society Technologies

- ***SINodes*** are mediator-based systems, including:
- A ***Virtual Data Store*** (VDS) 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.



- A ***Wrapper*** implements common communication protocols and translates to and from local access languages. There is one wrapper linked to each information source.
- The ***Ontology Builder*** performs semantic enrichment processes in order to create and maintain the current ***Ontology*** which is made up of the Global Virtual View of the sources and the mapping description between the GVV itself and the sources.
- The ***Metadata Repository*** holds the ontology and the knowledge required to establish semantic relationships between the SINode itself and the neighbouring ones.

Virtual Data Store



Global VDS model and language

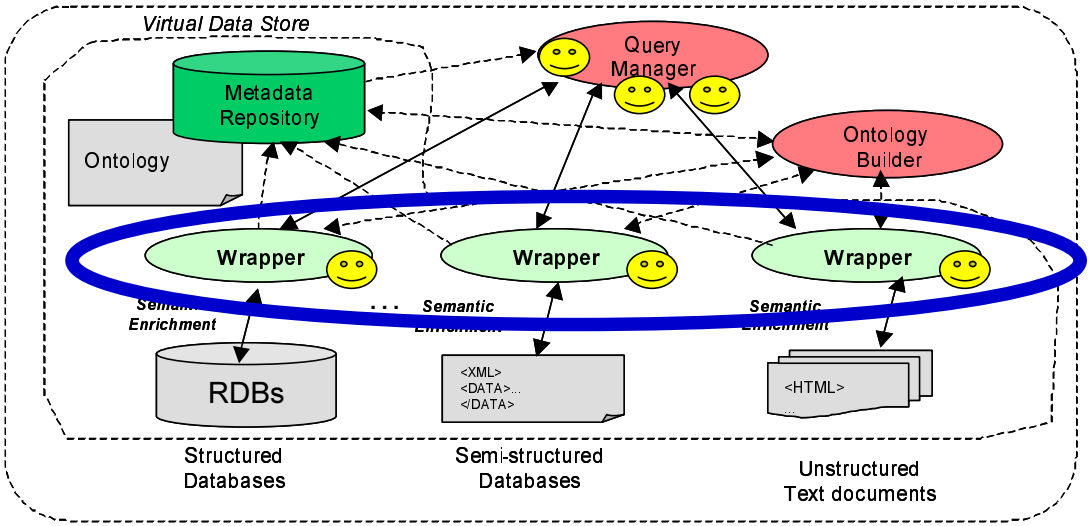
- The first tenet of the architecture within the VDS is a common model and the associated languages, travelling as payload on the global VDS infrastructure.
- The main requirements for a candidate language are
  - a rich syntax for ontology description, including mapping relations GVV/Sources
  - a flexible query language and tools for effective translation of queries and results among modules.
- One candidate for the data model and associated languages is ODMI3 (ODLI3), which was derived from ODMG specification; one candidate for the query language is OQLI3.
- Notice that the adoption of specific languages for intra-node communication does not avoid to put at the SEWASIE network disposal the information managed by SINode in other format.

Virtual Data Store

Global VDS infrastructure

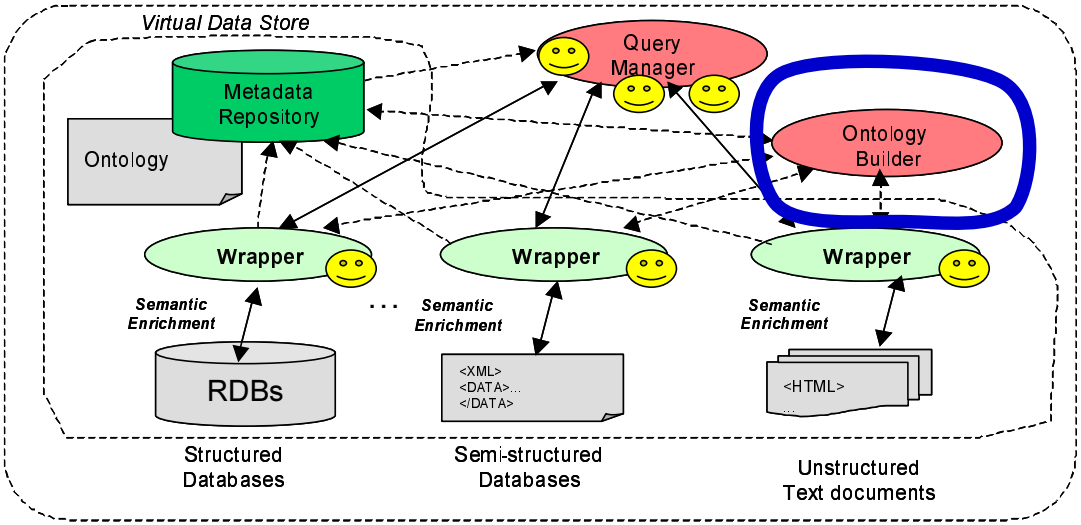
- The architecture of this module is inherently distributed (i.e. in most cases its functionality will be distributed among several host machines of different types). As a consequence, these components will all need to have inter-process communication functionalities to support the interaction. The first choice here is to use the TCP/IP family of protocols, which are universally supported at all levels. Above the basic network layer we need to select a proper enveloping mechanism to guarantee the higher level properties of the communication.
- We want to have
  - verified point-to-point communications
  - no special requirements to pass across common boundaries like firewalls (at least those with typical policy definitions)
  - an option to use reserved (encrypted) communications
  - standardisation and widespread availability (at least in a medium term perspective)
- A natural candidate for such a protocol family is provided by the SOAP/WSDL environment. An alternative to the SOAP/WSDL/UDDI family is given by the CORBA architecture. Based on these protocols the VDS will have an API made available to applications (agents or others) in order to
  - Query the content of the SINode: receive a query and return the corresponding results
  - Manage the semantic profile of the SINode in the SEWASIE network: keeping the up to date semantic profile, and updating it

Wrappers



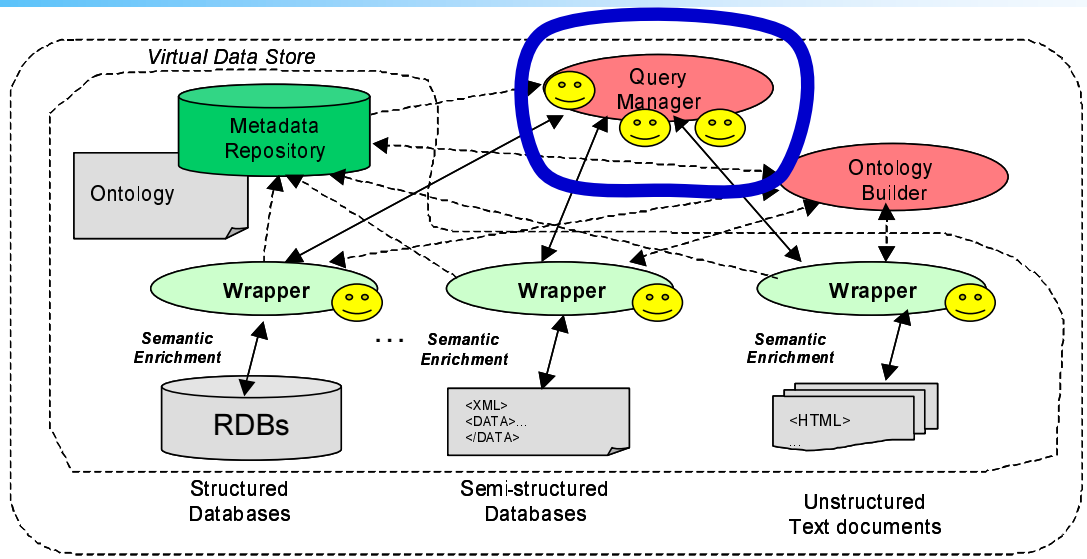
- Wrappers are the "docking stations" of the heterogeneous data sources contributing content to SEWASIE. They are software modules in charge of the mediation between the internals of each data source and the functionalities of the SINode.
- Different wrappers have to be defined to cover structurally diverse sources.
- The internals of the wrapper will need to be modular. However, the interface of these modules will be uniform and independent of the underlying source type.
- Two major functions need to be performed by these wrappers:
  - to support the translation of the structure of the information managed by local sources into the SINode description language
  - to support the translation of the queries from the SINode query language into the specific query language of the underlying source.
- To this aim, functionalities and protocols will need to be made available in order to enable the communication between wrappers and the Query Manager and the Ontology Builder.

Ontology Builder



- The Ontology Builder (OB) is the collective name of a set of functionalities which will support the creation and maintenance of the GVV of the SINode. Given common model and languages, we need to establish tools for synthesizing ontologies and merging them into a GVV, with the final goal of developing a shareable ontology at the SINode level.
- The ontology building process is a cooperative one, involving the designers, the wrappers of the sources providing raw data to the OB, which performs the integration, saves the results in the Metadata Repository, and publishes them to the BAs.
- The building process begins with the creation of a common thesaurus of the information provided by wrappers, that is terminological intensional and extensional relationships describing intra-schema knowledge about classes and attributes of each source schemas.

Query Manager



- The Query Manager is the coordinated set of functions which take an incoming query, define a decomposition of the query according with the mapping of the global virtual view of the SINode onto the specific data sources available (GAV approach) and relevant for the query, sends the queries by means of local Query Agents to the wrappers in charge of the data sources, collects their answers, performs any residual filtering as necessary, and finally delivers whatever is left to the requesting SEWASIE Query Agent.

Ontology Builder

- Based on such information and on designer supplied relationships capturing specific domain knowledge, the OB performs semiautomatic inter-schema analysis by:
  - exploiting lexicon derived relationships, which are based on processes like synonyms identification or generalisation-specialisation relations, and
  - inferring new relationships.
- All these relationships are considered in the subsequent phase of construction of the ontology. Such an activity is based on hierarchical clustering techniques and supports the emergence of a number of global classes (GVV) representative of all the classes coming from the sources and of a mapping description between the GVV and the local sources.
- Most of ideas for SINodes comes from the MOMIS project and the MIKS project

http://www.dbgroup.unimo.it/Momis
http://www.dbgroup.unimo.it/Miks

The MOMIS project (bibliografy)

- S. Bergamaschi, S. Castano e M. Vincini "Semantic Integration of Semistructured and Structured Data Sources", SIGMOD Record Special Issue on Semantic Interoperability in Global Information, Vol. 28, No. 1, March 1999
- D. Beneventano, S. Bergamaschi, S. Castano, A. Corni, R. Guidetti, G. Malvezzi, M. Melchiori e M. Vincini: "Information Integration: the MOMIS Project Demonstration", International Conference on Very Large Data Bases (VLDB'2000), Cairo, Egypt, Settembre 2000
- S. Bergamaschi, S. Castano, D. Beneventano e M. Vincini: "Semantic Integration of Heterogeneous Information Sources", Special Issue on Intelligent Information Integration, Data & Knowledge Engineering, Vol. 36, Num. 1, Pages 215-249, Elsevier Science B.V. 2001
- D. Beneventano, S. Bergamaschi, F. Guerra, M. Vincini: "The MOMIS approach to Information Integration", IEEE and AAAI International Conference on Enterprise Information Systems (ICEIS01), Setúbal, Portugal, 7-10 July, 2001.
- Silvana Castano, Valeria De Antonellis, Sabrina De Capitani di Vimercati: Global Viewing of Heterogeneous Data Sources. TKDE 13(2): 277-297 (2001)

Components (actors and stages)

- There is another family of scenarios of interest, that is those concerning the creation of a new node, the update of an existing node, and the cancellation of a node. These scenarios describe the structural life of a SEWASIE system, namely its growth and change in time.
- The **creation of a new** node is the acquisition of new information sources and the organisation of them into an information unit (SINode). This is a semi-automatic process with the goal of
  - configuring the appropriate wrappers allowing access to the data and their structures,
  - building an ontology, that is a global virtual view (GVV) and the mapping description between the GVV itself and the integrated sources,
  - configuring the query manager for optimal handling of queries within this node, and
  - notifying the brokering agents network about the new node (or instantiating a new brokering agent for the new node)
- The **update of an existing node** concerns structural changes within the node, i.e.
  - changes of the ontology,
  - changes of source structure which imply adaptation at the node level
  - addition/deletion of a source which imply a change of the ontology and adaptation at the brokering agent level
- Notice that the above cited changes do not concern changes of the data content.
- The **deletion of a node** concerns the removal of the references to the node from all the brokering agents in the network, and the subsequent termination of the activities of the node.

Agents

- The SEWASIE project will develop a FIPA compliant trusted agent network, featuring completely open, scalable and secure-oriented architecture issues with the aim of making available the knowledge as synthesized in semantically enriched nodes of a virtual network.
- The advantages of an agent architecture in a context like SEWASIE are given by
  - savings of bandwidth: the agents can move locally to the resources they want to use and carrying along the code to manage them
  - ability to deal with non-continuous network connections, and therefore be intrinsically suited for mobile computing
- On the other end, the use of mobile, autonomous agents may add some complexity to the overall picture, due to the potential autonomy and indeterminacy of their plans of action.

Query Agents

- Query Agents play the role of query managers and "motion item" of the system, therefore a query agent should be able to do several jobs:
  - carrying a query plus the relevant pieces of the user ontology/profile
  - interacting with brokering agents in order to:
    defining the query plan (X), doing the query rewriting for a specific SINode and merging the result from several SINodes
  - carrying back the results (both data and metadata)
  - moving from location (SINode/Brokering Agent) to location (SINode/Brokering Agent)

Query Agents are instantiated by the users for each request to the system, but also by the system itself

Brokering Agents

- The brokering agents are responsible for maintaining the knowledge about the SEWASIE network as well as the information on the specific content of the SINodes.

  A brokering agent
  - knows about the ontologies which are present in the underlying SINode,
  - has some information about related ontologies in other nodes, and
  - has generic information about other ontologies

- The depth of the information of the BA becomes more and more shallow with the distance (with respect to some metrics) between the ontologies where it is "expert" (those of the underlying SINode) and other ontologies covered within the system. Its information on other (non local) ontologies is incomplete.

Brokering Agents

- The brokering agent is able to meet a query agent and recognise that the query presented by the latter is within scope for its local ontologies. If the query presented by the query agent also matches the strains of ontologies which are known to the brokering agent as being present on other nodes, then the brokering agent will also direct the query agent towards such brokers for further processing. When the info comes back to the query agent from the local node, then the query agent may need to interact with the brokering agent to clarify the semantics and context of the result and possibly integrate it with the results from other nodes.
- Whenever a match of the incoming query ontology does not occur with the local ontologies, then the brokering agent will provide routing information towards other brokers to the query agent, which will then leave and move to other nodes.
- The second main functionality of the brokering agent is to receive and classify its local ontologies and the references to other ontologies in the system. This means that whenever a new node is born, or changes, or disappears, then the ontologies used within the node have to be published to the local broker, which will update its internal information and then broadcast to other brokers a manifesto of its available ontologies.
- Reinforcements are possible for specific brokers when the incoming ontological info has a strong correlation with local ontologies; in this way specialist brokers may arise within the SEWASIE system, as well as pure informants on a topic or range of related topics.

Sewasie in a P2P architecture

(i.e. the case of a dynamic world)

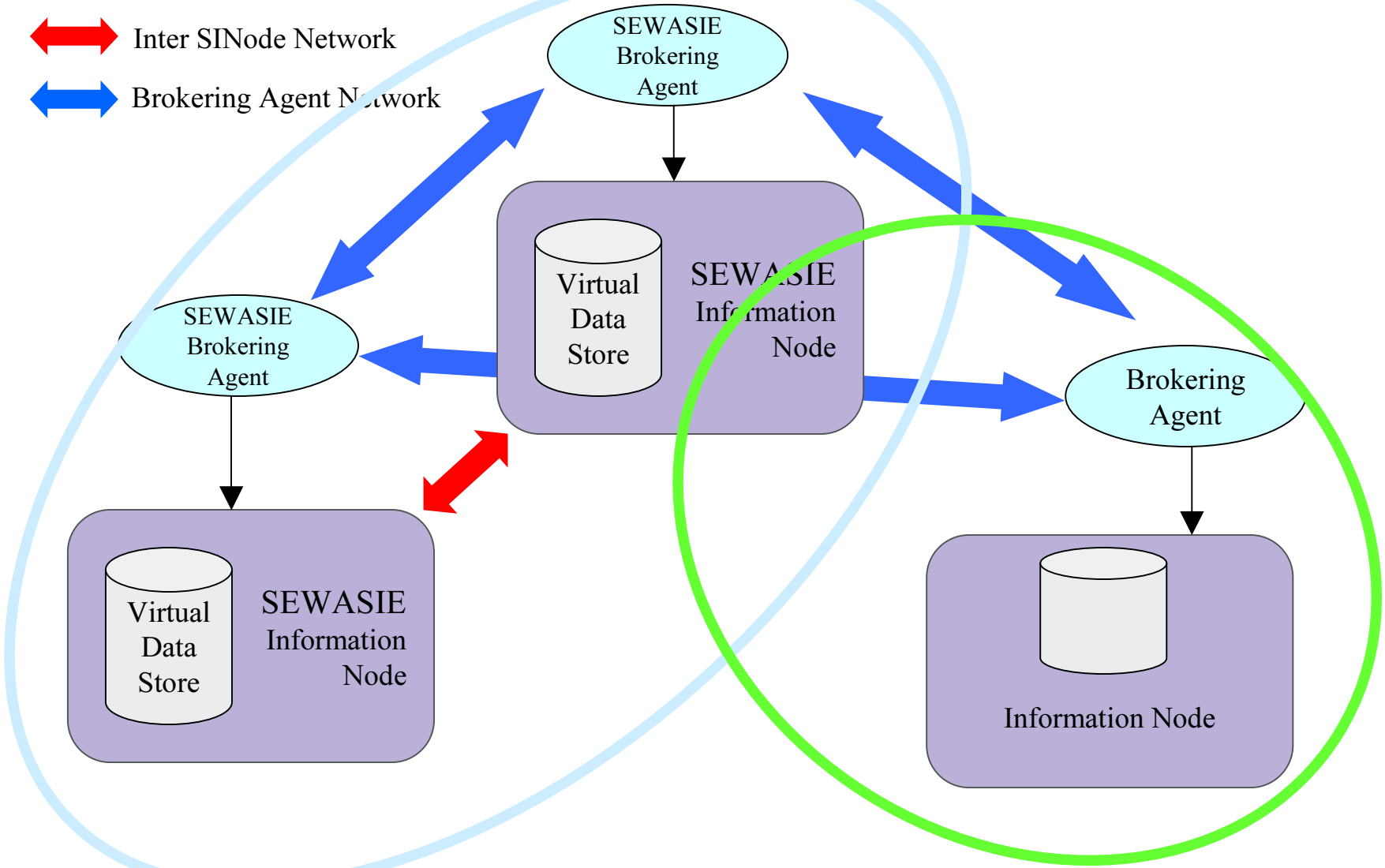
- P2P computing consists of an open-ended network of distributed computational peers, where each peer can exchange data and services with a set of other peers called acquaintances.
  In the general case, a P2P system has no centralized schema and no central administration.
- In the SEWASIE architecture, we rely on two centralized aspects:
  - The brokering agent (global control) that holds the knowledge of the overall network
  - The global schema or data repository of the network
- We can define two alternative P2P networks:
  - INTER SINode Network
  - Brokering Agent Network

[S. Bergamaschi, F. Guerra, **Peer to Peer Paradigm for a Semantic Search Engine**, in proceedings of the International Workshop on Agents and Peer-to-Peer Computing, to appear in LNCS 2530, Springer]

Sewasie in a P2P architecture

- The INTER SINode network allows all the SINodes to exchange information
  - A SINode provides to other SINodes the knowledge about the involved information sources.
  - It is possible to specify coordination formulas that explain how the data in one peer must relate data in a acquaintance.
- The Brokering Agent Network
  - Within the Brokering Agent Network, each Brokering Agent communicates with other peers in order to have information about the involved sources.

SEWASIE in a P2P architecture



Sewasie in a P2P architecture

- This architecture generates a distributed knowledge about the involved information sources
- The Brokering Agent P2P network may provide a support for generating coordination formulas (e.g. by using schema matching, by deriving relations among the peers using inference techniques).
- The Brokering Agent P2P network supports the generation of the query plan in order to identify which are the SINodes to be queried. In particular, the P2P Network can:
  - Generate interest groups with nodes that have similar content.
  - Help the query optimization, by giving information about the "data placement". A peer knows how is distributed data and in this way the query plan may take into account the existing resource and bandwidth constraints.

Sewasie in a P2P architecture

- SINode network is an alternative approach: we maintain a single brokering agent, holding the knowledge of the network topology and we need a P2P layer in each SINode with the following functionalities:
  - The P2P layer needs a protocol for establishing an acquaintance dinamically
  - The P2P layer offers semi-automated support for generating coordination formulas
  - The P2P layer uses approaches for query processing of multi-database systems
  - The P2P layer should be able to advertise its ontology