IST-2001-34825

Marseilles, 27 November, 2003

SEWASIE: a semantic search engine

Sonia Bergamaschi - project Coordinator
DBGroup (www.dbgroup@unimo.it)
Università di Modena e Reggio Emilia

Outline of the talk

- What is SEWASIE?
  - The project at a glance
  - The Business Scenario
  - The SEWASIE Architecture
  - A querying scenario
  - Architecture Guidelines
  - Resulting Product
  - SINodes & MOMIS
- A P2P paradigm for SEWASIE (for later discussion)
  - INTER SINode Network
  - Brokering Agent Network

The SEWASIE Consortium

- Università degli Studi di Modena e Reggio Emilia (I, coordinator)
- CNA Servizi Modena S.c.a.r.l. (I)
- Università degli Studi di Roma “La Sapienza” (I)
- Rheinisch-Westfaelischen Technischen Hochschule Aachen (D)
- Libera Università di Bolzano (I)
- Thinking Networks AG (D)
- IBM Italia S.p.A. (I)
- FhG/FIT (D)

SEWASIE
The Project at a glance

- SEWASIE is implementing an advanced search engine that provides intelligent access to heterogeneous data sources on the web via semantic enrichment.
- 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

The Business Scenario

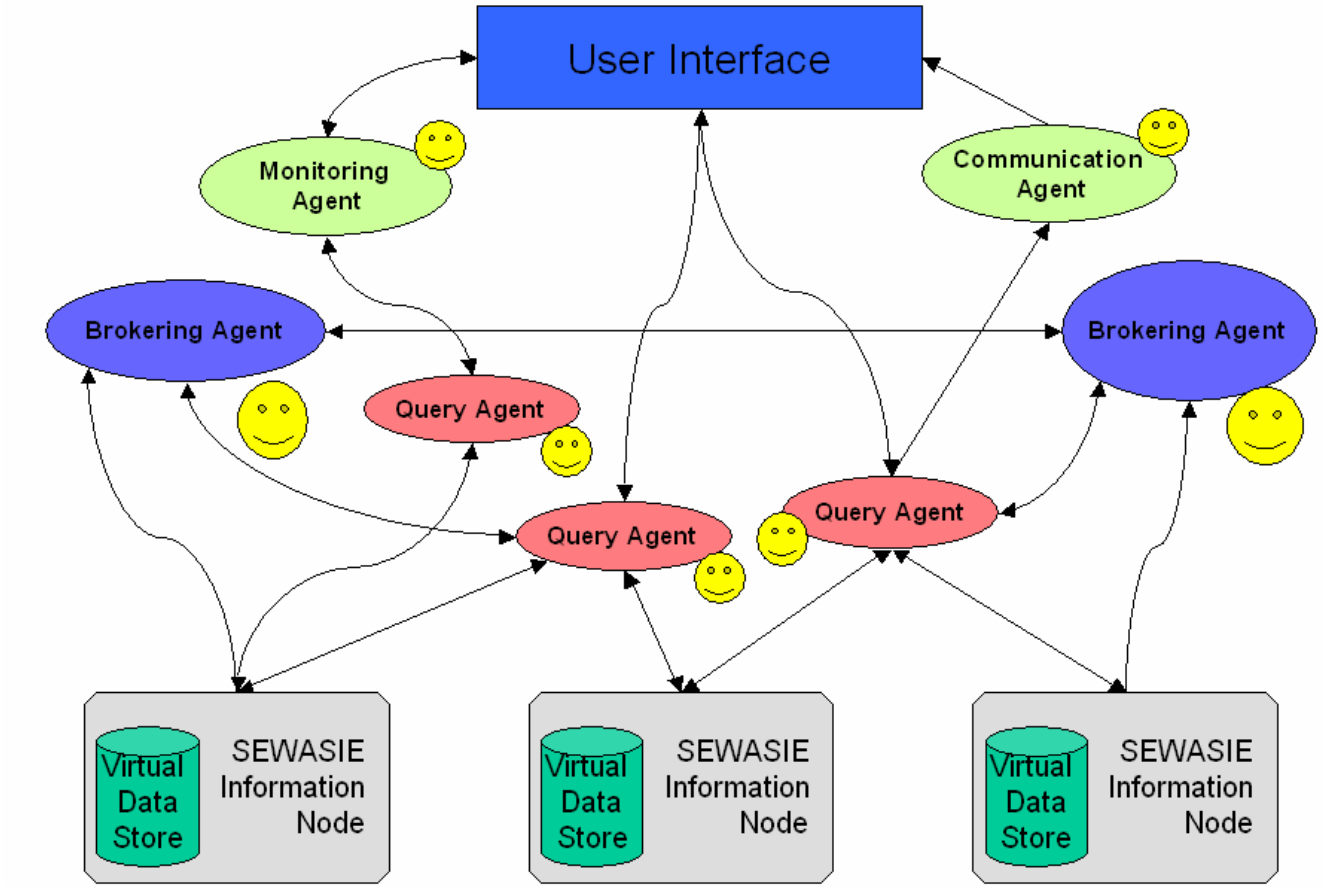
- One of the keys to sustainability and success is being able to access information. Throughout Europe, much of the industrial fabric is made of small and medium-sized enterprises (SMEs) in fields such as agriculture, manufacturing, commerce and services. This could be a cheaper supplier, an innovative 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.

The Business Scenario

- 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 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?

The SEWASIE architecture



The SEWASIE architecture

The SEWASIE project aims to develop an advanced search engine enabling intelligent access to heterogeneous data sources on the web, via semantic enrichment.

- 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 (with a domain ontology), 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
  - Ontologies are multilingual

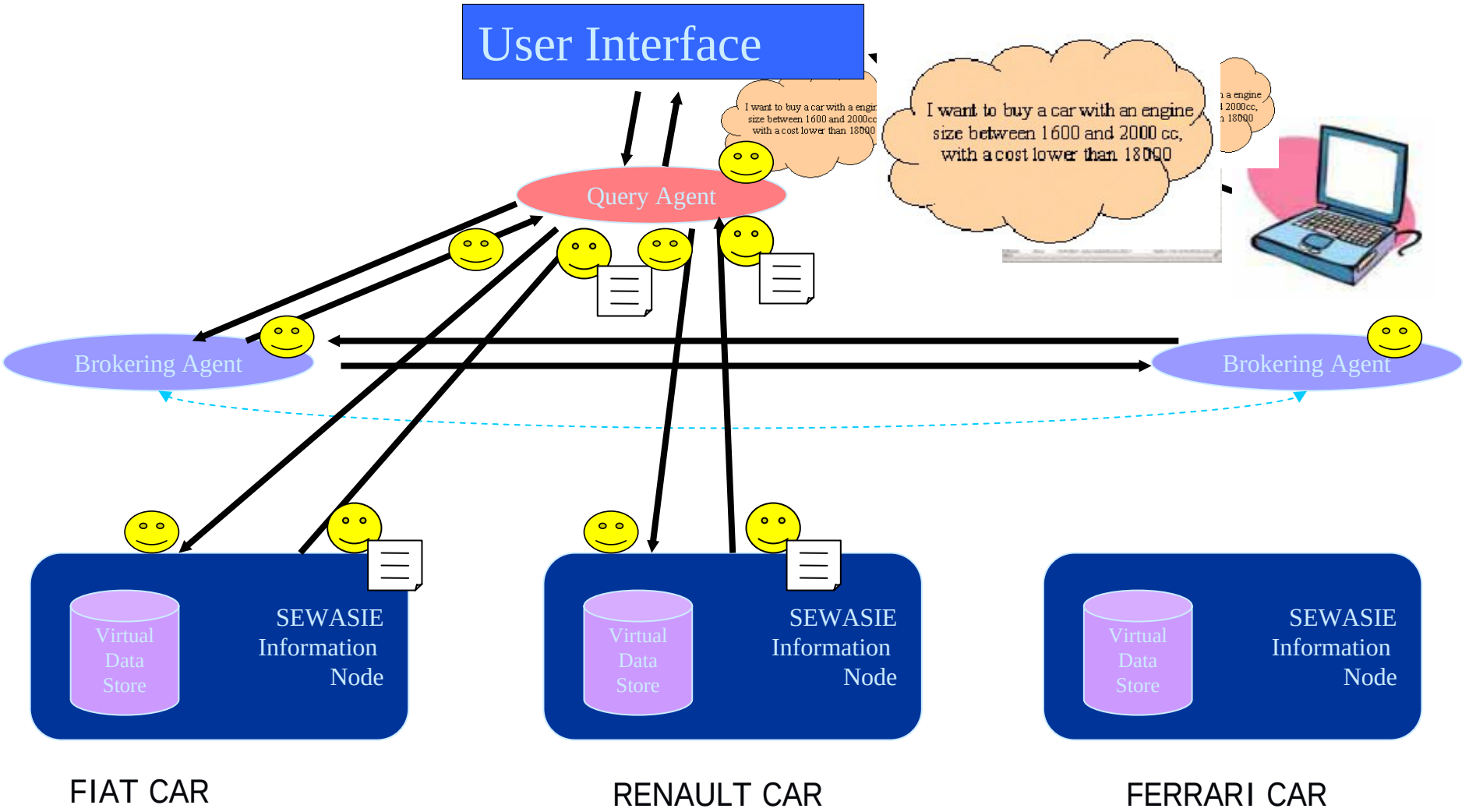
The SEWASIE Architecture

- The SEWASIE project will develop a **FIPA compliant trusted agent network**, open, scalable and secure with the aim of making available the knowledge as synthesized in semantically enriched nodes of a virtual network.

  The **Brokering Agents** maintain the knowledge related to the SEWASIE Virtual Network and the user profiles, classify SINodes, are 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 collects partial answers.
  - To select SINodes useful to solve a query, a **Query Agent** interacts with a **Brokering Agent**.

First Scenario- querying SEWASIE (example)



Global Project Development Guidelines

- Definition of the main features of the architecture completed
- Main features of field and development environments identified: standard commercial environments (UNIX, Windows), web services, SOAP+UDDI+WSDL
- Global rules for low-overhead controlled development established: development language (Java); Rational Unified Process (RUP), UML; integration and testing procedures at module and component level (with possible escalation and regression)

Main innovations: Ontologies and multi-lingual issues

- ODLI3 will be used as common ontology description language
- The need for ontology management tools (creation, modification, mapping) has emerged within several modules in the SEWASIE architecture, namely at the
  - Data providers level (SINodes)
  - Intermediaries (brokering agent) level
  - User level
  - and has been handled in a unified way
- SINodes will include Virtual Data Stores defined with multiple languages and will be unified in a unique ontology expressed in an intermediate language for publication by the Brokering Agents
- The user will be able to query the system in natural language, guided by his local ontology; the user interface will convert the query into the intermediate language terminology as part of the initial formalisation of the query

Main Innovations: Agents

- The working hypothesis for the basic agent architecture is to adopt the standard framework laid out by FIPA

  Main advantages:

  - Savings of bandwidth
  - Ability to deal with non-continuous network connections
  - Autonomy
- Agent classification[1]
  - Query Agents: provider agents
  - Brokering Agents: middle agents
  - Monitoring Agents: request agents
  - Communications Agents: request agents

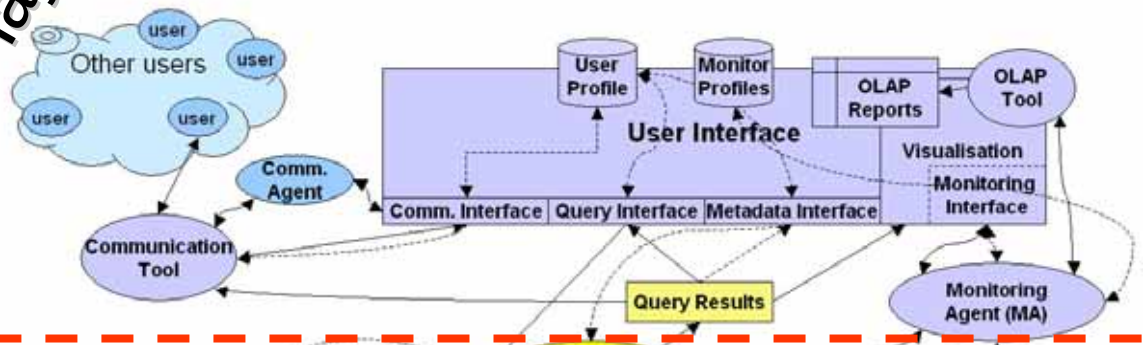
[1] The classification of agents is taken from: **M. Klusch editor,** Information Agent Technology for the Internet: a survey. *Data and Knowledge Engineering, 36(3)*, 337-372, 2001.

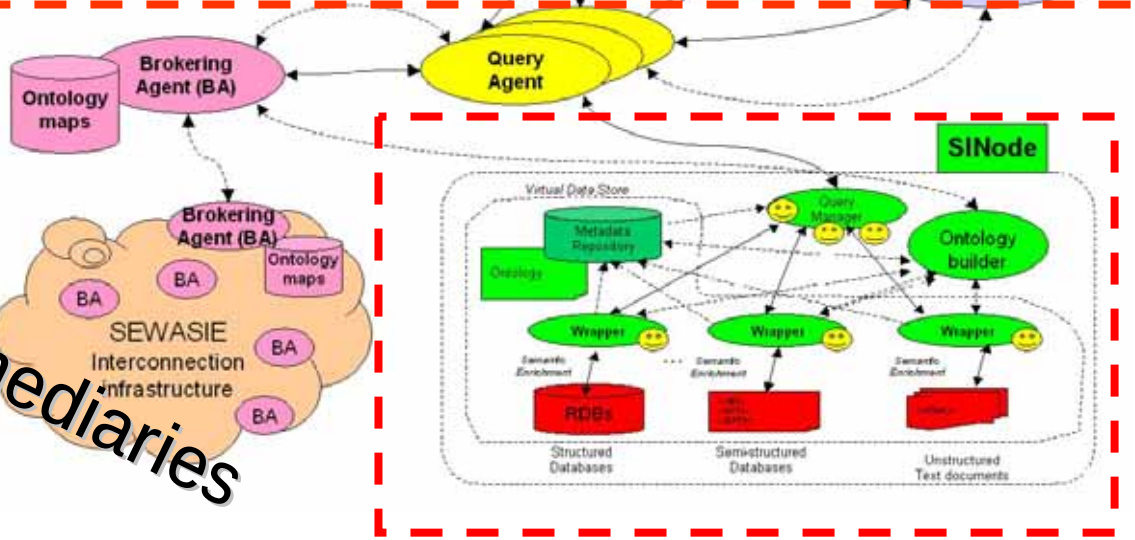
The SEWASIE user interface

- Goals
  - Offer an integrated, homogeneous environment where information may flow and be manipulated by different tools
  - Support the introduction of the semantic dimension in the non-technical-user scenario
- Innovative methods and technologies
  - Graphical user interface for semantic query formulation
  - Integration of OLAP tools and domain ontologies
  - Ontology-based negotiation support

SEWASIE detailed architecture

The user interface layer

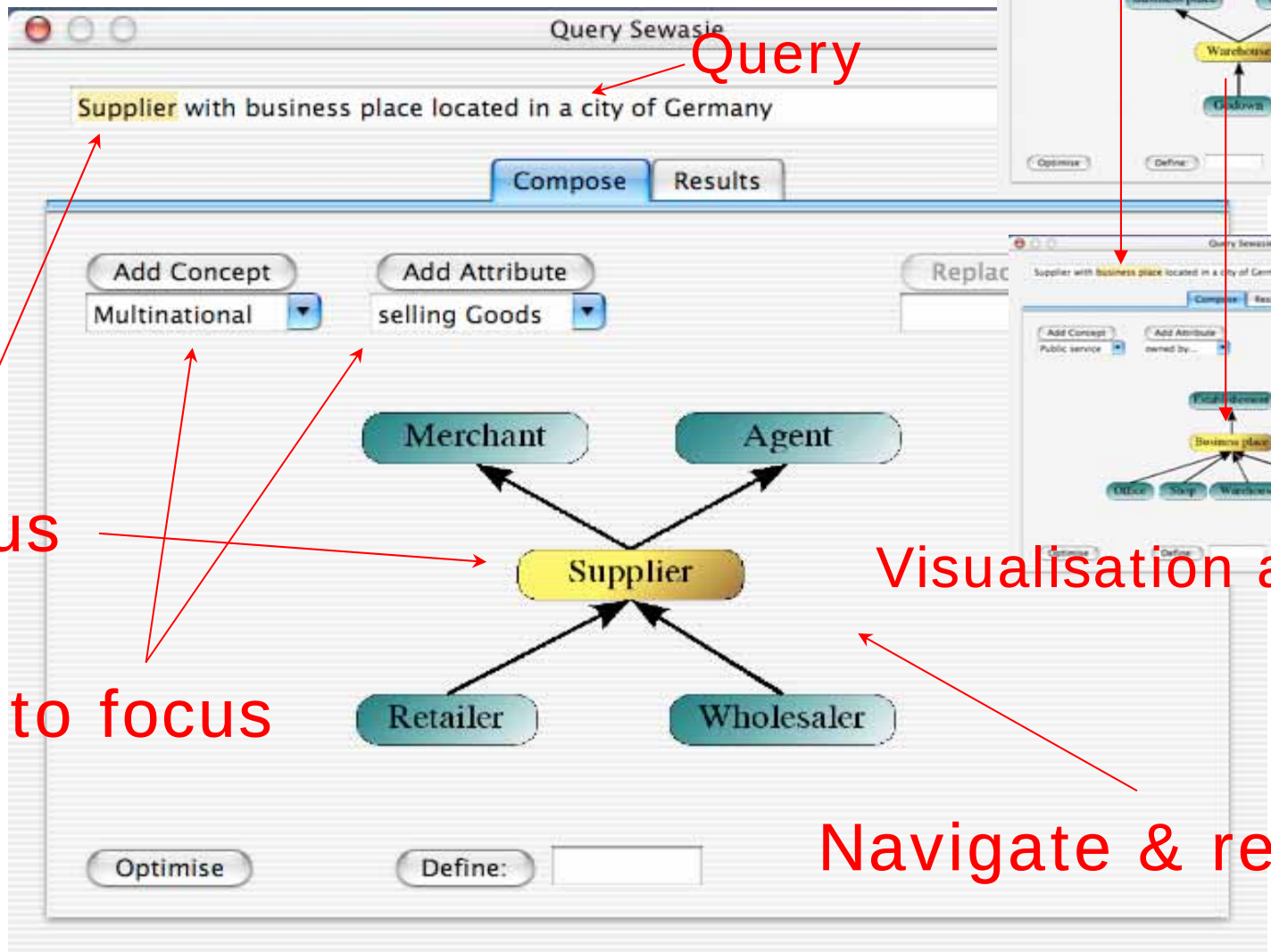




The intermediaries layer

The information layer

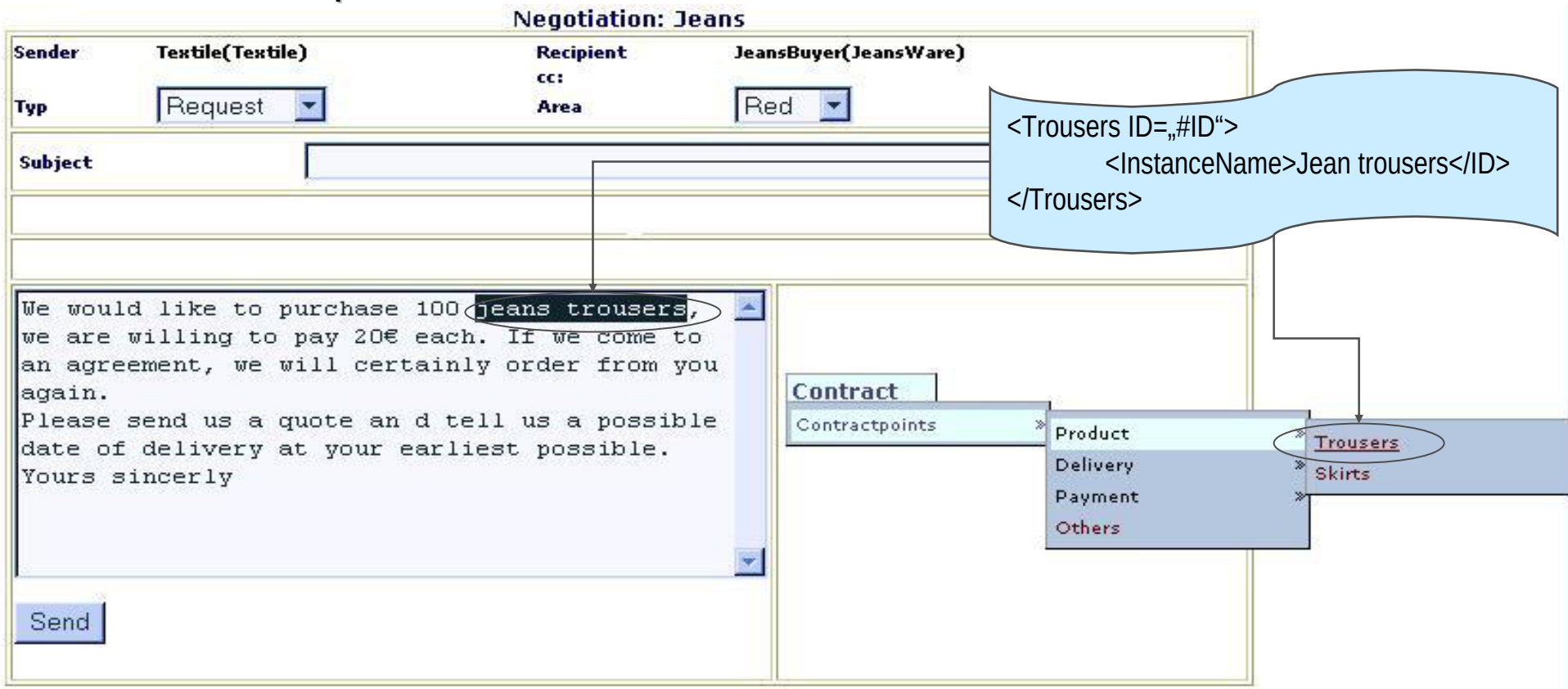
The semantic query interface



The user interface: negotiation

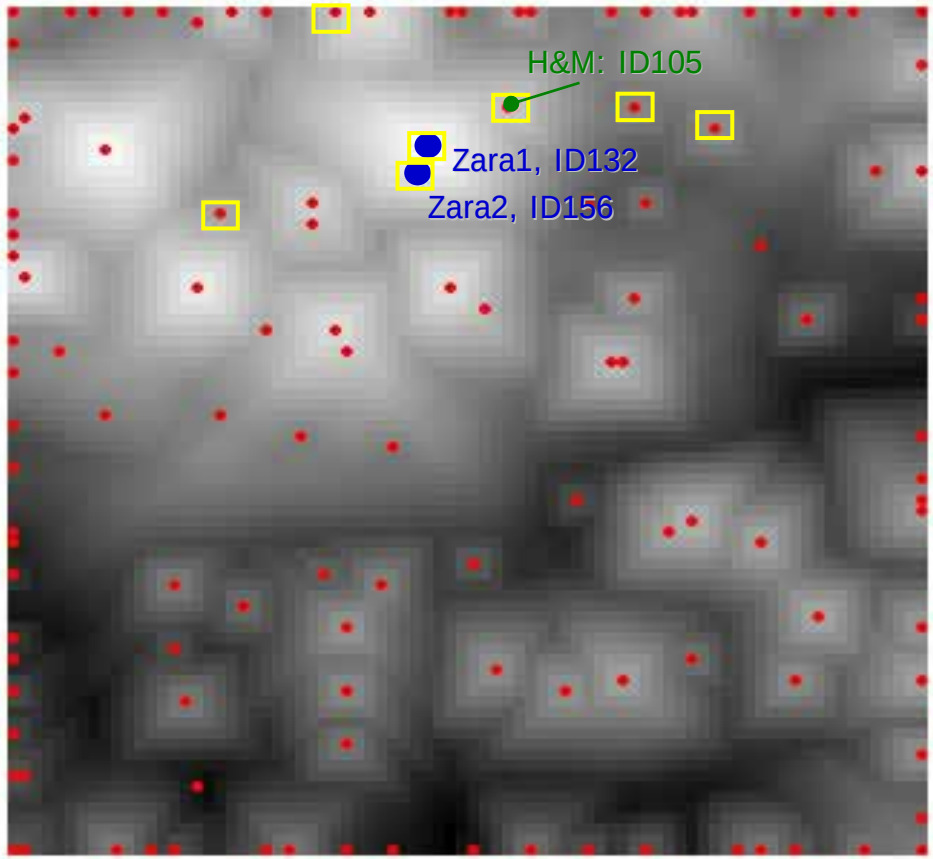
- An example of an ontology-enriched negotiation

"A client requests a textile company to deliver 100 Jeans trousers with a unit price of 20€ "



The user interface: OLAP and model based text exploration

- End-Customer
  - B2B
    - Retailer
    - Wholesaler
  - B2C
- Garments
  - Outerwear
    - Jeans
    - Dresses
    - Suits
  - Underwear
  - Footwear



Observation:

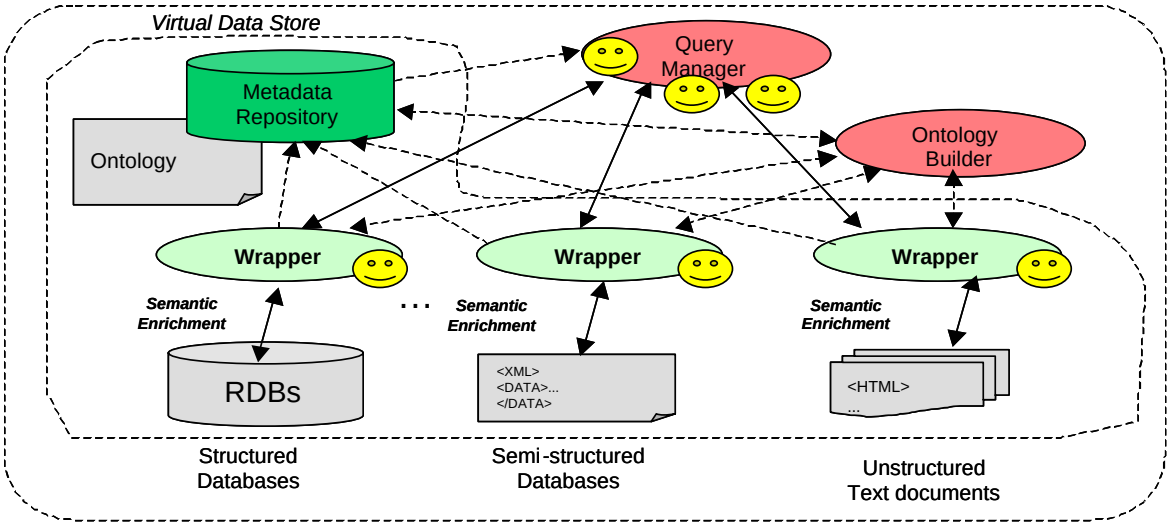
Major competitor H&M performed well in Germany.
In particular, the jeans sales were profitable.

User requirements and resulting product profile

- Three kinds of utilisation have been detected:
  - *"**End users**"*: simple *searchers*, using fee-based or free searching services. The latter if the problem that requires a search activity does not need specific answers but only statistically reliable ones. In this case, the system supplies answers containing little information.
  - *"**Middle users**"*: domain experts using SEWASIE to create a service. They *create an ontology* which has to be reliable and precise. The created services are used by the end users.
  - *"**SEWASIE as integrator of Information Systems in a middle-large enterprise**"*: the user (directly or through a software company which sells SEWASIE) is able to face the problem completely managing the SEWASIE software. He knows how to create the necessary ontologies exploiting all the tools SEWASIE offers. This kind of user purchases the software and the know-how to manage it.

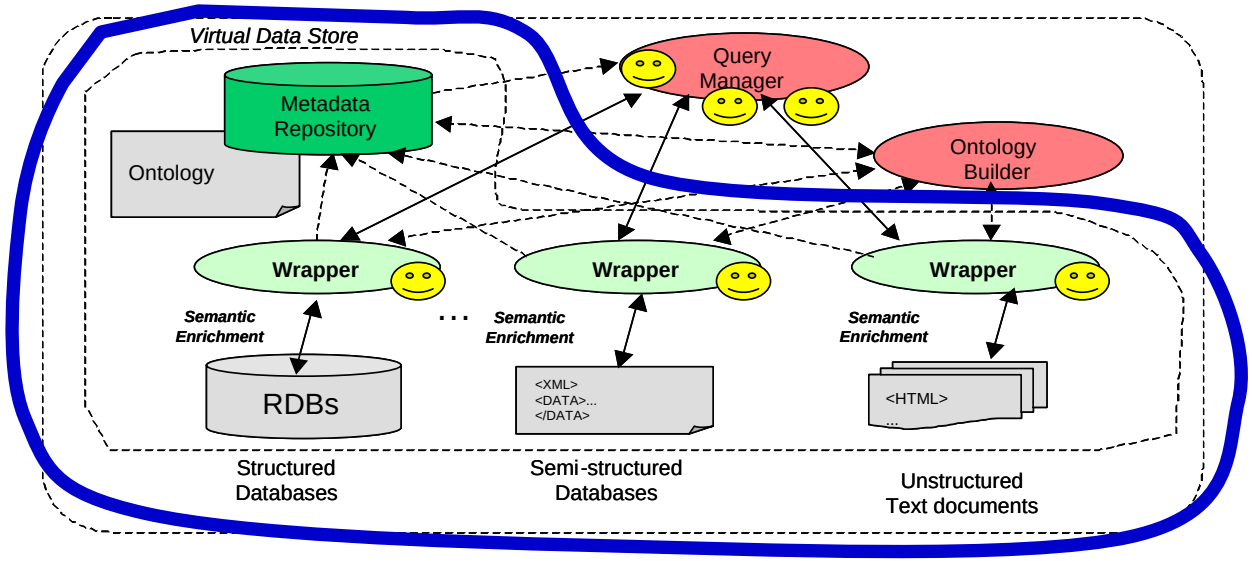
The SINode module

- ***SINodes*** are mediator-based systems, including:
- A ***Virtual Data Store*** (VDS) represents a Global Virtual View (GVV) 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 annotated Global Virtual View of the sources and the mapping descriptions between the GVV itself and the annotated sources.
- The ***Metadata Repository*** holds the ontology and the knowledge required to establish semantic relationships between the SINode itself and the neighbouring ones.

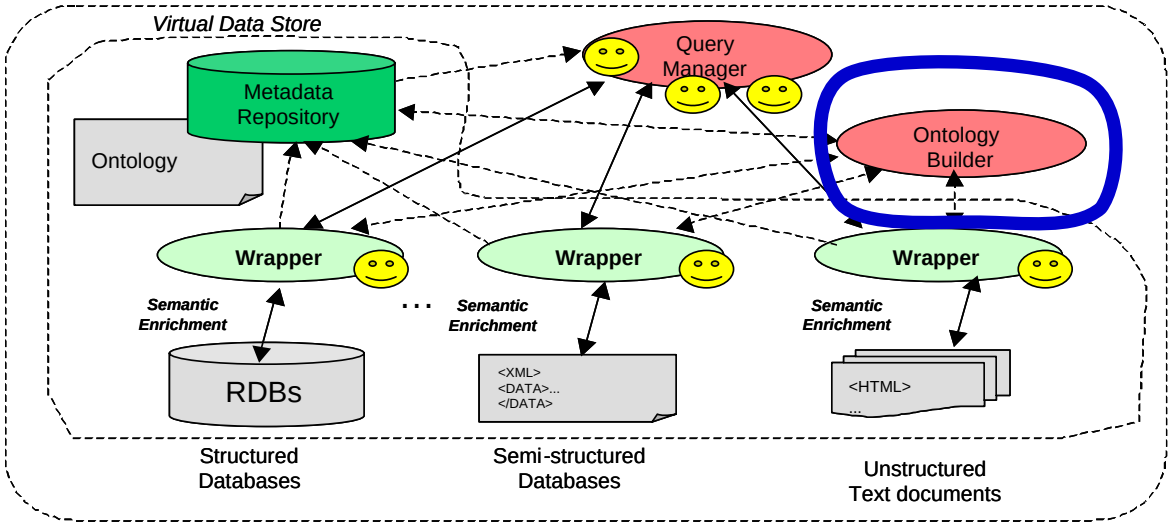
Virtual Data Store



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 the 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.
- The selected data model and associated languages is ODMI3 (ODLI3), which was derived from ODMG specification; the selected query language is OQLI3.
- 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 (at the moment an XML exportation of the metadata is available).

Ontology Builder



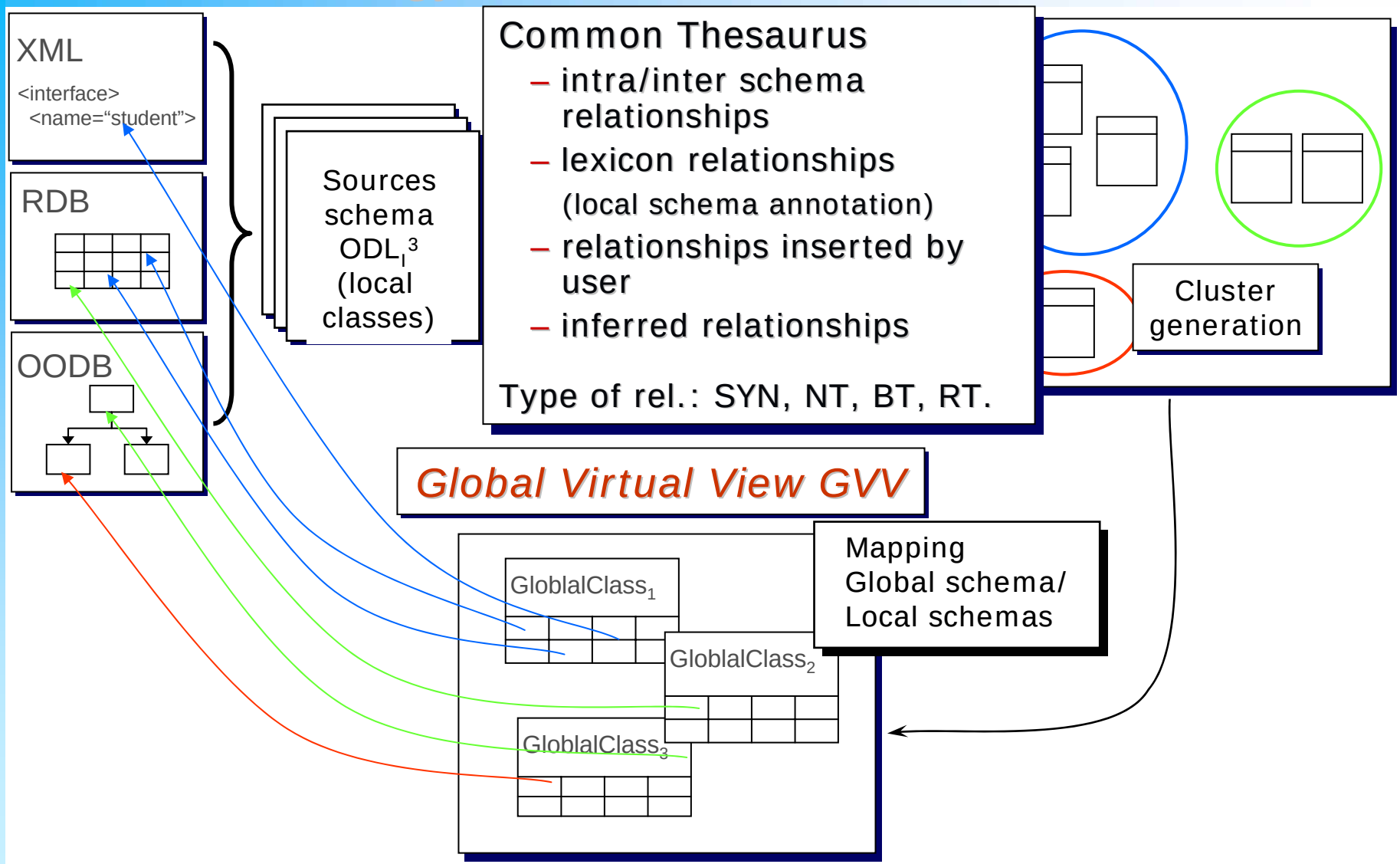
- The Ontology Builder (OB) is the collective name of a set of functionalities which will support the creation and maintenance of the GVV of the SINode. Given common model and languages, we need to establish tools for synthesizing ontologies and merging them into a GVV, with the final goal of developing 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 extraction of the structures of the data sources(local schemata) and the creation of a Common Thesaurus of the information extracted by wrappers, that is relationships describing intra-schema knowledge about elements (classes and attributes) of each source schemas.

MOMIS Ontology Builder

- Performs the integration of the schemas of heterogeneous data sources, such as relational, object, XML, or semistructured sources in a semiautomatic way.
- Starting from local source descriptions, the OB generates an integrated view (Global Virtual View (GVV)) of all data sources and expresses those views using both ODLI3 and XML.
- OB creates the GVV in different stages, first by creating a Common Thesaurus (CT) of intra-schema relationships. After this initial phase, OB enriches the CT with interschema relationships obtained using the lexical WordNet system (www.cogsci.princeton.edu/wn), which identifies the affinities between interschema concepts on the basis of their lexicon meaning. Moreover, the Common Thesaurus contains relationships inferred by the interaction with ODB-Tools (a tool based on Description Logics techniques), and relationships added by the user interaction.
- OB enriches the Common Thesaurus by evaluating structural affinities among interschema concepts (by using ARTEMIS).
- Most of ideas for SINodes comes from the MOMIS project

  http://www.dbgroup.unimo.it/Momis
  and MIKS (MOMIS + agents)

MOMIS Ontology Builder



Multilingual functionalities

- SEWASIE multilingual technologies will allow users to share information and resources available all over the world, but also to preserve their original local qualities
- the Ontology Builder must be able to represent and fully preserve multilingual information
- A *multilingual lexical system* will be integrated with the Ontology Builder in order to provide a resource for cross-language analysis of elements describing the meta-information of the data sources (MultiWordnet/Eurowordnet)

Multilingual functionalities: source schemata annotations

- An *annotation* process is needed to render explicit the meaning of multilingual elements describing the meta-information of source schemata with respect to a common *multi-lingual lexicon ontology*
- An *annotation* is a mapping of a given element into a well-defined set of concepts of the adopted multi-lingual lexicon ontology

Two levels of annotations are needed:

- Annotations of the Local classes (in local languages)
  - Word form selection
  - Meaning selection
  - Language selection

  For example:

  USAwear.company = <Business_Organization,English,{synset#123}>

  Ingromarket.azienda = <Impresa,Italian,synset#123}>

  where

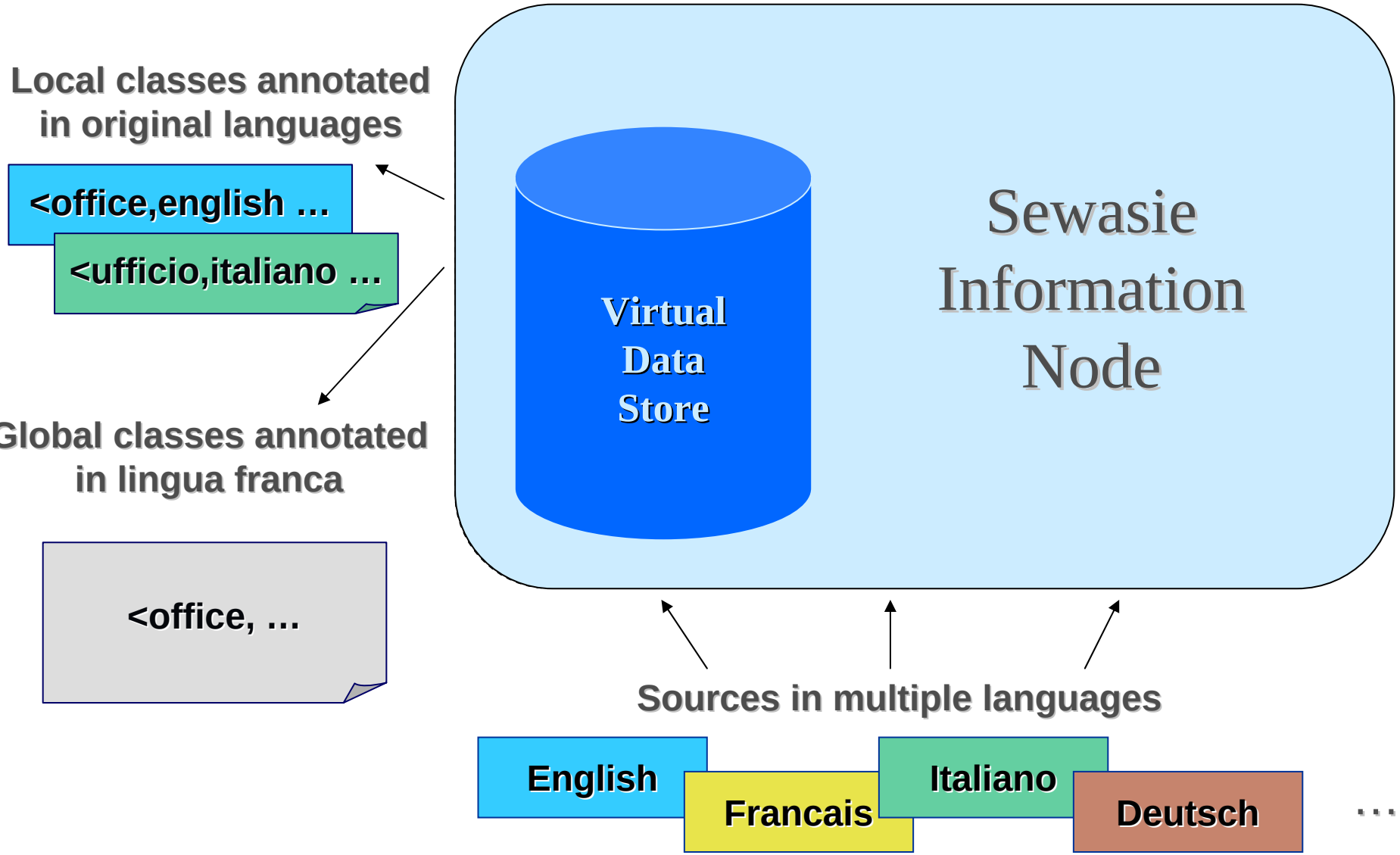
  synset#123 = "an institution created to conduct business"

- Annotation of the Global classes (in lingua franca)
  - Name selection
  - Meaning selection

  For example:

  $Global_0$ = <Business_Organization,{synset#123}>

Multilingual functionalities: annotations



Multilingual functionalities: future plans

- Second release of the system prototype for semantically enriched data stores (Month 25: May 2004):
  - Multilingual ontologies and annotations in SINodes exploiting a multi-lingual lexicon ontology (EuroWordnet, MultiWordnet)
- Final release of the system prototype for semantically enriched data stores (Month 35: March 2005):
  - Enrichment of multi-lingual lexicon ontology with the aid of statistical analysis techniques for multilingual text corpora (for example with techniques for the generation of multilingual dictionaries)
  - Automatic language identification for local annotations
  - Automatic word sense disambiguation techniques could provide useful suggestions to speed up the process of meaning selection (at least for global classes annotations in lingua franca, i.e. in english)

Local sources annotation (in English)

- The integration designer has to manually choose the appropriate WordNet (www.cogsci.princeton.edu/~wn/) meaning for each element of the conceptual schema provided by wrappers.
- The annotation phase is composed of two steps:
  1. Word Form choice. The WordNet morphologic processor aids the designer by suggesting a word form corresponding to the given term.
  2. Meaning choice. The designer can choose to map an element on zero, one or more senses. Notice that the user can choose a sense among the existing ones in WordNet and **he can add new senses in the DB**.

Annotation: Extending WordNet

If a source description element (i.e. a class or an attribute name) has no correspondent in the reference lexical ontology (WordNet in our case), the designer may add a new meaning and proper relationships to the existing meanings.

The designer inserts

- the new meaning (and a new wordform if not present)
- zero or more synonyms.

The designer may add a new sense to an existing lemma by using an existing sense or a new one:

- chooses a sense from a list of candidates obtained by inserting one or more keywords
- writes the English gloss explicitly.

Annotation: Extending WordNet

A synset relationships editor has been developed to allow the designer to view, analyze and modify the actual state of relationships involving the current synset as the source synset, i.e. the first member of any relation.

In particular, two different ways are available to find another target synset:

- by an explicit search of the target sense definitions using one or more keywords
- by exploiting similarity search for the target synset, i.e. a search to get a set of synsets similar to the current one up to a certain degree.

Global Virtual View annotation

- The GVV has to be annotated to become ”exportable knowledge”.
- Annotating a GVV means to provide Global Classes with a name and with lexical meanings.
- By starting from annotations of local sources and mappings between the GVV and the local ontologies, we have developed a semi-automatic methodology to generate the annotations of the GVV.

  D. Beneventano, S. Bergamaschi, F. Guerra, M. Vincini: “Synthesizing an Integrated Ontology”, IEEE Internet Computing, Vol.7,N.5, September/October 2003.

Global Virtual View annotation

Annotated Local classes

CS.Essay=<essay, {essay#1}>
CS.Publication=<publication,{publication#2}>
UNI.Article=<article,{article#1}>

A Global class

GlobalClass1
- CS.Essay
- CS.Publication
- UNI.Article

The CT relationships

UNI.Article	NT	CS.Publication
CS.Essay	NT	CS.Publication

The annotated Global class

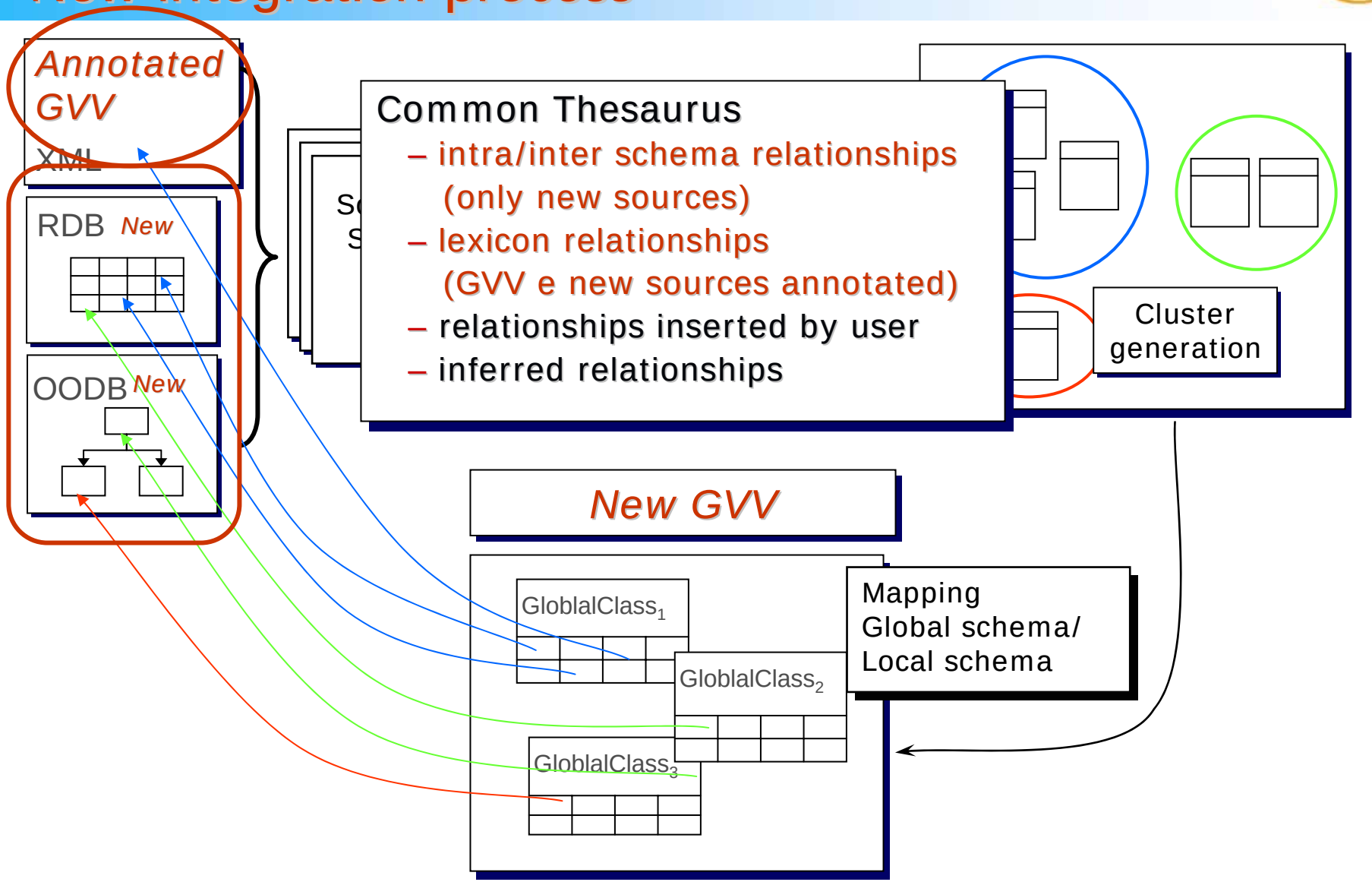
meanings

GlobalClass1 = <publication, {essay#1, publication#2, article#1}>
name

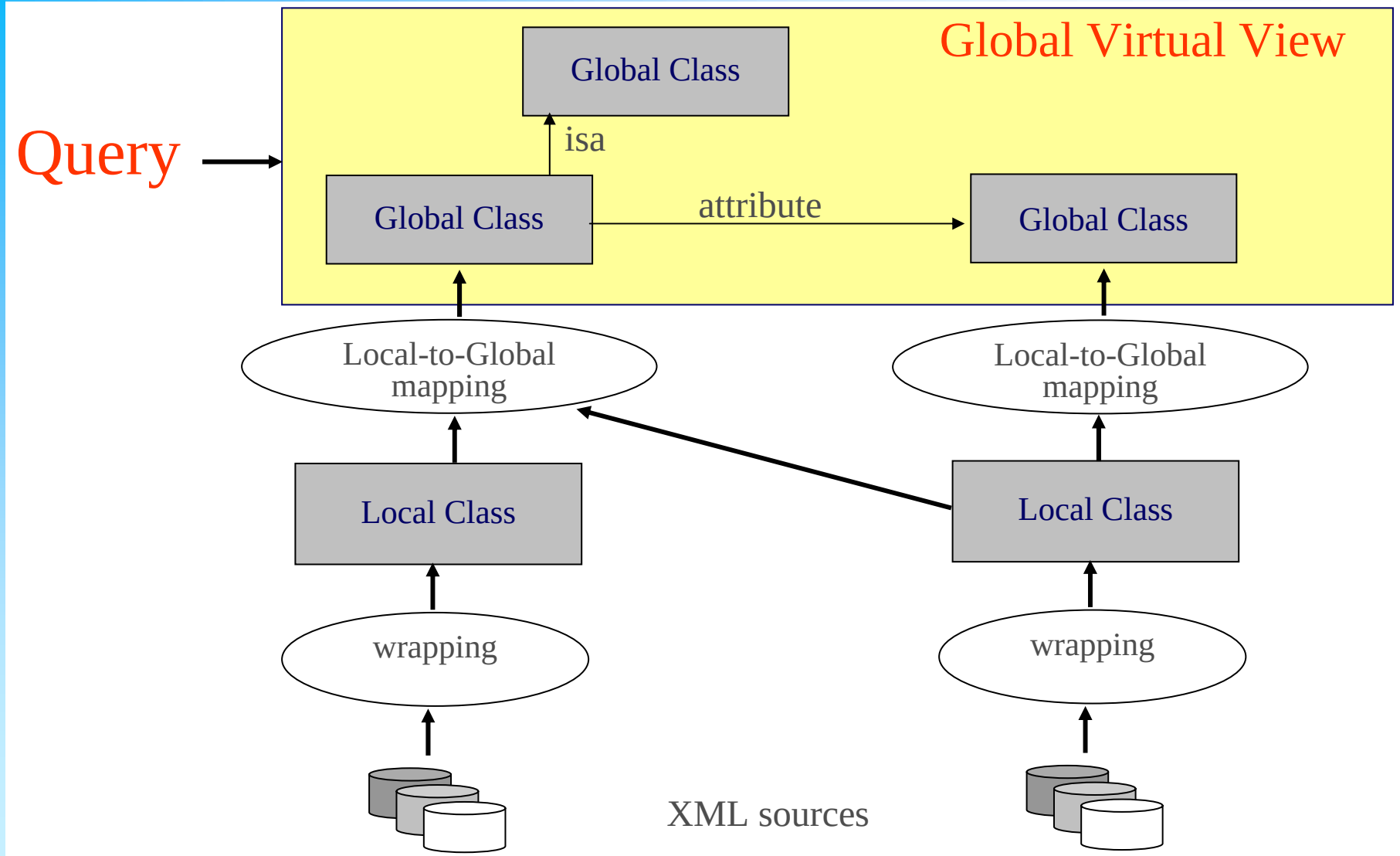
essay#1 = an analytic or interpretive literary composition
publication#2 = a copy of a printed work offered for distribution
article#1 = nonfictional prose forming an independent part of a publication

Word Form Meaning

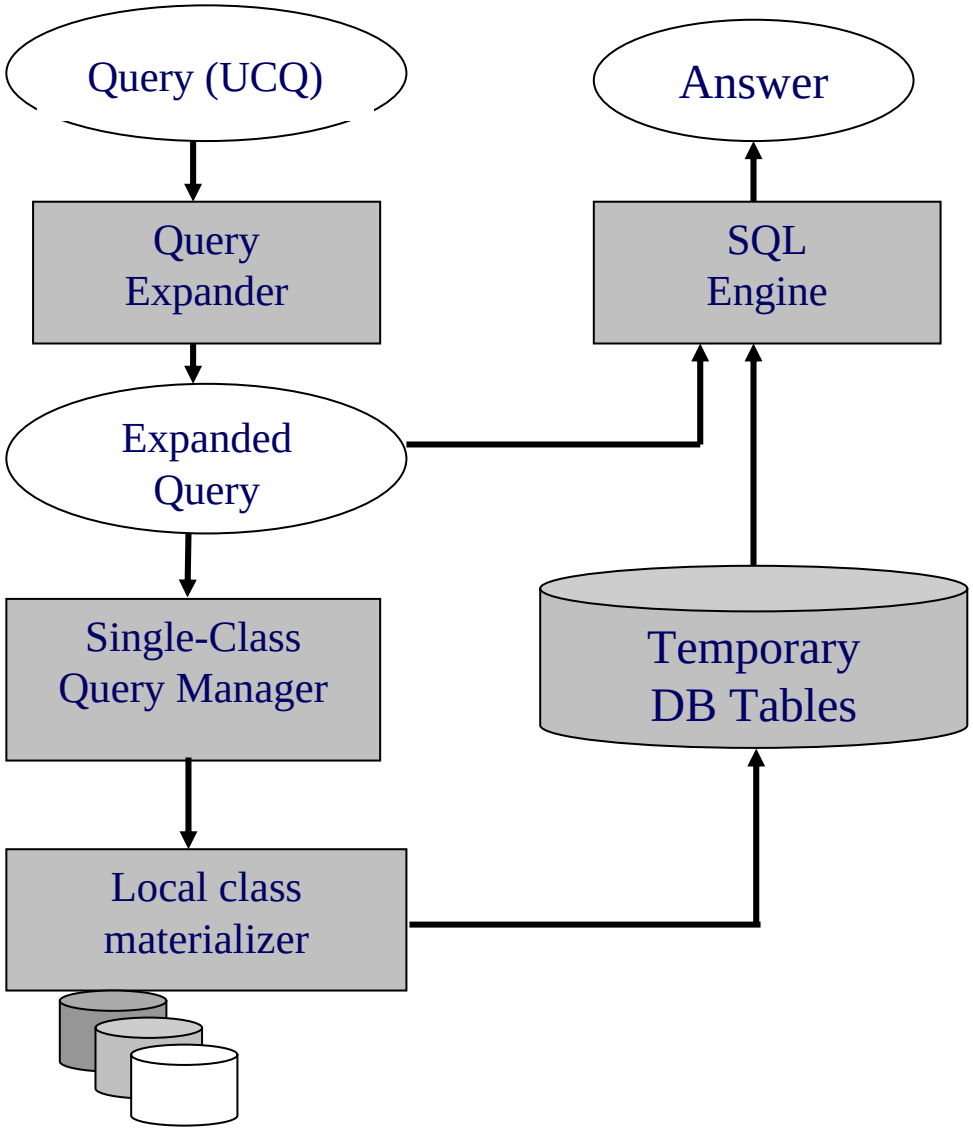
New integration process



Techniques for query management within one SINode



Query processing within one SINode



Query processing within one SINode

- Technique for GAV (Global-as-view) data integration system (can be easily extended to GLAV)
- Semantics of the Global Virtual View is taken into account by a novel technique for dealing with integrity constraints in the global schema (called query expansion)
- Rewriting algorithm for re-expressing the expanded query in terms of the local classes (which are connected to the data sources)
- Data reconciliation performed in the materialization of the global classes

next period activities

- Definition of techniques for query answering in the context of more than one SINode
- Definition of techniques for handling sets of queries, and for information reconciliation
- More sophisticated techniques for information reconciliation
- Second release of the protoype for query management

Other efforts in the SEWASIE arena (1/2)

- The I3 (Intelligent Information Integration) Reference Architecture
  - The I3 Reference Architecture (I3-RA) was developed in the early 90's to provide a general functional description of a general information integration infrastructure with high level capabilities
- The JXTA search architecture
  - The strong point of JXTA is the *Query Routing Protocol* (QRP) , which is an XML based protocol for defining queries, responses, and registrations. Each hub builds a *queryspace*, similar to XML namespaces, with the XML structures of valid queries for each provider.
  - The most prominent element of the system is the *Search Hub Service*. This service performs the routing of queries from consumers to providers. It accepts queries, resolves queries identifying potential providers, routes the requests, and sends the collated responses back to the customer

Other efforts in the SEWASIE arena (2/2)

- InfoSleuth (MCC Corp, USA)
  - The goals of InfoSleuth:
    - Map or relate data sources to ontological concepts
    - Classify information according to a hierarchy of concepts
    - Provide delivery of personalised information and alerts (info channels)
    - Extract or aggregate information and resolve semantic heterogeneity
  - Wants to develop an agent-based system for information discovery and retrieval in an open and dynamically changing environment.
- SWAP (EU-IST-2001-34103)
  - The SWAP project (EU-IST-2001-34103) aims to develop P2P architectures for Semantic Web systems. The P2P environment and support for Knowledge Management functionalities are the qualifying points of this project.

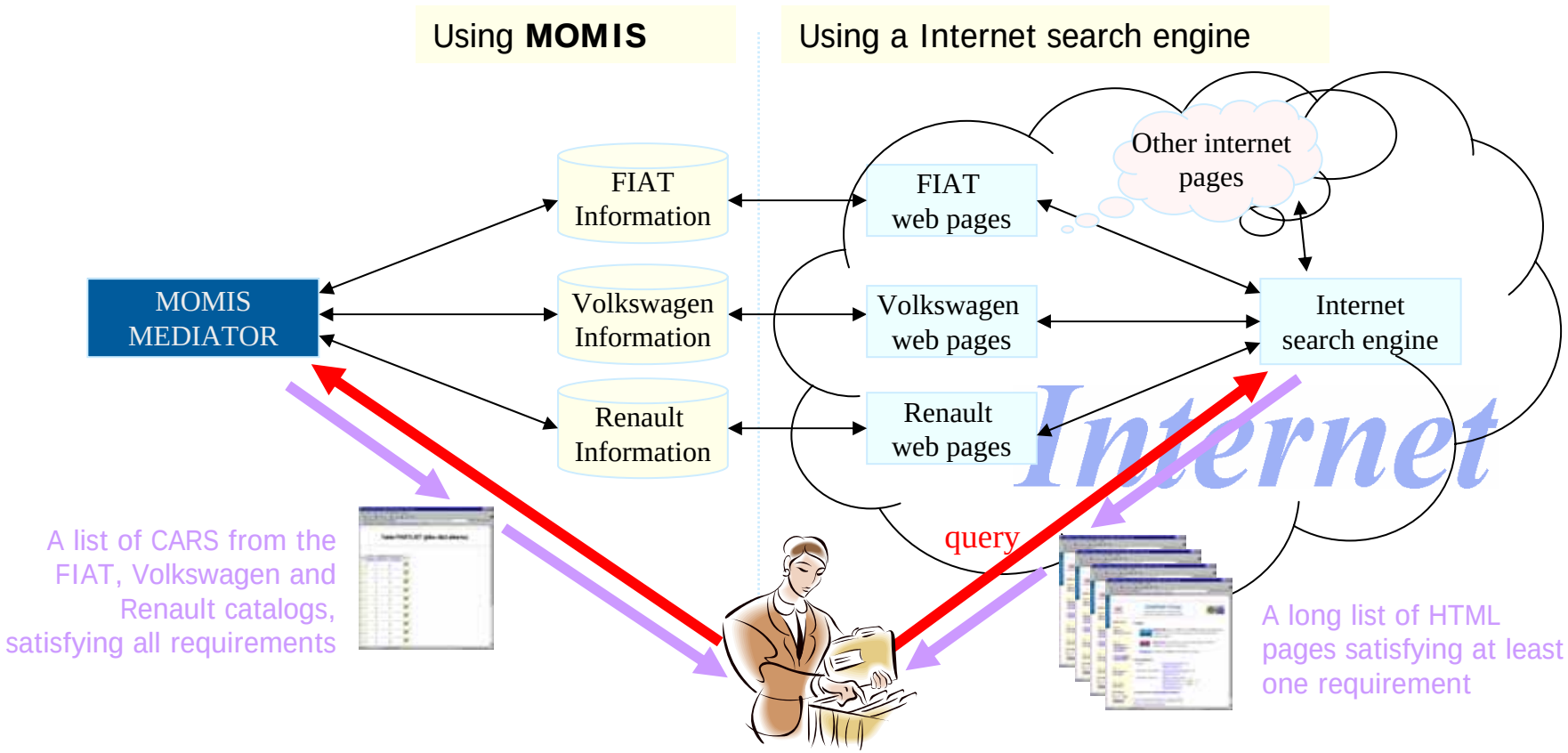
Future Work

- First tests will consider limited functionalities of the system (i.e. the prototypes developed in 2003) and will be run on a limited set of information sources. Following the outcome of these tests, the system will be revised and eventual problems solved.
- Later on, system functionalities will be expanded (2 further releases of the system are expected throughout the project: May 2004 & March 2005) and the context for information retrieving made more complex. Progressive steps will always be tested, so that the research and technological effort will always keep the focus on end users needs.

The MOMIS Project: overview

- Momis (*M*ediator envir*O*nment for *M*ultiple *I*nformation *S*ources) is a mediator-based system for information extraction and integration that works with structured and semistructured data sources.
- MOMIS was developed as a joint collaboration among the University of Modena and Reggio Emilia, the University of Milano and the University of Brescia under the direction of Professor S. Bergamaschi.
- Sources integration is based on the individuation of an ontology shared by each source; the ontology is represented as a set of terminological relationships called *Common Thesaurus*.

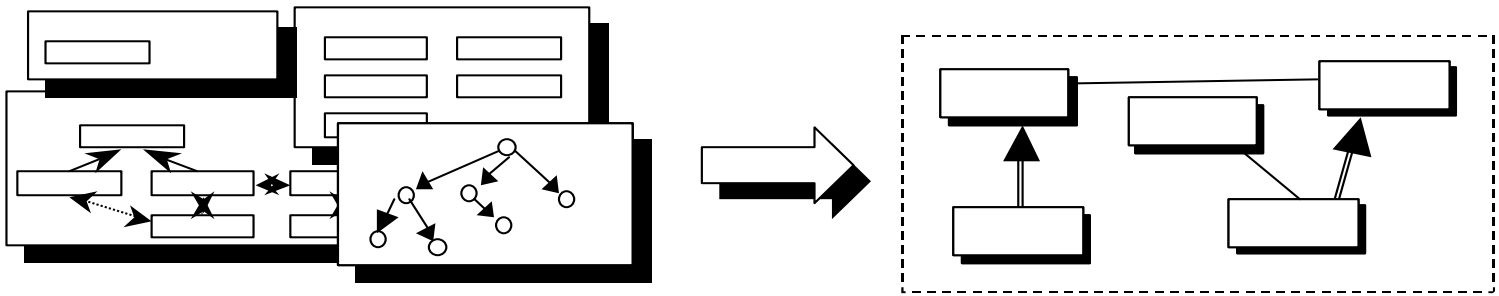
Integration in the e-commerce environment



I want to buy a car with a engine size between 1600 and 2000cc, with a cost lower than 18000

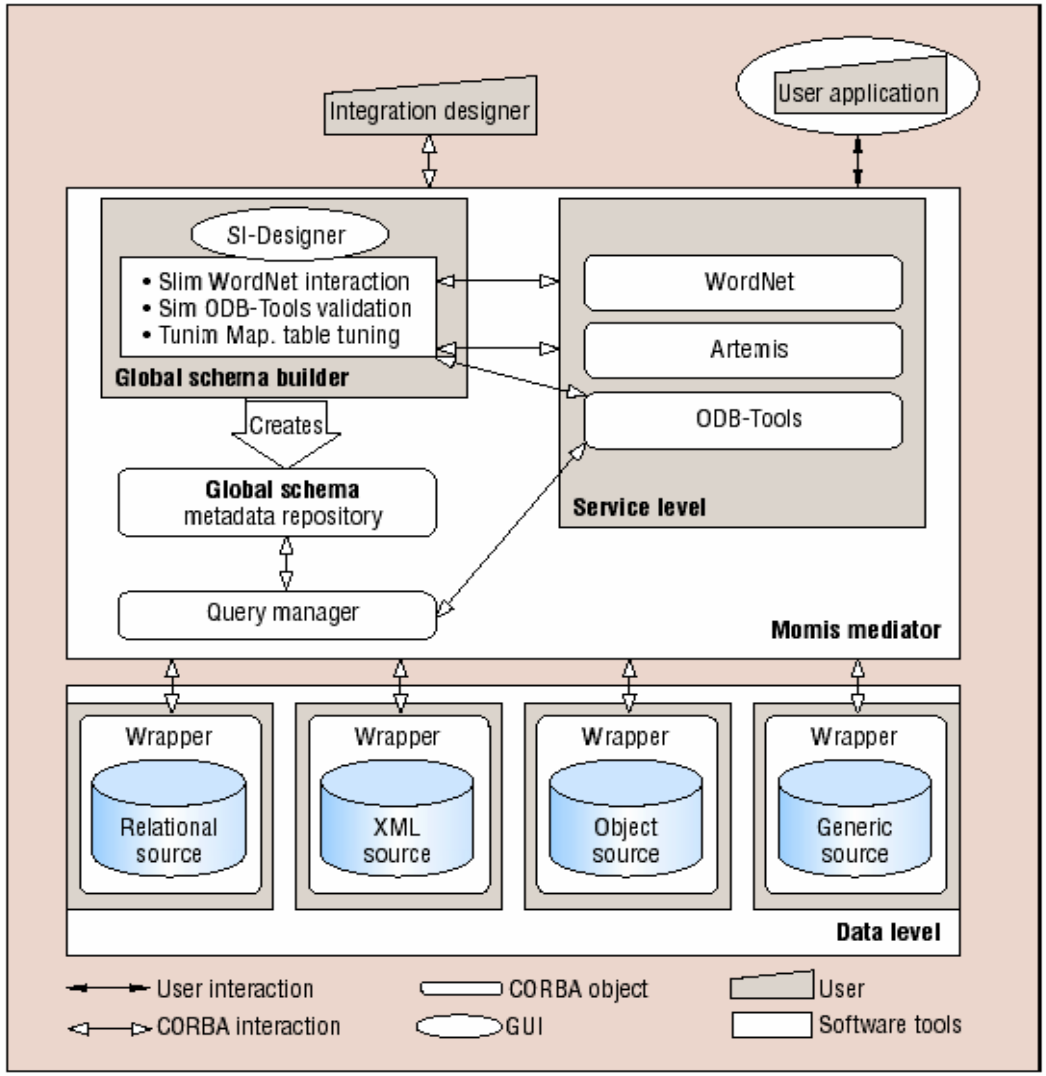
The MOMIS Project: overview

Distributed information stored in multiple, heterogeneous sources



- Sources integration provides a Global Schema (which is a virtual view)
- the Global Schema allows the user to send a query and get a unified answer from all the involved sources (transparently)
- All information in http://www.dbgroup.unimo.it
- INTERDATA (1999-2000); D2I (from Data to Information) (2001-2002) – “Programmi di ricerca scientifica di rilevante interesse nazionale” MIUR

The MOMIS Project



The MOMIS Project: the functional architecture

- Momis relies on a semantic approach based on the conceptual schema - or metadata - of the information sources, and on the I^3 architecture.
- The system consists of the following functional elements that communicate using the Corba standard: a **common data model**, **data wrappers**, and a **mediator**. Momis defines the ODL_{I3} language, according to the common data model, ODM_{I3}, which describes source schemas for integration purposes. Momis treats ODM_{I3} and ODL_{I3} as subsets of the corresponding languages in the ODMG specification— according to the proposal for a standard mediator language developed by the I3-POB working group.
- Momis uses data wrappers, over each source data to translate metadata descriptions of the source into common ODL_{I3} representations. Wrappers then serve to export the resulting data set.
- The global schema builder and the query manager. The global schema builder processes and integrates ODL_{I3} descriptions received from wrappers to derive the information source representations. The query manager module performs query processing and optimization, generates OQL_{I3} subqueries for the sources, and synthesizes a unified global answer for the user.
- Momis creates an integrated view of all sources and performs revision and validation of the various kinds of knowledge used for the integration. To accomplish this, Momis combines the reasoning capabilities of description logics with affinity-based clustering techniques. Momis then exploits a common ontology for the sources constructed using lexical knowledge derived from WordNet and validated integration knowledge.

The MOMIS Project (bibliografy)

MOMIS

- 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

SI-Designer? Ontology Builder

- 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.
- D. Beneventano, S. Bergamaschi, F. Guerra, M. Vincini: “Synthesizing an Integrated Ontology”, IEEE Internet Computing, Vol.7,N.5, September/October 2003.

ARTEMIS

- Silvana Castano, Valeria De Antonellis, Sabrina De Capitani di Vimercati: Global Viewing of Heterogeneous Data Sources. TKDE 13(2): 277-297 (2001)

SEWASIE in a P2P architecture

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

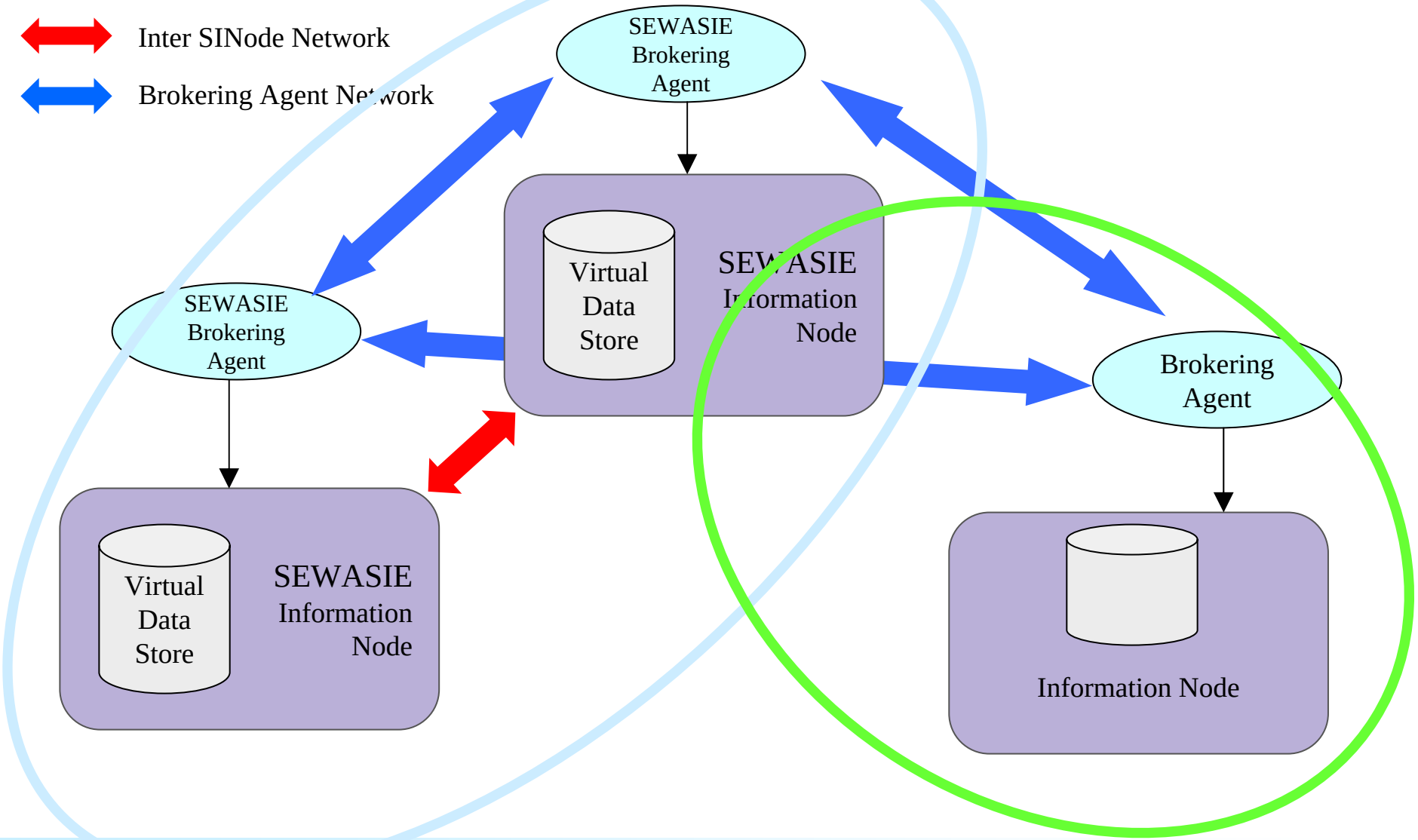
[S. Bergamaschi, F. Guerra, **Peer to Peer Paradigm for a Semantic Search Engine**, in proceedings of the International Workshop on Agents and Peer-to-Peer Computing, LNCS 2530, Springer]

[S. Bergamaschi, F. Guerra, M. Vincini, A peer-to-peer information system for the semantic web, in proceedings of the International Workshop on Agents and Peer-to-Peer Computing, held in AAMAS 2003 International Conference on Autonomous Agents and MultiAgent Systems Melbourne, Australia, July 14, 2003, LNCS 2872]

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 or not 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 dynamically
  - 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