PROGETTO DI UNA UNITÀ DI RICERCA - MODELLO B
Anno 2004 - prot. 2004095494_001

1.1 Tipologia del programma di ricerca

Interuniversitario

Aree scientifico disciplinari

Area 09: Ingegneria industriale e dell'informazione (100%)

1.2 Durata del Programma di Ricerca

24 Mesi

1.3 Coordinatore Scientifico del Programma di Ricerca

BERGAMASCHI ***SONIA*** ***sonia.bergamaschi@unimo.it***

ING-INF/05 - Sistemi di elaborazione delle informazioni

Università degli Studi di MODENA e REGGIO EMILIA

Facoltà di INGEGNERIA

Dipartimento di INGEGNERIA DELL'INFORMAZIONE

1.4 Responsabile Scientifico dell'Unità di Ricerca

BERGAMASCHI ***SONIA***

Professore Ordinario ***01/07/1953*** ***BRGSNO53L41F257K***

ING-INF/05 - Sistemi di elaborazione delle informazioni

Università degli Studi di MODENA e REGGIO EMILIA

Facoltà di INGEGNERIA

Dipartimento di INGEGNERIA DELL'INFORMAZIONE

059 2056132	***059 2056126***	***sonia.bergamaschi@unimo.it***
(Prefisso e telefono)	*(Numero fax)*	*(Email)*

1.5 Curriculum scientifico del Responsabile Scientifico dell'Unità di Ricerca

Testo italiano
Sonia Bergamaschi è nata a Modena ed ha ricevuto la Laurea in Matematica presso la Facoltà di Scienze dell'Università degli Studi di Modena nell'anno 1977. È professore ordinario presso la Facolta` di Ingegneria dell'Università di Modena e Reggio Emilia (sede di Modena) e guida il gruppo di ricerca su Database. La sua attività di ricerca è stata principalmente rivolta alla rappresentazione ed alla gestione della conoscenza nelle Basi di Dati di elevate dimensioni, con particolare attenzione sia agli aspetti teorici e formali

sia a quelli implementativi. Dal 1985 è stata molto attiva nell'area dell'accoppiamento di tecniche di Intelligenza Artificiale, Logiche Descrittive (DL) e Basi di Dati al fine di sviluppare Sistemi di Basi di Dati Intelligenti. Su tali argomenti sono stati ottenuti rilevanti risultati teorici ed è stato sviluppato il sistema ODB-Tools per il controllo di consistenza di schemi e l'ottimizzazione semantica delle interrogazioni.
Recentemente si è occupata di Integrazione Intelligente di Informazioni, proponendo un sistema a mediatore, chiamato MOMIS, per fornire un accesso integrato a sorgenti di informazioni strutturate e semistrutturate che consenta all'utente di formulare una singola interrogazione e di ricevere una risposta unificata. Le tecniche di DL, clustering e ontologie di linguaggio costituiscono la base teorica del sistema.
Dal 2001 e` coordinatore del SIG "Agenti Informativi Intelligenti" nella rete di eccellenza europea AgentLinkII e dal 2002 e` coordinatore del progetto di ricerca europeo SEWASIE che ha come obiettivo lo sviluppo di un motore di ricerca semantico. Ha pubblicato piu` di novanta articoli su riviste e conferenze internazionali e le sue ricerche sono state finanziate da MURST, CNR, ASI e Comunità Europea. È stata membro nel comitato di programma di numerose conferenze nazionali ed internazionali di Basi di Dati e Intelligenza Artificiale. E` membro di IEEE Computer Society e di ACM.

Testo inglese
Sonia Bergamaschi was born in Modena (Italy) and received her Laurea degree in Mathematics from Università di Modena on 1977. She is currently full professor at the Engineering Faculty and leads the "DBGROUP", at the Dipartimento di Ingegneria dell'informazione.
Her research activity has been mainly devoted to knowledge representation and management in the context of very large databases facing both theorical and implementation aspects.
Since 1985 she was very active in the area of coupling artificial intelligence (Description Logics) and database techniques to develop Intelligent Database Systems. On this topic very relevant theoretical results have been obtained and a system ODB-Tools performing consistency check and semantic query optimization In Object Oriented Databases, based on this theoretical results, has been developed.
More recently, her research efforts have been devoted to the Intelligent Information Integration (I3) topic. An I3 system, called MOMIS, to provide an integrated access to structured and semistructured data sources and to allow a user to pose a single query and to receive a single unified answer has been proposed. Description Logics plus clustering techniques constitute the theoretical framework and are exploited for constructing a common ontology, i.e. an integrated view of the information in the separate sources, and for query processing and optimization.
Sonia Bergamaschi is coordinator since 2001 of the Intelligent Information Agents group of the european network of excellence AgentLinkII and since 2002 of the European Research project SEWASIE whose aim is to develop a semantic search engine. She has published about ninety international journal and conference papers and her researches have been founded by the Italian MURST, CNR, ASI institutions and by European Community projects. She has served on the committees of international and national database and AI conferences. She is a member of the IEEE Computer Society and of the ACM.

1.6 Pubblicazioni scientifiche più significative del Responsabile Scientifico dell'Unità di Ricerca

1. *D. BENEVENTANO; BERGAMASCHI S.; C. SARTORI (2003).* ***Description Logics for Semantic Query Optimization in Object-Oriented Database Systems*** *ACM TRANSACTIONS ON DATABASE SYSTEMS. (March 2003).*
2. *D. BENEVENTANO; BERGAMASCHI S.; F. GUERRA; M. VINCINI (2003).* ***Synthesizing an Integrated Ontology*** *IEEE INTERNET COMPUTING. (vol. 7 pp. 42-51)*
3. *BERGAMASCHI S.; I. BENETTI; D. BENEVENTANO; F. GUERRA; M. VINCINI (2002).* ***An Information Integration Framework for E-Commerce*** *IEEE INTELLIGENT SYSTEMS.*
4. *BERGAMASCHI S.; S. CASTANO; D. BENEVENTANO; M. VINCINI (2001).* ***Semantic Integration of Heterogeneous Information Sources*** *DATA & KNOWLEDGE ENGINEERING. (vol. 36 pp. 215-249) Special Issue on Intelligent Information Integration, Elsevier Science B.V.*
5. *BENEVENTANO D; BERGAMASCHI S.; LODI S.; SARTORI C. (1998).* ***Consistency Checking in Complex Object Database Schemata with Integrity Constraints*** *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. (vol. 10 (4) pp. 576-598)*

1.7 Risorse umane impegnabili nel Programma dell'Unità di Ricerca

1.7.1 Personale universitario dell'Università sede dell'Unità di Ricerca

Personale docente

nº	Cognome	Nome	Dipartimento	Qualifica	Settore Disc.	Mesi Uomo	
						1° anno	2° anno
1.	*BERGAMASCHI*	*Sonia*	*Dip. INGEGNERIA*	*Prof. Ordinario*	*ING-INF/05*	*6*	*6*

			DELL'INFORMAZIONE				
2.	*VINCINI*	*Maurizio*	*Dip. INGEGNERIA DELL'INFORMAZIONE*	*Ricercatore Universitario*	*ING-INF/05*	*4*	*4*
3.	*TIBERIO*	*Paolo*	*Dip. INGEGNERIA DELL'INFORMAZIONE*	*Prof. Ordinario*	*ING-INF/05*	*2*	*2*
	TOTALE					**12**	**12**

Altro personale

nº	Cognome	Nome	Dipartimento	Qualifica	Mesi Uomo 1° anno	2° anno
1.	*Guerra*	*Francesco*	*Dip. INGEGNERIA DELL'INFORMAZIONE*	*Contratto*	*4*	*4*
2.	*Miselli*	*Daniele*	*Dip. INGEGNERIA DELL'INFORMAZIONE*	*Contratto*	*4*	*4*
	TOTALE				**8**	**8**

1.7.2 Personale universitario di altre Università

Personale docente

Nessuno

Altro personale

Nessuno

1.7.3 Titolari di assegni di ricerca

Nessuno

1.7.4 Titolari di borse

nº	Cognome	Nome	Dipartimento	Anno di inizio borsa	Durata(in anni)	Tipologia	Mesi Uomo 1° anno	2° anno
1.	*Benassi*	*Roberta*	*Dip. INGEGNERIA DELL'INFORMAZIONE*	*2004*	*3*	*Dottorato*	*6*	*6*
2.	*Martoglia*	*Riccardo*	*Dip. INGEGNERIA DELL'INFORMAZIONE*	*2003*	*3*	*Dottorato*	*6*	*6*
	TOTALE						**12**	**12**

1.7.5 Personale a contratto da destinare a questo specifico programma

nº	Qualifica	Costo previsto	Mesi Uomo 1° anno	2° anno	Note
1.	*Borsista*	*18.000*	*6*	*6*	*Analista programmatore (realizzazione prototipi)*
2.	*Borsista*	*18.000*	*6*	*6*	*Analista programmatore (realizzazione prototipi)*
	TOTALE	**36.000**	**12**	**12**	

1.7.6 Personale extrauniversitario indipendente o dipendente da altri Enti

nº	Cognome	Nome	Nome dell'ente	Qualifica	Mesi Uomo 1° anno	2° anno
1.	*Corni*	*Alberto*	*DEMOCENTER - Modena*	*Analista*	*4*	*4*
2.	*Montanari*	*Daniele*	*ENIDATA*	*Project Manager*	*4*	*4*
	TOTALE				**8**	**8**

2.1 Titolo specifico del programma svolto dall'Unità di Ricerca

Testo italiano
Generazione e mapping di ontologie ed elaborazione di interrogazioni distribuite su siti web

Testo inglese
Ontologies generation and mapping and distributed query processing on web-site

2.2 Settori scientifico-disciplinari interessati dal Programma di Ricerca

ING-INF/05 - Sistemi di elaborazione delle informazioni

2.3 Parole chiave

Testo italiano
ONTOLOGIA DI DOMINIO ; SISTEMI INFORMATIVI SU WEB ; SELEZIONE DI SORGENTI WEB ; INTERROGAZIONE SU ARCHITETTURE DISTRIBUITE ; MAPPING TRA ONTOLOGIE

Testo inglese
DOMAIN ONTOLOGY ; WEB INFORMATION SYSTEMS ; WEB SOURCE SELECTION ; QUERYING ON DISTRIBUTED ARCHITECTURE ; MAPPING AMONG ONTOLOGIES

2.4 Base di partenza scientifica nazionale o internazionale

Testo italiano
La necessità di fornire servizi web efficaci ed efficienti per la ricerca automatica di informazione, estratta dall'enorme disponibilità di dati on-line, ha portato allo sviluppo di una nuova area di ricerca, chiamata Semantic Web (Berners-Lee, 2001). L'obiettivo del Semantic Web e' quello di rendere le pagine Web riconoscibili attraverso procedure automatizzate, con l'introduzione di markup semantici (metadati). Attualmente, gli approcci al Semantic Web consentono l'annotazione semantica di risorse ipotizzando l'esistenza a-priori di ontologie in grado di descrivere il dominio di interesse. Maggiore è l'accuratezza dell'ontologia, maggiore è la precisione dell'annotazione. Il sistema MOMIS (Mediator Environment for Multiple Information Sources) sviluppato dall'unità di Modena si pone l'obiettivo di generare una descrizione sintetica ed integrata delle informazioni provenienti da sorgenti di informazione eterogenee, in modo che l'utente abbia a disposizione una vista globale virtuale (GVV) sulle sorgenti coinvolte senza conoscerne l'effettivo grado di eterogeneità. La GVV rappresenta una concettualizzazione del dominio di interesse, cioè una ontologia di dominio, ottenuta a partire dalle sorgenti stesse. Per ottenere la GVV, MOMIS utilizza le relazioni esistenti tra le informazioni presenti nelle singole sorgenti espresse sia nella forma di relazioni estensionali (ovvero relazioni tra insiemi di istanze dei singoli oggetti rappresentanti i dati) sia nella forma di relazioni intensionali/terminologiche (ovvero relazioni derivanti dal significato attribuito ai termini che descrivono i dati). In particolare, per l'individuazione di relazioni terminologiche, si utilizza l'ontologia di linguaggio Wordnet, un database lessicale sviluppato dalla Princeton University (http://www.cogsci.princeton.edu/~wn/), la cui struttura è ispirata alle teorie psicolinguistiche relative alla memoria lessicale umana. WordNet organizza i nomi, i verbi, gli aggettivi e gli avverbi in insieme di sinonimi (synsets) ognuno dei quali rappresenta un concetto. I synset sono legati fra loro attraverso relazioni di iponimia, iperonimia, meronimia, olonomia.
In un ambiente dinamico come il Web, un ulteriore obiettivo è quello di consentire l'estensione e la modifica di una ontologia: due sono gli approcci proposti, uno basato sull'evoluzione (Motik, 2002), il secondo sul versioning (Klein, 2001). Per approccio basato sull'evoluzione si intende l'adattamento dei concetti di una ontologia alle variazioni del dominio modellato. La modifica di uno o più concetti di una ontologia può generare delle inconsistenze sia all'interno della stessa, sia nelle applicazioni che sono basate su quella ontologia. Il numero elevato di cause che generano un cambiamento e la complessità nella valutazione delle sue conseguenze (occorre considerare infatti i possibili effetti in cascata) rende questo approccio complesso. Una infrastruttura nota che si basa su un approccio evolutivo per la gestione degli aspetti dinamici è KAON (http://kaon.semanticweb.org), sviluppata dall'Università di Karlsruhe. In KAON sono state individuati alcune specifiche di progettazione per gestire più agevolmente l'evoluzione: l'infrastruttura deve essere in grado di risolvere ogni possibile modifica dell'ontologia e di assicurare la consistenza dell'ontologia precedente e di tutte le applicazioni che dipendono da essa; l'infrastruttura deve fornire strumenti che agevolino l'utente nella gestione dei cambiamenti; a fronte della necessità di un cambiamento, l'infrastruttura deve avvisare l'utente della eventuale necessità di raffinamenti successivi.
Con l'approccio versioning si ha la possibilità di gestire i cambiamenti nelle ontologie creando differenti versioni della stessa ontologia. In tal caso occorre riconoscere e distinguere le diverse versioni e definire le procedure per la produzione e il mantenimento delle ontologie.

Un aspetto preliminare all'introduzione di nuove sorgenti in una ontologia è l'individuazione di quelle più utili ad essere integrate. In particolare, uno dei maggiori limiti è l'incapacità di fornire una rappresentazione dei documenti web in cui siano preservati i significati espressi nel testo e in cui siano mantenute le relazioni semantiche tra di essi. L'unità di Modena ha proposto TUCUXI (Benassi, 2004), un sistema che sfrutta la teorizzazione linguistica delle proprietà di coesione e di coerenza fondamentali per i testi scritti. Più precisamente in (Halliday 1976) vengono descritti alcuni aspetti del linguaggio: vi è un'importante differenza fra un

insieme di frasi ed un testo scritto. Infatti, un insieme di frasi diventa testo quando un lettore, considerando i termini utilizzati nella loro successione, ne comprende il significato e si rende conto di come siano più o meno dipendenti (dipendenza intrafrasale e interfrasale). Ogni lingua ha un insieme di strumenti per far sì che i concetti espressi risultino coesi e coerenti. In particolare, (Hoey, 1991) ha dimostrato che lo strumento che maggiormente contribuisce alla coesione di un testo è la coesione lessicale, che si ottiene tramite reiterazione di un concetto (uso di termini sinonimi o più generici/specifici) oppure tramite collocation (uso di termini specifici per lo stesso dominio lessicale). (Halliday, 1976) introduce l'idea di costruire gruppi di parole (catene lessicali) fra loro semanticamente correlate a partire dal testo scritto. Le catene lessicali rappresentano pertanto una sintesi del testo in cui la semantica è preservata. In questo senso, il contributo fondamentale di (Morris, 1991) è l'introduzione di un'ontologia lessicale come base di conoscenza tramite la quale ricavare le relazioni semantiche fra termini. WordNet rappresenta uno degli strumenti maggiormente utilizzate per la costruzione delle catene lessicali, le quali possono essere considerate come cluster di synset (Barzilay, 1997)(Silber, 2002) (Galley, 2003).

In applicazioni nelle quali il numero di ontologie cresce a dismisura, poter disporre di tecniche ed algoritmi che permettano mapping tra di esse diviene un fattore cruciale per una soluzione di Semantic Web. In letteratura sono presenti differenti criteri attraverso i quali classificare gli strumenti software che hanno come obiettivo la creazione delle relazioni di mapping fra gli schemi (Rahm, 2001). In particolare, per quanto riguarda il trattamento dell'eterogeneità delle sorgenti, i sistemi per la generazione di match possono essere suddivisi in due categorie. Nella prima categoria si collocano i sistemi in cui le relazioni di mapping sono definite e calcolate direttamente tra uno schema sorgente e uno schema destinazione; appartengono a tale categoria i sistemi CLIO (Miller, 2001), GLUE (Doan, 2002) e PROMPT (Noy, 2000). Nei sistemi dell'altra categoria gli schemi coinvolti vengono rappresentati in un modello intermedio comune ed i mapping generati sono riferiti a tali rappresentazioni; in questa categoria ci sono sistemi quali DIKE (Palopoli, 2003), LSD (Doan, 2000) e MOMIS (Beneventano, 2003).

In una architettura complessa di sorgenti eterogenee e/o distribuite ed ontologie distribuite la gestione e l'elaborazione, anche approssimata, di query, con particolare riferimento a query che esprimono condizioni non solo sui dati ma anche sulla componente strutturale, risulta un interessante ed attuale tema di ricerca. In tale contesto i mapping semantici tra le ontologie rivestono un ruolo fondamentale in quanto costituiscono la base di partenza sia per poter riscrivere una data query originariamente espressa con riferimento a una ontologia locale, sia per quantificare le similarità tra i diversi concetti descritti nelle ontologie distribuite. Tra i sistemi più interessanti in grado di essere applicati a schemi o grafi citiamo COMA (Do et al., 2002), che supporta la combinazione di diverse tecniche di schema matching, Cupid (Madhavan et al., 2001), che combina algoritmi di matching per nomi e strutture, e Similarity Flooding (SF) (Garcia-Molina et al., 2002), che fornisce un algoritmo per il matching tra grafi di grande versatilità. Un altro importante problema da affrontare nella elaborazione di interrogazioni su sorgenti eterogenee e/o distribuite è quello della object fusion, ovvero quello di raggruppare assieme informazioni relative allo stesso oggetto del mondo reale memorizzate nelle differenti sorgenti. La fusione di dati provenienti da differenti sorgenti richiede che le differenti rappresentazioni dello stesso oggetto siano identificate; tale processo è chiamato object identification (Naumann 2002). Inoltre, dopo aver risolto il problema della object identification, si deve effettuare la fusione di dati provenienti dalle differenti sorgenti considerando problemi di inconsistenza tra le sorgenti (Bertossi, 2003) (Greco, 2003) (Naumann 2002) (Lin, 1998). Infine, allo scopo di sintetizzare in un unico risultato tutte le informazioni relative allo stesso oggetto, provenienti dalle differenti sorgenti, un operatore proposto in lettaratura, che sembra essere particolarmente promettente, è quello di full-disjunction (C. Galindo-Legaria) (1994, Ullman 1996).

Testo inglese

The need to provide effective and efficient on-line services, especially concerning information search, has contributed to refine already existing Information Retrieval (IR) techniques (Baeza-Yates, 1999) and study new tools for the semantic search (Heflin, 2000). Moreover, to fully and automatically exploit the enormous availability of on-line data, a new vision of the Web, called Semantic Web, arises (Berners-Lee, 2001). The goal of the Semantic Web is to make Web pages processable with automatic procedures, by means of the introduction of semantic markups (metadata).
Now, Semantic Web approaches allow to semantically annotate data sources by assuming the existence of a-priori ontologies that describe the interest domain. The more accurate is the ontology, the better is the annotation accuracy. The MOMIS (Mediator Environment for Multiple Information Sources) system, developed by the Modena unit, aims at generating a synthetic and integrated description of information coming from heterogeneous data sources, so the user has at his disposal a Global Virtual View (GVV) on the involved sources without knowing exactly the heterogeneity data degree. The GVV represents a conceptualization of the domain, i.e. an ontology of the domain of interest created from the involved local sources. To obtain the GVV, MOMIS exploits the relations between information in the individual data sources, expressed in the form of extensional relationships (i.e. relationships between set of instances of single objects describing data) and in the form of intensional/lexical relationships (i.e. relations coming from the meaning assigned to terms that describe data). In particular, to identify lexical relations the WordNet lexicon ontology is exploited. WordNet is an electronic lexical database which is considered as the most important resource available, for computational linguistic, text analysis and other associated fields. WordNet, developed by the Congnitive Science laboratory at the Princeton University (http://www.cogsci.princeton.edu/~wn/), has a structured inspired to the current psycholinguistics theories about human lexical memory. it organizes nouns, verbs, adjectives and adverbs into synonims sets (synsets), each of them represents an underlying concept. Synsets are connected each other by means of hyponymy, hypernymy, meronymy, holonymy.

In a dynamic environment as the Web is, a further goal is to extend a previously created GVV by introducing new data sources. In literature, two main methods for ontology evolution exists: the first one is based on the evolution (Motik, 2002), the second on the versioning (Klein, 2001). The approach based on the evolution means that the ontology is accurately adapted to the changes. The most famous evolution-based infrastructure for ontology dynamics KAON (http://kaon.semanticweb.org), developed at the University of Karlsruhe. In particular, during the KAON design, some features to simplify the evolution problem management has been identified: the infrastructure must be able to solve every possible changes in the ontology and ensure the consistency of the previous ontology and of all the applications that depend on it; the infrastructure must make available tools to ease the user during the changes management; to face the need of a change, the infrastructure most inform the user about possible refinements.
The versioning approach concerns the ability to manage changes in the ontologies, by creating and defining different versions of the same ontology. To obtain the ability to manage several copies of the same ontology, it is necessary to develop a method for recognizing and distinguishing between different ontology versions and procedures to produce and maintain the ontologies.

A preliminary problem related to the introduction of new data sources is the identification of the ones that are suitable to be

integrated. Nowadays, keyword-based search engines suffer from the information overload problem. In particular, one of the greater limitation of current search engines is the incapacity for providing a documents representation where meanings of the texts are preserved and semantic relations between meanings are maintained. The unit of Modena has proposed TUCUXI (Benassi, 2004), a system that exploits the linguistic theories about coherence and cohesion properties of written texts. More precisely, in (Halliday 1976) some aspects about human languages are described: there is a strong difference between a random set of sentences and a written text. In fact, a set of sentences becomes a texts only when a reader, considering the sequence of terms, understands the underlying meaning and recognizes that they are each other dependent (intra and intrasentences dependence). Each language has a set of possibilities to achieve cohesion and coherence between the expressed concepts. In particular, (Hoey, 1991) showed that cohesion in text is mainly achieved through lexical cohesion, that can be obtained through reiteration of a concept (use of synonyms, broader/narrower terms), or through collocation (use of terms that tend to co-occur in similar lexical context) (Halliday, 1976) introduces the idea that, starting from the written text, it is possible to construct group of words (lexical chains) that are sematically related each-others. Thus, lexical chains represent a text synthesis where semantics is preserved. The main contribution in (Morris, 1991) is the introduction of a lexical ontology as a knowledge base from which semantic relations between terms can be extracted. WordNet, is one of most important tool for constructing lexical chains that can be seen as clusters of synsets (Barzilay, 1997)(Silber, 2002)(Galley, 2003).

Within applications, when the number of ontologies grows out of all proportion, having at one's disposal techniques and algorithms to obtain mappings amongst them is a crucial aspect for obtaining a Semantic Web solution. In literature, several criteria have been described to classify software tools that aim at creating mapping relations amongst schemas (Rahm, 2001). In particular, the systems proposed for the generation of matchings can be subdivided in two categories: 1) the mapping relations are calculated between a source schema and a destination schema (CLIO – Miller, 2001, GLUE – Doan, 2002, PROMPT – Noy, 2000); 2) the involved schemata are translated in an intermediate common model and the generated mappings about involved sources are referred to thier representations in the common model (MOMIS – Beneventano, 2003, LSD Doan, 2000, DIKE – Palopoli, 2003).

In a complex architecture of heterogeneous/distributed ontologies, approximate query evaluation is a challenging research topic. In this context, semantic mappings among ontologies play a fundamental role as they constitute the starting point both for being able to rewrite a query, originally expressed with reference to a local ontology, and (to estimate the similarity of concepts expressed in the distributed ontologies) for performing approximate query evaluation. Among the most interesting systems that can be applied to the problem of schema/graphs matching we cite COMA (Do et al., 2002), which supports the combination of different schema matching techniques, Cupid (Madhavan et al., 2001), which combines matching algorithms based on name and structural matching and Similarity Flooding (SF) (Garcia-Molina et al., 2002), which exploits a particularly versatile graph matching algorithm.

Another relevant problem to be faced in query processing over heterogeneous/ distributed sources is that of "object fusion", that is grouping together information related to the same object stored in different sources. The fusion of data coming from different sources requires that the different istantiations of the same object be identified: this process is called "object identification" (Naumann 2002). Moreover, after having solved the object identification problem, we have to perform the fusion of data coming from different sources taking into account the problem of inconsistent information among sources (Bertossi, 2003, Greco, 2003, Naumann 2002, Lin, 1998). An operator proposed in the literature which seems promising in order to synthetize in a unique result all the information related to the same object, coming from different sources, is "full-disjunction" (C. Galindo-Legaria, 1994, Ullman 1996).

2.4.a Riferimenti bibliografici

(Baeza-Yates, 1999) Ricardo A. Baeza-Yates, Berthier A. Ribeiro-Neto: Modern Information Retrieval. ACM-Press/Addison-Wesley. (1999)

(Barzilay, 1997) Regina Barzilay and Michael Elhadad. Using Lexical Chains for Text Summarization, in ACL Workshop on Intelligent Scalable Text Summarization, July 1997.

(Bertossi, 2003) L. E. Bertossi, J. Chomicki: Query Answering in Inconsistent Databases. Logics for Emerging Applications of Databases 2003: pagine 43-83, Springer 2003.

(Budanitsky, 2001) Budanitsky, A., Hirst, G.: Semantic distance in wordnet: An experimental application-oriented evaluation of five measures. In: Workshop on WordNet and Other Lexical Resources. NAACL 2001. (2001)

(Benassi, 2004) Roberta Benassi, Sonia Bergamaschi, Maurizio Vincini, Web Semantic Search with TUCUXI, Submitted paper to SEBD 2004, 2004.

(Beneventano, 2003) Domenico Beneventano, Sonia Bergamaschi, Francesco Guerra, Maurizio Vincini, Synthesizing an Integrated Ontology, IEEE Internet Computing Magazine, September-October 2003, 42-51.

(Bergamaschi, 2001) BERGAMASCHI S.; S. CASTANO; D. BENEVENTANO; M. VINCINI Semantic Integration of Heterogeneous Information Sources DATA & KNOWLEDGE ENGINEERING. Special Issue on Intelligent Information Integration, Elsevier Science B.V. (vol. 36 pp. 215-249) (2001).

(Berners-Lee, 2001) Tim Berners-Lee, James Hendler, Ora Lassila. The Semantic Web, Scientific American, May, 2001

(Do et al., 2002) H. Do, E. Rahm: COMA - A system for flexible combination of schema matching approaches. In Proc. of the 28th VLDB, 610-621, 2002.

(Doan, 2000) A. Doan, P. Domingos, A. Halevy. Learning Source Description for Data Integration, Proceedings of the Third International Workshop on the Web and Databases, WebDB 2000, Dallas, Texas, USA.

(Doan, 2002) A. Doan, J. Madhavan, P. Domingos, A. Halevy. Learning to map between ontologies on the semantic web, Proceedings of the 11th International WWW Conference, 2002.

(Galindo-Legaria, 1994) C. Galindo-Legaria. Outerjoins as disjunctions. Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data, Minneapolis, Minnesota, USA, 1994, 348-358.

(Galley, 2003) Michel Galley, Kathleen McKeown. Improving Word Sense Disambiguation in Lexical Chaining. In the proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03). August 2003. Acapulco, Mexico.

(Garcia-Molina et al., 2002) Garcia-Molina, S. Melnik, E. Rahm: Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching. In Proc. of the 18th ICDE, 2002.

(Greco, 2003) G. Greco, S. Greco, E. Zumpano: A Logical Framework for Querying and Repairing Inconsistent Databases. IEEE Trans. Knowl. Data Eng. 15(6): 1389-1408 (2003)

(Halliday, 1976) M.A.K. Halliday, R. Hasan: Cohesion in English, Longman 1976

(Heflin 2000) Jeff Heflin, James Hendler. Searching the Web with SHOE, In AAAI-2000 Workshop on AI for Web Search. 2000.

(Hirst, 1998) G. Hirst, D. St-Onge: Lexical chains as representations of context for the detection and correction of malapropisms WordNet: An electronic lexical database, Christiane Fellbaum(editor), Cambridge, MA: The MIT Press, 1998.

(Hoey, 1991) M. Hoey: Patterns of Lexis in Text. Oxford University Press, Oxford, 1991.

(Klein, 2001) M. Klein and D. Fensel,Ontology Versioning on the Semantic Web, First Int'l Semantic Web Working Symp., Stanford University Press, 2001, pp. 75–91.

(Lin, 1998) J. Lin, A. O. Mendelzon: Merging Databases Under Constraints. Int. J. Cooperative Inf. Syst. 7(1): 55-76 (1998)

(Madhavan et al., 2001) J. Madhavan, P. A. Bernstein, E. Rahm: Generic Schema Matching with Cupid. In Proc. of the 27th VLDB, 49-58, 2001

(Miller, 1995) George A. Miller: WordNet: A Lexical Database for English. Commun. ACM 38(11): 1995, 39-41

(Miller, 2001) R. J. Miller, M. A. Hernández, L. M. Haasand, L. Yan, C. T. H. Ho, R. Fagin, L. Popa: The Clio Project: Managing Heterogeneity. SIGMOD Record, Vol 30(1): 2001, 78-83

(Morris, 1991) Jane Morris, Graeme Hirst: Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. Computational Linguistics 17(1): 21-48 (1991)

(Motik, 2002) B. Motik at al: User-driven Ontology Evolution Management, Proc. 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW 02), LNCS 2473, Springer, 2002, 285-300.

(Naumann 2002) F. Naumann, M. Haussler: Declarative Data Merging with Conflict Resolution. International Conference on Information Quality (IQ 2002), 212-224.

(Noy, 2000) N. F. Noy, M.A. Musen: PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment, Seventeenth National Conference on Artificial Intelligence (AAAI-2000), Austin, Texas, USA, 2000.

(Okamura, 1994) M. Okamura, T. Honda Word sense disambiguation and text segmentation based on lexical cohesion,Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-94), volume 2, 755-761, 1994.

(Palopoli, 2003) L. Palopoli, G. Terracina, D. Ursino Experiences using DIKE, a system for supporting cooperative information system and data warehouse design, IEEE Transaction on Knowledge and Data Engineering, vol. 15(2), pp 835-865, 2003.

(Rahm, 2001) E. Rahm, P.A. Bernstein A survey of approaches to automatic schema matching, VLDB Journal, vol. 10(4), pp 334-350, 2001.

(Silber, 2002) H. Gregory Silber, Kathleen F. McCoy: Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization. Computational Linguistics 28(4): 487-496 (2002)

(Ullman, 1996) J. D. Ullman A. Rajaraman,: Integrating Information by Outerjoins and Full Disjunctions. PODS-1996, 238-248.

2.5 Descrizione del programma e dei compiti dell'Unità di Ricerca

Testo italiano
L'Unita' di Modena lavorerà a tutti e tre i temi del progetto. All'interno del TEMA1 si occupera' delle problematiche connesse alla creazione ed estensione di una ontologia di dominio;
Nell'ambito del TEMA 2 collaborera' alla definizione di una architettura di riferimento per la scoperta e la gestione di mapping semantici tra ontologie; relativamente al TEMA 3 studierà e svilupperà, in collaborazione con l'unità di Bologna, la traduzione (riscrittura) automatica di query formulate su una data ontologia rispetto alle altre ontologie; studierà e svilupperà tecniche per la definizione di condizioni di join tra le sorgenti locali per definire coppie di classi che fanno riferimento allo stesso oggetto del mondo reale.
Infine, congiuntamente a tutte le altre Unità si lavorerà alla realizzazione dei prodotti comuni.
In particolare, nella prima fase l'attività congiunta ha come obiettivo la definizione dell'architettura metodologica e funzionale di riferimento per l'intero progetto (prodotto D0.R1).
D0.R1 Rapporto sull'architettura metodologica e funzionale di riferimento (BO, MO, RM, TN)
Durante la seconda fase si procederà alla definizione delle interfacce dei componenti per il prototipo integrato (prodotto D0.R2).
D0.R2 Specifiche delle interfacce dei componenti del prototipo integrato (BO, MO, RM, TN)
Infine nella terza fase del progetto si collaborerà con le altre Unità all'integrazione
dei prototipi realizzati durante il progetto.
D0.P1 Prototipo integrato di sistema (BO, MO, RM, TN)

L'attività specifica di ricerca dell'unità di Modena prevede l'articolazione seguente.

TEMA 1
1.1 – Definizione di un linguaggio di Ontologia con aspetti/concetti estensionali
Il linguaggio di ontologia che l'unità di Modena contribuirà a definire si baserà sul linguaggio ODLI3, sviluppato in precedenza dall'unità nell'ambito del sistema MOMIS, reso compatibile con gli standard W3C. L'unità si concentrerà sul problema di rendere tale linguaggio sufficientemente espressivo per poter esprimere: mapping fra ontologie eterogenee indipendentemente sviluppate in modo da facilitare il compito di riscrittura di query; concetti estensionali in modo da facilitare il compito di reperimento di sorgenti utili all'esecuzione di una query.

1.2 – Aggiunta di una nuova sorgente informativa ad una Ontologia di Dominio
A partire dal sistema MOMIS, l'unità di Modena studierà il problema dell'evoluzione della GVV e dell'ontologia di riferimento dovuta all'integrazione di una nuova sorgente informativa.
Infatti, una modifica in uno o più concetti dell'ontologia può causare diverse inconsistenze sia in concetti collegati, sia in altre ontologie che sono collegate alla prima attraverso i mapping. L'approccio che si perseguirà mira ad integrare la descrizione di una nuova sorgente informativa all'interno di una ontologia esistente utilizzando un processo semi-automatico, basato sul lessico, che calcolerà le affinità tra elementi della descrizione da inserire e l'ontologia esistente. Un elemento verrà aggiunto all'ontologia solo nel caso in cui non siano presenti elementi affini. L'ontologia dovrebbe crescere in modo monotono, minimizzando le modifiche all'esistente, evitando le inconsistenze interne e riducendo il rischio di inconsistenze dovute ai mapping con altre ontologie (si rendono espliciti nuovi elementi dell'ontologia, mentre quelli preesistenti mantengono lo stesso significato).

1.3 – Individuazione di una nuova sorgente informativa relativa ad un'Ontologia di Dominio
L'unità di Modena collaborerà allo studio e allo sviluppo di strumenti di natura semantica in grado di migliorare l'efficacia delle tecniche attualmente utilizzate dai motori di ricerca keyword-based, come ad esempio Google. La ricerca di nuove sorgenti Web sarà coadiuvata da tecniche di comprensione del linguaggio naturale, alcune delle quali sono già implementate in TUCUXI (Benassi, 2004). Lo scopo è quello di ottenere una rappresentazione sintetica dei significati espressi in un testo e di mantenere le relazioni semantiche fra termini. L'idoneità della sorgente sarà valutata grazie ad una misura di similarità semantica fra ontologia e catene lessicali che verrà sviluppata nell'ambito del TEMA 2.

FASE 1
Si valuteranno, in collaborazione con tutte le altre unità, in modo critico le proposte di standard e linguaggi emergenti per la definizione ed il trattamento delle ontologie con particolare riferimento al problema dell'evoluzione delle ontologie.

PRODOTTI
D1.R1: Analisi Critica dei linguaggi e standard emergenti per le ontologie (BO,MO,ROMA,TN)

FASE 2
Si definirà un linguaggio per la definizione ed il trattamento delle ontologie con particolare riferimento alla descrizione dell'evoluzione delle ontologie. L'attività sarà svolta in collaborazione con le unità di BO e TN e verrà sviluppato un prototipo per l'aggiunta di una nuova sorgente informativa alla Ontologia di dominio.
Inoltre si produrrà un'analisi critica delle tecniche esistenti per l'estrazione delle catene lessicali.

PRODOTTI
D1.R2: Definizione del linguaggio per la specifica di una ontologia di dominio (BO, MO, TN)
D1.R4: Analisi critica delle tecniche esistenti per l'estrazione delle catene lessicali (MO)
D1.P1: Prototipo per l'aggiunta di una nuova sorgente informativa alla Ontologia di Dominio (MO)

FASE 3
Si studierà l'implementazione di nuovi algoritmi per l'estrazione delle catene lessicali. A partire dall'analisi critica delle tecniche esistenti, si individueranno le caratteristiche che i nuovi algoritmi dovranno presentare, in relazione a tre aspetti ritenuti di

maggiore interesse: il primo riguarda il tipo di documenti trattati nell'ambito del progetto (pagine Web); il secondo è inerente alla complessità computazionale ed in particolare si definiranno algoritmi di complessità lineare; il terzo è relativo all' accuratezza che la rappresentazione sintetica e semantica dei documenti dovrà garantire. A tale riguardo, si porrà particolare attenzione alla disambiguazione dei termini (word sense disambiguation) come fase preliminare alla costruzione di catene lessicali correttamente rappresentative delle sorgenti analizzate. Il prototipo realizzato estenderà in maniera significativa le funzionalità di TUCUXI (Benassi, 2004) e completerà RoadRunner (sviluppato dall'unità di Roma), migliorandone la capacità di assegnare semantica alle informazioni estratte da siti data intensive.
PRODOTTI
D1.R6 Definizione di tecniche per associare semantica allo schema di un sito data-intensive basate su catene lessicali (RM, MO)
D1.P2 Prototipo per l'estrazione di catene lessicali da siti web (MO)

TEMA 2
L'attività sarà rivolta alla definizione di linguaggi, tecniche ed algoritmi che esprimano mapping tra diverse ontologie. All'interno del progetto saranno definiti algoritmi di matching fra le sorgenti coinvolte che tengano conto delle tematiche dettagliate nel seguito.
I sistemi che usano vincoli derivati dal lessico cercano di sfruttare in via prioritaria i nomi degli elementi degli schemi per trovare elementi simili. La similarità dei nomi degli elementi degli schemi può essere individuata in differenti modi tra i quali ricordiamo: l'uguaglianza dei nomi, l'uguaglianza del nome canonico che si ottiene dopo un'operazione di pre-processing, l'uguaglianza degli ipernimi, l'uguaglianza di nomi sulla base di indicazione fornita dall'utente. È necessario osservare che non sempre l'uguaglianza tra due nomi (o tra due ipernimi) si può ricondurre a un mero confronto meccanico fra stringhe. A nomi possono corrispondere significati differenti (polisemia) e viceversa nomi differenti possono avere lo stesso significato (sinonimia). Per generare tali tipi di relazioni lessicali è quindi fondamentale fare riferimento a ontologie lessicali che cataloghino i vocaboli sulla base del significato e in questo modo possano effettuare i corretti raffronti.
Si può osservare che in alcune circostanze i nomi associati agli elementi degli schemi di alcune basi di dati non sono semanticamente rilevanti. In questi casi è opportuno utilizzare delle tecniche ausiliare per estrarre semantica dall'analisi dei dati e/o tecniche che consentono di analizzare e sfruttare i commenti espressi in linguaggio naturale dal progettista della sorgente. Altri vincoli possono essere ricavati dall'analisi di schemi: ad esempio l'uguaglianza può essere ricavata basandosi sull'equivalenza dei tipi di dato e sul dominio delle chiavi, sulla cardinalità delle relazioni e sulle relazioni is-a. Altre relazioni possono essere ricavate da una analisi specifica basata sul modello logico dei dati. Infine, sulla base dell'estrazione della rappresentazione sintetica delle sorgenti tramite la tecnica delle catene lessicali, l'unità di Modena studierà e svilupperà misure di similarità semantica per quantificare l'attinenza delle sorgenti rispetto all'ontologia di dominio di riferimento come evoluzione di quelle presentate in (Budanitsky, 2001). Tali misure dovranno considerare due diversi aspetti: il primo relativo al matching "esatto" di significati/concetti, il secondo inerente alla similarità e alle relazioni semantiche fra concetti.
FASE 1
Verranno analizzati in modo critico i principali algoritmi di matching presenti in letteratura, considerando con attenzione quelli che prevedono tecniche per la risoluzione dei conflitti tra le varie rappresentazioni e le proposte di standard per il mapping tra ontologie.
PRODOTTI
D2.R1: Analisi Critica di linguaggi e tecniche di mapping (MO, TN)
FASE 2
Si concorrerà alla definizione di un linguaggio per il mapping e di un algoritmo di matching che proponga in modo automatico mapping basati sulla similarità semantica.
PRODOTTI
D2.R2: Definizione del linguaggio per la specifica di mapping semantici (MO, TN)
D2.R3: Valutazione empirica di misure di similarità semantica (MO)
FASE 3
L'attività svolta durante la fase 3 sarà rivolta all'estrazione della rappresentazione sintetica delle sorgenti tramite la tecnica delle catene lessicali. In particolare, verranno valutati gli algoritmi di cui al D1.R6 secondo vari parametri, tra i quali alcuni di natura tecnologica (robustezza del processo estrattivo, complessità computazionale,…) altri di natura qualitativa (espressività delle catene lessicali come metodologia descrittiva delle sorgenti, possibilità di estensione in ambito multilinguistico, efficacia delle tecniche proposte in relazione all'assegnazione di semantica ai dati estratti tramite RoadRunner). Verranno studiate e svilupperemo misure di similarità semantica fra le catene lessicali e l'ontologia di riferimento stessa. Tali misure dovranno considerare due diversi aspetti: il primo relativo al matching "esatto" di significati/concetti, il secondo inerente alla semantic relatedness fra concetti (Budanitsky, 2001). Ad esempio, le misure di similarità semantica dovranno considerare diversamente il caso in cui l'ontologia di riferimento e la sorgente, tramite le catene lessicali, condividono il concetto di libro ed il caso in cui la seconda non contenga il concetto di libro ma esprima il concetto di volume.
PRODOTTI
D2.P1 Prototipo della piattaforma per la generazione/gestione automatica di mapping tra ontologie di domino eterogenee (MO, TN)

TEMA 3
Un primo obiettivo specifico è quello di mettere a punto tecniche e strumenti per la traduzione (riscrittura) automatica di una query, formulata con riferimento a una ontologia di dominio locale, in forme che siano adeguate anche rispetto alle altre ontologie disponibili nell'ambiente distribuito. Tale processo è chiaramente necessario nell'ottica di rispondere nel modo più efficace e completo possibile alle query poste dagli utenti, sfruttando così appieno le potenzialità delle informazioni messe a disposizione dalle sorgenti dati. Non è infatti plausibile pensare che tutte le informazioni utili a soddisfare il fabbisogno informativo dell'utente che esegue un'interrogazione provengano dalla sorgente rispetto alla quale la query è stata formulata; piuttosto, occorre cercare di sfruttare tutte le sorgenti utili, interrogando quindi anche quelle che sono integrate in ontologie diverse da quella su cui è posta la query originaria. Lo scopo è quindi quello di ottenere delle tecniche che, sfruttando le informazioni sulla semantica dei singoli concetti descritti nelle ontologie di riferimento e il contesto in cui sono inseriti, riscrivano la query verso le altre ontologie, in una forma che sia il più possibile equivalente a quella originaria.
Un secondo obiettivo è relativo allo studio di tecniche per la creazione della Istanza Globale della GVV. Tale istanza, viene calcolata sulla base delle estensioni delle sorgenti locali, unicamente in fase di risoluzione delle interrogazioni e sulla base dei

seguenti elementi: il mapping tra la GVV e le sorgenti locali, l'identificazione degli oggetti delle sorgenti locali che rappresentano gli stessi oggetti del mondo reale (Join Map) e l'operazione di full-disjunction che permette di sintetizzare un unico risultato per ogni oggetto instanziato in piu' sorgenti. L'attività di ricerca sarà rivolta allo studio di tecniche per la definizione semiautomatica delle Join Map e all'estensione dell'operazione di full-disjunction. In particolare, verranno sviluppate soluzioni che sono valide sia sotto l'ipotesi di "omogeneità semantica", cioè valori uguali per attributi locali comuni in differenti sorgenti e relativi allo stesso oggetto reale, che nel caso generale.

FASE 1
La prima fase prevedrà un'analisi critica delle tecniche di riscrittura delle query basate su ontologie (deliverable D3.R1).
PRODOTTI
D3.R1: Analisi critica di linguaggi di interrogazione e tecniche di riscrittura basati su ontologie (BO, MO, TN)

FASE 2
Verrà scelto l'approccio per la riscrittura di query, unitamente a una serie di tecniche propedeutiche ad una riscrittura il più possibile efficace. In particolare, i mapping semantici tra le ontologie costituiranno una importante base di partenza per poter riscrivere una data query originariamente espressa con riferimento a una ontologia locale. L'idea alla base sarà quella di partire da tali mapping per quantificare le similarità tra i diversi concetti descritti nelle ontologie di riferimento. Le tecniche per valutare la similarità tra i concetti coinvolti non sono proprie della fase di riscrittura in sé ma ne costituiscono una fondamentale fase preliminare e propedeutica. Tali tecniche dovranno essere studiate in modo che non si limitino all'utilizzo delle informazioni semantiche legate al significato dei vari concetti presenti nelle ontologie, ma che tengano anche conto del contesto (struttura) in cui tali concetti sono inseriti, ispirandosi ad altri approcci presentati recentemente (Garcia-Molina et al., 2002). Le similarità individuate mediante tali approcci verranno quindi sfruttate per la fase di riscrittura verso le altre ontologie, che sarà in grado tanto di riadattare la struttura della query quanto di riscriverne in modo consistente i valori. Tali similarità verranno inoltre utilizzate per valutare e quantificare la verosimiglianza tra le query ottenute tramite riscrittura e quella originaria. Per ottenere una risposta completa e minimale ad una query verrà applicato ed esteso il metodo della full-disjunction.

PRODOTTI
D3.R3: Definizione del linguaggio di interrogazione e delle tecniche di riscrittura basate su ontologie (BO, MO, TN)

FASE 3
La FASE 3 vedrà l'effettiva implementazione delle tecniche di riscrittura proposte all'interno un prototipo per la formulazione di interrogazioni.
PRODOTTI
D3.P1 Prototipo per la formulazione di interrogazioni (BO, MO)

Testo inglese
The unit of Modena will work on all the three project themes; within THEME 1 it will deal with issues concerning the creation and the extension of a domain ontology; within THEME 2 it will co-operate to the definition of a reference architecture for discovering and managing semantic mappings among ontologies; concerning THEME 3, it will partecipate to the automatic translation (rewriting) of queries expressed on a given ontology into an appropriate form for other different ontologies, and to the study of techniques for computing a unique result concerning the same object instantiated in multiple sources.
We will work, together with all the other Units, on the definition of common products. in the first phase, this activity aims at defining a methodological and functional reference architecture for the whole project (product D0.R1).

D0.R1 Report on the methodological and functional reference architecture (BO, MO, RM, TN)

In the second phase, we will work, together with all the other Units, on the components interfaces definition of the integrated prototype (product D0.R2)

D0.R2 Specifications of the components interfaces of the integrated prototype (BO, MO, RM, TN)

Finally, in the third phase of the project, we will work together with all the other Units on the integration of the prototypes developed during the project.

THEME 1
Concerning THEME 1, the research activity will focus on the following topics:
1.1 – Definition of an ontology language making up extensional aspects/concepts

The ontology language that the unit of Modena will contribute to define will be based on the ODLI3 language, made compatible with the W3C standards. In addition, the unit will focus on making the language expressive enough to represent mappings between heterogeneous independently developed ontologies and to ease the query rewriting.

1.2 - Adding a new information source to the domain ontology

Starting from the MOMIS system, the unit of Modena will study the aspect concerning the GVV and reference ontology evolution due to the integration of a new information source. In fact, a change in one or more concepts within the ontology can cause several inconsistencies both on related concepts in the same ontology and on the ontologies connected to the first one by means of mappings. The approach aims at integrating the description of a new information source into a pre-existing ontology, by exploiting a semi-automatic lexicon-based process that calculates the affinity among the description elements to be inserted and the ontology. An element will be added to the ontology only if there are no similar elements within it. The ontology should grow monotonously, minimizing the changes to the existing ontology and avoiding internal inconsistencies. In addition, the risk of propagating

inconsistencies to mappings is reduced.

1.3 – Identification of a new information source concerning the domain ontology

The unit of Modena will collaborate to the study and development of semantic tools able to improve the effectiveness of current keyword-based techniques of search engines, e.g. Goolge. The search of new Web sources will be assisted by natural language comprehension techniques, some of them are already implemented in TUCUXI (Benassi, 2004). The purpose is to obtain a synthetic representation of the meanings contained in a text and to maintain the semantic relations between terms. The relevance of the source will be evaluated by means of a semantic similarity measure (developed within THEME 2) between the ontology and the lexical chains.

According to the whole project, each activity is divided into three separated phases.
PHASE 1
Concerning 1.1 and 1.2 activities and together with the other units, the proposal of standards and emerging languages for ontology definition and treatment, with respect to the description of the ontologies' evolution will be analyzed.
PRODUCTS
D1.R1: Critical Analysis of the emerging ontology languages and standards. (BO, MO, ROMA, TN)

PHASE 2
A language for ontology definition and treatment, with particular attention to ontology evolution, will be defined; the activity will be done together with the units of Bologna and Trento. A prototype to add a new information source to the domain ontology will be developed. Concerning the activity 1.3 a critical analysis of existing techniques for lexical chains extraction will be produced.

D1.R2: Definition of the language for domain ontology specification (BO, MO, TN)
D1.R4: Critical analysis of existing techniques for lexical chains' extractions (MO)
D1.P1: Prototype for adding a new information source to the domain ontology (MO)

PHASE 3
During this phase and with respect to the activity 1.3, the unit of Modena will focus on the study and implementation of new lexical chains extraction algorithms. Starting from the critical analysis of existing techniques (D1.R4), the unit of Modena will individuate the features that new algorithms have to implement, with respect to three main aspects of interest: the first concerns the kind of documents (Web pages), the second is about the computational complexity, in particular linear complexity algorithms will be defined; the third deals with the accuracy that the synthetic and semantic representation has to ensure. In particular, the attention will be posed on the word sense disambiguation as a preliminary phase for building lexical chains representative of the analysed sources. The prototype will strongly extend the TUCUXI functions (Benassi, 2004) and complete RoadRunner (developed by the unit of Roma), improving its ability to assign semantics to information extracted from data intensive web sites.

PRODUCTS
D1.R6: Definition of lexical-chains techniques to associate semantics to a data-intensive site schema (RM, MO)
D1.P2: Prototype to extract lexical chains from web sites (MO)

THEME 2
Concerning THEME 2, the activity will focus on the definition of languages, techniques and algorithms to obtain mappings among different ontologies. In general, we can have matching relations identified by the instances or by the structures, relations coming from the analysis of the involved sources, matching derived from external tools, e.g. lexical analysis or logic inference.
Within the project, the Unit of Modena will cooperate in designing matching algorithms that take into account topics detailed hereafter.
Systems that use constraints coming from lexicon try to exploit the schema elements names to find similar elements.
The similarity between the schema elements names can be identified in different ways, such as the name identity, the identity of canonical names obtained after a preprocessing step, the hypernyms identity, the name identity on the basis of the user's suggestions. It is worth noting that the identity between two names (or hypernyms) is not always a mere string comparison. A name can be associated to one or more meanings (polysemy) and, viceversa, different names can have the same meaning within the source context (synonymy). To generate such kinds of relations, it is necessary to refer to lexicon ontologies that catalogue terms on the basis of their meaning, so it is possible to perform correct comparisons.
Analogously, it is worth noting that, sometimes, names associated to some database schema elements are not semantically relevant. In these cases, it is advisable to adopt ausiliary techniques for semantics extraction from data analysis and/or techniques for analysing and exploiting the natural language comments expressed by the source designer.
Other constraints can be derived from the analysis of the two schemas: for example, the identity can be derived on the basis of the data type equivalence or on the keys domain, on the cardinality of the relations and on the IS-A relations.
Finally, on the basis of the synthetic source representation coming from the lexical chains techniques, the unit of Modena will study and develop semantic similarity measures (evolution of the ones presented in (Budanitsky, 2001)) to quantify the data source relevance with respect to the reference domain ontology. Such measures will take into account two different aspects: the first concerns the "exact" matching between meanings/concepts, the second is about the semantic similarity between concepts.
According to the whole project, each activity wil be divided into three different phases:
PHASE 1
During the first phase, the most important matching algorithms in the literature will be critically analysed, with particular attention to the one that adopt techniques for conflict resolution among different representations and the ontology mapping standard proposals.
PRODUCTS
D2.R1: Critical analysis of languages and mapping techniques (MO, TN)
PHASE 2
During the second phase, we will collaborate to define a language for the mapping and the matching algorithm that will

automatically suggest mapping relations on the basis of semantic similarity.
PRODUCTS
D2.R2: Definition of the language for semantic mappings specification (MO,TN)
D2.R3: Empirical evaluation of semantic similarity measures (MO)

PHASE 3
The activity carried out during phase 3 will be devoted to extract a synthetic source representation by means of the lexical chains technique. In particular, the algorithm in D1.R6 will be evaluated under several parameters, such as technological ones (robustness of the extraction process, computational complexity,...) and qualitative ones (expressiveness of the lexical chains as a methodology to describe sources, possibility to extend them in a multilingual environment, effectiveness of the proposed techniques in order to assign semantics to the data extracted by RoadRunner).
In addition, the unit of Modena, as far as the problem to quantify the relevance of the sources with respect to the reference domain ontology, will study and develop semantic similarity measures between the lexical chains and the reference ontology. Such measures will take into account two different aspects: the first concerning the "exact" matching between meanings/concepts, the second is about semantic relatedness between concepts (Budanitsky, 2001). For example, the semantic similarity measure will differently consider when the reference ontology and the source, by means of lexical chains, share the concept of book and when the source does not contain the concept of book but the concept of volume.
PRODUCTS
D2.P1 Prototype of the platform for the automatic generation/management of mappings between heterogeneous domain ontology (MO, TN)

THEME 3
Concerning THEME 3, the first objective is to study techniques for the automatic translation (rewriting) of a given query, formulated w.r.t. a local domain ontology, in order to be compatible with the other ontologies available in the distributed environment. Such process is necessary if we want to be able to answer queries in the most effective and complete way, thus taking advantage of the full potentialities of the information available in the data sources. In fact, it is not reasonable to think all the information useful to satisfy the users' informative needs to come from the source on which the query has been formulated; rather, we must exploit all the useful sources, querying also the ones which are integrated in ontologies different from the one on which the original query is formulated. Therefore, the goal is to deliver techniques which, taking advantage of the semantic information (concepts and mappings) of the involved ontologies, be able to rewrite a given query towards the other ontologies, in the most possibly faithful way w.r.t. the original query.
A second objective is to study techniques for the GVV Global Instance computation. The Global Instance is computed within the query resolution phase and on the basis of the following elements: the mappings between the GVV and the local sources, the identification of the local sources' elements representing the same real word object (join map), and the full-disjunction operation which allows to obtain a unique result for the same object instantiated in different sources. The specific activity of this THEME is to study and develop techniques for the semi-automatic definition of join maps and the generalization of the full-disjunction operation.In particular, we will develop solutions which are valid both under the hypothesis of "semantic homogeneity", i.e. equal values for common local attributes belonging to different sources and referring to the same real world object, and in the general (inconsistent) case.

THEME 3 activities will also be divided in three phases.
PHASE 1
The first phase will produce a critical analysis of the ontology-based query rewriting techniques (deliverable D3.R1).
PRODUCTS
D3.R1: Critical analysis of query languages and ontology-based query rewriting techniques (BO, MO, TN)

PHASE 2
The query rewriting operation has to take into account the global classes involved by the query. In order to obtain a complete e minimal answer, we have to compute the global instance.
This phase activities involve the choice of the query rewriting approach, together with the definition of a series of techniques preliminar to an effective rewriting (deliverable D3.R3). In particular, the semantic mapping between the ontologies (THEME 2) will constitute a good starting point in order to be able to rewrite a given query, originally expressed w.r.t. a local ontology. The idea will be to exploit such mappings in order to quantify the similarities between the various concepts described in the involved ontologies. The techniques used to estimate the similarities between the involved concepts are not strictly included in the rewriting phase but they constitute one fundamental and preliminary phase for it. Such techniques should not be limited to the exploitation of the semantic information about the concepts available in the ontologies, but they should also make use of the context (structure) in which such concepts are inserted, following other recently presented approaches (Garcia-Molina et al., 2002). The similarities extracted by means of such approaches will provide a good basis for the rewriting phase; such rewrite will not only adapt the structure of the query but it also will consistently rewrite its values. Moreover, the extracted similarities will be used in order to estimate and quantify the distance between the rewritten queries and the original one. In order to obtain a complete and minimal answer, we will exploit and extend the full-disjunction method.
PRODUCTS
D3.R3: Definition of the query language and of the ontology-based query rewriting techniques (BO, MO, TN)

PHASE 3
In PHASE 3, the proposed rewriting techniques and the full-disjunction operation will be implemented in the prototype for query formulation (deliverable D3.P1).
PRODUCTS
D3.P1 Prototype for query formulation (BO, MO)

2.6 Descrizione delle attrezzature già disponibili ed utilizzabili per la ricerca proposta con valore patrimoniale superiore a 25.000 Euro

Testo italiano

Nessuna

Testo inglese

Nessuna

2.7 Descrizione delle Grandi attrezzature da acquisire (GA)

Testo italiano

Nessuna

Testo inglese

Nessuna

2.8 Mesi uomo complessivi dedicati al programma

Testo italiano

		Numero	Mesi uomo 1° anno	Mesi uomo 2° anno	Totale mesi uomo
Personale universitario dell'Università sede dell'Unità di Ricerca		*5*	*20*	*20*	*40*
Personale universitario di altre Università		*0*	*0*	*0*	*0*
Titolari di assegni di ricerca		*0*			
Titolari di borse	*Dottorato*	*2*	*12*	*12*	*24*
	Post-dottorato	*0*			
	Scuola di Specializzazione	*0*			
Personale a contratto	*Assegnisti*	*0*			
	Borsisti	*2*	*12*	*12*	*24*
	Dottorandi	*0*			
	Altre tipologie	*0*			
Personale extrauniversitario		*2*	*8*	*8*	*16*
TOTALE		**11**	**52**	**52**	**104**

Testo inglese

		Numero	Mesi uomo 1° anno	Mesi uomo 2° anno	Totale mesi uomo
University Personnel		*5*	*20*	*20*	*40*
Other University Personnel		*0*	*0*	*0*	*0*
Work contract (research grants, free lance contracts)		*0*			
PHD Fellows & PHD Students	*PHD Students*	*2*	*12*	*12*	*24*
	Post-Doctoral Fellows	*0*			
	Specialization School	*0*			
Personnel to be hired	*Work contract (research grants, free lance contracts)*	*0*			
	PHD Fellows & PHD Students	*2*	*12*	*12*	*24*

	PHD Students	*0*			
	Other tipologies	*0*			
No cost Non University Personnel		*2*	*8*	*8*	*16*
TOTALE		**11**	**52**	**52**	**104**

3.1 Costo complessivo del Programma dell'Unità di Ricerca

Testo italiano

Voce di spesa	Spesa in Euro	Descrizione
Materiale inventariabile	*20.000*	*Personal computer, portatili, stampanti, software.*
Grandi Attrezzature		
Materiale di consumo e funzionamento	*4.000*	*Carta, cancelleria, fotocopie, supporti magnetici*
Spese per calcolo ed elaborazione dati		
Personale a contratto	*36.000*	*Progetto e sviluppo di software*
Servizi esterni		
Missioni	*40.000*	*Missioni nazionali ed internazionali collegate alle attivita' di ricerca.*
Pubblicazioni	*5.000*	*Pubblicazioni a stampa.*
Partecipazione / Organizzazione convegni	*10.000*	*Partecipazione / Organizzazione convegni nazionali ed internazionali*
Altro		
TOTALE	**115.000**	

Testo inglese

Voce di spesa	Spesa in Euro	Descrizione
Materiale inventariabile	*20.000*	*Personal computers, laptop, printer, software*
Grandi Attrezzature		
Materiale di consumo e funzionamento	*4.000*	*Papers, photocopies, diskettes, etc*
Spese per calcolo ed elaborazione dati		
Personale a contratto	*36.000*	*Software tool design and development*
Servizi esterni		
Missioni	*40.000*	*National and International meeting expenses related to the research activity.*
Pubblicazioni	*5.000*	*Publications .*
Partecipazione / Organizzazione convegni	*10.000*	*National and International meeting and Conferences participation/organization*
Altro		
TOTALE	**115.000**	

3.2 Costo complessivo del Programma di Ricerca

		Descrizione
Costo complessivo del Programma dell'Unità di Ricerca	*115.000*	
Fondi disponibili (RD)	*11.500*	*Fondi di Dipartimento denominati WINK*
Fondi acquisibili (RA)	*23.000*	*Fondi provenienti da Ateneo in caso di approvazione del progetto*
Cofinanziamento di altre amministrazioni		
Cofinanziamento richiesto al MIUR	*80.500*	

3.3.1 Certifico la dichiarata disponibilità e l'utilizzabilità dei fondi di Ateneo (RD e RA)

SI

(per la copia da depositare presso l'Ateneo e per l'assenso alla diffusione via Internet delle informazioni riguardanti i programmi finanziati e la loro elaborazione necessaria alle valutazioni; legge del 31.12.96 n° 675 sulla "Tutela dei dati personali")

Firma _____ Data 20/03/2004 ore 19:24