Masse de Données et Connaissances - Appel à projets 2007 (ANR-07-MDCO) Fiche B : Description technique détaillée du projet	
Acronyme du projet : MEDEOR	

1 Résumé du projet en français

1 page au maximum. Sauf exigence particulière de confidentialité à mentionner dans le formulaire excel, le résumé ci-dessous pourra être diffusé par l'ANR ou par l'Unité Support.

Un grand nombre de sources de données biomédicales sont publiquement accessibles via Internet. Elles concernent une grande variété de données incluant en particulier les expressions et les séquences des gènes, les fonctions et les structures moléculaires et les caractéristiques des maladies. Ces sources de données sont des intermédiaires entre l'observation expérimentale et la capacité de synthétiser à grande échelle la compréhension des systèmes biomédicaux et des diverses interactions dans lesquelles ils participent et auxquelles ils répondent. Développées indépendamment, distribuées géographiquement et gérées de manière autonome, ces sources de données sont fortement hétérogènes [MMB03, BML+04, Suj01, GP+01].

L'interrogation, le croisement et l'analyse de ces sources de données sont essentiels pour la recherche sur la détermination des fonctions des gènes, sur l'analyse du génome, sur l'origine génétique des maladies et pour la découverte de nouveaux traitements. Pour relever ce challenge, le développement d'une infrastructure globale est nécessaire [BML+04, GP+01, HR+01, PRM+04, GRM+04, LAS+04, MMB03]. Cependant, des barrières scientifiques et technologiques fortes limitent l'interopérabilité et la navigation efficace entre ces sources de données. Ces barrières comprennent, outre l'hétérogénéité pour représenter les données, la complexité et la nature dynamique des données, ainsi qu'un très grand nombre de sources de données. Le nombre de sources de données biomédicales est actuellement de l'ordre de plusieurs centaines et croit régulièrement. De plus, la nature des données est dynamique (c'est-à-dire, fréquemment mises à jour) en raison de la rapidité d'événements venant d'un effort de recherche mondial [BLM+04, MMB03, LAS+04].

L'objectif du projet MEDEOR est de concevoir une infrastructure logicielle pour interroger de manière efficace des sources de données biomédicales hétérogènes, dynamiques et distribuées à grande échelle. Plus exactement, nous aborderons les problèmes suivants :

- La reformulation de requêtes qui sélectionne les sources de données appropriées et qui décompose la requête en sous-requêtes en prenant en compte les différences de nommage, la polysémie, les différences sémantiques et la nature hautement dynamique des données.
- L'amélioration des performances en proposant des méthodes d'optimisation dynamique qui s'adaptent aux évolutions des sources de données et de leurs ressources ainsi que des politiques de cache matérialisant les résultats des sous-requêtes les plus fréquentes.
- La définition d'un processus de découverte des connaissances efficace qui permet à l'utilisateur final d'exploiter ce processus d'une façon transparente pour le libérer de la charge de leur développement.

Pour réaliser ce projet, nous voulons utiliser l'expertise de chaque partenaire. Les partenaires (LIRIS, IRIT, LIFL et I3S), très complémentaires, sont déjà fortement impliqués dans des initiatives en systèmes d'information, extraction de connaissances, algorithmique parallèle et distribuée et en grilles de calcul. Nous validerons nos travaux à partir de sources disponibles sur l'Internet et de sources fournies par des équipes biomédicales avec lesquelles nous entretenons d'anciennes collaborations comme les "Génopôles de Lille et Lyon", l'INRA et "les Hospices Civils de Lyon".

2 Summary in English

1 page au maximum. Sauf exigence particulière de confidentialité à mentionner dans le formulaire excel, le résumé ci-dessous pourra être diffusé par l'ANR ou par l'Unité Support.

A large number of biomedical data sources are publicly available over the Internet and have a large variety of data including in particular gene expressions and sequences, molecular structures and functions and diseases characteristics. These data sources are intermediaries between experimental observation and the ability to synthesize large scale understanding of biomedical systems and manifold interactions in which they participate, and to which they respond. Developed independently, geographically distributed and managed autonomously, these data sources are highly heterogeneous [MMB03, BML+04, Suj01, GP+01].

The querying, the crossing and the analysis of these data sources are essential for the research on the gene functions, genome analysis, genetic origin of the diseases and for the discovery of new drug. To raise this challenge, the development of a global infrastructure is necessary [BML+04, GP+01, HR+01, PRM+04, GRM+04, LAS+04, MMB03]. However, significant scientific and technological, barriers impede the interoperability and efficient navigation among these data sources. The barriers include besides representational heterogeneity, complexity and dynamic nature of data, as well as the very large number of the data sources. The number of biomedical data sources is currently in the hundreds and growing. Moreover, the nature of the data is dynamic (i.e., frequently updated) due to the rapidity of developments coming from a global research effort [BLM+04, MMB03, LAS+04].

The objective of MEDEOR project is to design a software infrastructure to efficiently query the dynamic heterogeneous and widely distributed biomedical data sources. More exactly, we would like to tackle the following problems:

- The query reformulation which selects the appropriate data sources and decomposes the query into subqueries must take into account the differences of naming, polysemy, the semantic differences and the highly dynamic nature of data.
- The performance improvement by proposing dynamic query optimization methods that adapt to the evolution of the data sources and their resources and cache policies which materialize the result of the most frequent sub-queries.
- To define efficient knowledge discovery processes which allow the end-user to exploit this process in a transparent way in order to free him/her from the burden of their development.

To achieve this project, we would like to use the expertise of each partner (LIRIS, IRIT, LIFL and I3S). These complementary partners are already strongly involved in the initiatives in information systems, data mining, parallel and distributed algorithmics, and in GRID computing. We will validate our work from sources available over the Internet and from sources supplied by biomedical teams with which we maintain the old collaborations such as the "Génopôles de Lille et Lyon", the INRA institute and the "Hospices Civils de Lyon".

3 Introduction

2 pages au maximum. On décrira brièvement le projet, les enjeux scientifiques - techniques - économiques associés, les verrous à lever, les résultats attendus et les perspectives ouvertes sur le plan scientifique et/ou en termes d'applications. On discutera la pertinence par rapport à l'appel à projets.

This project tackles the integration, mediation and analysis of dynamic heterogeneous and widely distributed biomedical data sources. Indeed a large and increasing number of biomedical data sources are publicly available over the Internet. However, there does not exist a global comprehensive infrastructure that allows to query and to mine this disseminated information. Existing infrastructures INDUS, ARAMEDIA-II, GGM, INBIOMED, BACIIS, MediaGrid suffer from a lack of scalability (in terms of number and size of data sources) and dynamicity (in term of evolving data sources and queries).

The number of biomedical sources is large and growing, and the nature of biomedical data is dynamic (i.e., frequently updated) due to the rapidity of developments coming from a global research

effort. Data sources are daily added and updated. The various systems [BML+04, GP+01, HR+01, GRM+04, LAS+04, CP+05, CB+04, PG+07] that have been developed in order to access data from these sources can be broadly categorized in three approaches: information linkage, data warehousing and mediation. The information linkage approach [BML+02, GP+01, HR+01, CW99] can be easily implemented for simple operations. However it does not allow handling complex operations and is limited regarding the scalability of an integration system. The data warehousing approach [BML+02, BML+04, DC+01, PG+ 07] replicates the local data in shared sources and provides better query performance. However maintaining the warehouse up-to-date against the large number of dynamic sources increases the cost of query processing. The optimal approach for query processing among such sources of data is definitely mediation [GP+01, PS+99, CW99, HR+01, CP+05, CB+04]. This approach can provide scalability and dynamicity (adding data sources does not impact the infrastructure) while delivering up-to-date data (data are requested on demand from the data sources). However, the adding of new data sources needs complex human intervention and expertise. Furthermore the performance (query processing time) is compromised in a very large scale context.

In mediation approaches, an integrated global schema is used to model the data in the constituent sources. Users refer to such schema when querying the global information system. When a source is added or an existing source is updated the global schema will change. So, the later must be independent of local schemas in order to deal with scalability. Ontologies provide the semantic understanding of the relationships between different objects in different data sources. This feature is particularly important in biomedical informatics because of the complexity of this domain. An ontology, like UMLS [Bod04] as a global schema for an integrated system, is independent of the individual biomedical data sources and will only change when the biomedical domain evolves and needs to take into account new discoveries and concepts. The stability of the concepts in the ontology is not affected by the addition and/or deletion of data sources and their changes.

Nevertheless, if the earlier systems such as INDUS, ARMEDA-II, INBIOMED, BACIIS, MediaGrid are based on a mediation approach, several open research problems still remain before large scale deployment:

1- The query reformulation which selects the appropriate data sources and decomposes the query into subqueries must take into account the differences of naming, polysemy, and semantic differences. Naming differences means distinct lexical terms denoting the same semantic objects across data sources. These differences are also known as synonymy - multiple terms with the same meaning such as variations in the names of data values, for instance Doctor versus Physician. Semantic differences occur when the labels of data values are identical across multiple data sources, but their meanings are not precisely equivalent. For example, the attribute “list of synonyms for a gene'” indicates the “Unigene number” in the Genome Data-Base (GDB), the “International Protein Index (IPI)'” number in the Gene Ontology database, and the “Enzyme Commission number” in the SWISS-PROT database. Polysemy occurs when a data value has multiple meanings in various contexts. For instance, the gene symbol for the *fibrillin 1* gene is *FBN1*, while this for the *mouse fibrillin* gene it is *Fbn1* (note lower case *bn*). Similarly, the alternate gene symbol for the *fibrillin 1* gene is *MFS1*, while *MFS1* is also used as an abbreviation for the disease *Marfan* Syndrome. In a static context existing solutions deal with such type of interoperability issues. However, none of them can cope with highly dynamic environments.

2- Particular attention must be paid to performance improvement. Since biomedical sources are autonomous and heterogeneous, they are considered as black-boxes. As a consequence, the data retrieval rate from a particular source at the mediator is typically difficult to predict and control. It strongly depends on the complexity of the subquery assigned to the sources, the load on the source and the characteristics of the network. Delays in data delivery may stall the query engine, leading to a dramatic increase in response time. Moreover, the characteristics of the subquery results are difficult to assess, due to the autonomous nature of the data sources. The sizes of intermediate results used to estimate the costs of the integration query execution plan are then likely to be inaccurate. The query execution plan in a mediation environment, prepared at compile time (i.e., static query optimization), produces poor performance at runtime because the mediator has limited knowledge of the behavior of the remote data sources. Monitoring tools are mandatory in order to take into account the variability of the environment (e.g. cpu load and network bandwidth). Furthermore caching

methods should be developed in order to store at the mediator level a part of data frequently accessed in order to accelerate repetitive queries.

3- The addition of a new data source in an existing integrated system should not modify the translation process of a query expressed on a global schema into equivalent queries of constituent data sources. Hence, the declaration of the correspondence between the global concepts and objects, and local concepts and objects of the constituent data sources must be independent of the query reformulation process.

4- The mining of very large datasets, requires lot of storage and computation power resources. GRID platforms offer a very effective response to this issue. Indeed GRIDs can be used for the storage aspect as well as for the computation aspect. As public data available are sparse and dynamic it is not possible to store a copy of these data on a single site. But the analysis of these data will impose the construction of temporary datasets that will have to be stored in order to be available for the Datamining task. Moreover, with their high computation power resources, GRID offers a chance to develop very efficient datamining algorithms. With this environment, heuristics and exact methods will be developed and hybridized and new advanced distributed datamining algorithms will be proposed.

The contribution of the project will be:

1- Query reformulation: This reformulation consists in i) selecting the appropriate data sources, ii) decomposing the query into sub-queries among these sources, and iii) replacing the query terms by the terms used by the data sources taking into account the differences in naming, the polysemy, and the semantic differences.

2- Dynamic query optimization methods that adapt to the evolution of the data sources and the resources (e.g. CPU, network bandwidth) of the infrastructure. Thus these optimization methods will be coupled to a monitoring service.

3- Semantic collaborative query caching: The need for high performance leads to the deployment of caching strategies to cache the results of the queries on the system. The management of caches will rely on collaborative patterns that optimize the global system and not solely on partial and local view of the data being queried and the results being produced by those queries.

4- The distributed heterogeneous and dynamic natures of the data to be mined as well as the large size of datasets, bring to the datamining task a high combinatoric aspect. Therefore it becomes necessary to propose new efficient and advanced combinatorial optimization methods combining multi-objective meta-heuristics, parallel and distributed computing and decision-aided techniques. Such complex methods must be exploited by the end-user in a transparent way in order to free him/her from the burden of their development.

4 Contexte et état de l'art / Context and state-of-the art

2 pages au maximum. On précisera, en particulier, la position du projet par rapport à la concurrence nationale et internationale, en donnant les références nécessaires. Pour les projets à vocation appliquée, on décrira également le contexte économique dans lequel se situe le projet en présentant une analyse du marché, de ses tendances,…

In this section we first describe, the related works on existing bio-medical systems. Then, we detail those relative to the various specific points which will be developed in this project.

Related work on biomedical systems

The online availability of biomedical data sources has motivated researchers to discover new knowledge in functional genomics, proteomics and medicine. However, the data sources are autonomous and highly heterogeneous, so it is very difficult to navigate and process open-ended queries [BML+04, BML+02, MMB03, Suj01]. In order to overcome these limitations, various systems have been implemented: Bio-Kleisli [CW99], Object Protocol Model (OPM) [CM95, CK+97], DiscoveryLink [HR+01], TAMBIS [GP+01, PS+99], BACIIS [BML+04, BML+02], INDUS [CP+05], ARMEDA-II [GRM+04, PRM+04], INBOIMED [LAS+04], GGM [PG+07] and MediaGrid [CB+04].

It is interesting to observe that the existing systems in biomedical are confronted with several barriers such as:

1. The generation of a query by a user. For a user it is difficult to memorize all data attributes due to the domain complexity and the number of data sources. The number of biomedical data sources is currently in the hundreds and growing on. For example, UMLS [Bor04] integrates 900 000 concepts as well as 12 million relations among these concepts.
2. The reformulation of a user query. The major problem is to take into account the differences of naming, polysemy, and semantic differences during the reformulation process.
3. The building of an efficient execution plan. Since biomedical data sources are autonomous and heterogeneous, it is difficult at compile time to estimate accurately all parameters necessary for the optimization process.
4. The maintenance cost. The nature of the data is dynamic due to the rapidity of developments coming from a global research effort. Hence, adding, or updating of data sources do not have any impact on the global schema and on the programs.

Related work on query reformulation

The SIMS [AKS96, AC+03] provides an approach to declarative query reformulation for dynamic information integration. The system directly linked local concepts and objects defined in local terminology to domain concepts and objects in source models for the query reformulation. This strategy restricts scalability and flexibility of the integration system is the case of a large and dynamic context. The system handles a limited number of structured data sources. A similar approach for biomedical data sources has been used in [KPL03]. In [NF04], the system use ontology-based query reformulation to provide a user with meaningful answers to his/her queries. However this approach is limited to querying a single structured data source and does not reformulate a query into subqueries.

Project ARIANE [JA+01, AJ00, JF+98] employs knowledge representation system based on UMLS for semantic integration in Biomedicine. A user query and data sources are modeled with conceptual graphs. The system selects relevant data sources for a given query through conceptual graph matching. This system cannot select the data sources based on some criteria such as reliability and quality of data.

In BioRegistry [MD+05], for a given user query the most appropriate data sources are identified and selected among all the existing bioinformatics data sources. The system employs an information retrieval approach through Formal Concept Analysis where data sources instead of documents are searched and indexation is based on metadata reflecting information about sources rather than on data extracted from the documents. Their approach is limited to binary relationship between sources and metadata concepts and the relevance between a query and data sources is established by sharing concepts in their sets of metadata.

Related work on query optimization

From a declarative user query, an optimizer at the mediator can generate multiple access plans involving local operations at the data source level and global operations at the mediator level. However cost estimates based on query optimization of classical systems are hard since (i) data sources do not export needed statistical information (e.g., HTML files, object-oriented databases), (ii) cost formulas for processing an operator (e.g., selection or join) vary depending on the implementation of the wrapper and the underlying data source and (iii) processing and communication costs are difficult to determine and may vary over time according to the data availability and to the network or system loads.

Various solutions to the cost estimate problem have been proposed [AP+96, DKS92, GST96, ZMS03]. Whatever the solution of the cost model is, the statistics stored in the database catalogue are subject to obsolescence notably it is very difficult to estimate the processing and communication costs during the compile time in large-scale mediation systems. Hence, in [IHW04, KD98, KMS00, BBD05] centralized dynamic optimization methods are proposed in order to react to estimation errors (i.e. variation between the parameters estimated at compile-time and the parameters computed at run-time) and resource unavailability (i.e. data, CPU, memory, networks). In large-scale mediation systems, the centralization of dynamic query optimization methods generates a bottleneck and produces relatively significant message passing on a network with low band-width and strong latency [AH+04, OMH05].

The decentralized methods described in [CV04, IF+99, UF00] improve the cost of local processing by adapting the use of the CPU, I/O and memory resources with the changes of the execution environment (e.g. estimation errors, delays in data arrival rates). However, the methods proposed are focused mainly on the resources such as CPU, I/O and memory and do not take into account the network resource. In particular, these methods do not minimize the volume of data transferred on networks.

Related work on caching

We propose in this project to work on query cache management, using the semantics underlying the data bases structures. The idea of performing semantic caching is issued from the database community [CR94], which is not surprising since the schema of a database represents the semantics of the data being stored. Thus it is easy to find the relative degree of locality between the requests of the users and the semantic distance between the data as well. In general, the cache is constituted by a set of elements associated to an index of semantic relationships called semantic regions [DF+96]. At the reception of a new query [RD98], it is split in two parts, one being found locally in the cache, and the other being retrieved from the database.

Some work has addressed, for web based multi data sources, the use of semantic caching [BC+99, LC99, CRS99, SA05]. These solutions exploit either a pivot schema on which requests are cached [LC99], or the notion of semantic region with an emphasis on the possibility of exploiting the region in different contexts [CRS99].

None of the already proposed approaches (to the best of our knowledge) exploits the possibility for different semantic query caches to cooperate to solve user requests. Work has been conducted in the context of web caching, where documents are cached [Coq06]. We plan to extend these works for semantic collaborative query caching.

Related work on parallel and distributed data mining.

Data mining is a complex iterative and interactive process allowing the discovery of new surprising models from large datasets [HK00]. Different tasks have been identified and received research interest (e.g. ACM SIGKDD and PAKDD conferences, KDNet excellence network, etc.). The majority of these tasks are classification, clustering (unsupervised classification), association rule mining, attributes selection, etc. Much effort has been particularly devoted to association rule mining (see the 2002 report of the working sub-group “Usages Multiples des Motifs Fréquents” of the specific research group GaFoDonnées of CNRS-STIC). However, the major part of these efforts deals with the market basket problem for which the very classical Apriori algorithm, based on the support metric, has been proposed [AIS93]. More research is needed to tackle other problem instances especially discovering rules in medical and genomic data. The scientific challenge is to propose, in addition to the Apriori algorithm, new methods to deal with dense correlated (several thousands of correlated attributes), heterogeneous and evolving datasets.

Since the Apriori algorithm [AIS93], many variants have been proposed to deal with large databases [MTV94, BMS97], where the number of instances may be very high, but usually the number of attributes remains reasonable. Other variants have also been proposed to deal with dense and even correlated databases [BR01, BBR03]. However, gene expression datasets are often dense but not large in the sense that they contain a very small number of experiments compared with the large number of genes (up to 30 000) under study. Such an application is an unfavourable case even for the best algorithms as the support (frequency) metric is not yet adapted. The Lille partner team has experience in tackling such problems with multi-criteria meta-heuristics. Their approach consists in taking into account several criteria to evaluate the quality of a rule (not only the support). This allows to integrate other criteria (biological and others) and to reduce the complexity of the search space (due to the huge number of genes). Multi-criteria meta-heuristics are then developed to deal with such problems. On the other hand, only a few parallel and distributed approaches have been proposed to deal with the distributed location of the data sources [KP+99] and with the performance of the data mining process [Zak00, HKK00]. Another challenge is to propose parallel distributed meta-heuristics for rule mining in distributed genomedical data. These methods have to be designed with the perspective of deploy on grids.

5 Partenaires / Partnership

2 pages au maximum. On présentera les partenaires et on décrira aussi les compétences et savoir-faire des équipes impliquées vis-à-vis de l'état de l'art au niveau national et international. On mentionnera ici, pour chacune des équipes, son implication éventuelle dans d'autres projets. Les indications fournies serviront à apprécier la qualité du partenariat.

The PYRAMID team of the IRIT laboratory is specialized in the dynamic optimization of distributed query on a large scale network. They are editors of several special editions of international and national journals (international Journal of Computer Systems Science & Engineering, « Revue Ingénierie des Systèmes d'information », and « Revue des sciences et technologies de l'information, série Technique et science informatiques »). They also took part in the organization of the PaDD and Globe workshops in conjunction with the international conference DEXA (www.dexa.org). The following four references are examples of the scientific production of the team:

- B. Ozakar, F. Morvan, A. Hameurlain, "Mobile Join Operators for Restricted Sources", Mobile Information Systems: An International Journal, 1(3):167-184, 2005.
- B. Ozakar, F. Morvan, A. Hameurlain, "Query Optimization: Mobile Agents versus Accuracy of the Cost Estimation", International Journal of Computer Systems Science & Engineering, 20(3): 161 - 168, May 2005.
- J. Pierson, J. Gossa, P. Wehrle, Y. Cardenas, S. Cahon, E.S. Mahmoud, L. Brunie, C. Dhaenens, H. Abdelkader, N. Melab, M. Miquel, F. Morvan, T. El Gazali, A. Tchounikine, "GGM Efficient Navigation and Mining in Distributed Geno-Medical Data", IEEE Transactions on NanoBioscience, IEEE. 2007. (à paraître)
- J.-P. Arcangeli, A. Hameurlain, F. Migeon, F. Morvan, "Mobile Agent Based Self-Adaptive Join for Wide-Area Distributed Query Processing", International Journal of Database Management, 15(4): 25-44, October 2004.

The "Communicating Information Systems" of the LIRIS laboratory is specialized in the data management in large scale heterogeneous systems such as Grids. They organize annually a workshop on "Data Management in Grids" at the conference VLDB (Very Large Data Bases), and participate in a number of program committee on all aspects of data management and grid computing. Their interest is focused on security, mediation and cache management. The following four references are examples of the scientific production of the team on subjects related to the current project:

- Y. Cardenas, J. Pierson, L. Brunie, "Temporal Storage Space for Grids", In Second International Conference on High Performance Computing and Communications (HPCC 2006), Munich, Germany, LNCS, pp. 803-812, September 2006..
- Y. Cardenas, JM. Pierson, L. Brunie, "Uniform Distributed Cache Service for Grid Computing", In proc. of the Sixteenth International Workshop on Database and Expert Systems Applications (DEXA 2005), Copenhagen, Denmark, IEEE Computer Society, pp. 351-355, 2005.
- L Seitz, JM. Pierson, L. Brunie, "Encrypted Storage of Medical Data on a Grid", Methods of Information in Medicine (2):198-202, Schattauer GmbH. 2005.
- J Montagnat, F. Bellet, H. Benoit-Cattin, V. Breton, L. Brunie, H Duque, Y. Legre, I. Magnin, L. Maigne, S Miguet, JM. Pierson, L Seitz, T Tweed, "Medical images simulation, storage and processing on the European DataGrid testbed", Journal of Grid Computing 4(2):387-400, Springer Verlag, 2004.

The OPAC team of the LIFL laboratory is specialized in grid computing for combinatorial optimization and data mining. They organise in many conferences and workshops in areas such as the "Gestion et extraction parallèle et distribuée de connaissances" of the EGC conference. They deal with real applications in genomics and proteomics in collaboration with the network of Genopoles. The following 4 references are examples of their contributions:

- E-G. Talbi, A. Zomaya, "Grids for Bioinformatics and Computational Biology", John Wiley & Sons, USA, 2007.
- M. Khabzaoui, C. Dhaenens, E-G. Talbi, "A cooperative genetic algorithm for knowledge discovery in microarray experiments", in Parallel Computing for Bioinformatics and

Computational Biology, Edited by A.Y. Zomaya, John Wiley & Sons, chapter 13, pp. 303-324, USA, 2006.
- R. Bolze, F. Capello, E. Caron, M. Dayde, F. Desprez, Y. Jegou, P.Primet, E. Jeannot, S. Lanteri, M. Leduc, N. Melab, G. Mornet, R. Namyst, B. Quetier, O. Richard, E-G. Talbi, I. Touche, « GRID'5000: A large scale and highly reconfigurable Grid», International Journal of high Performance Computing applications (IJHPCA)), accepted.
- N. Melab, S. Cahon, E-G. Talbi, « Grid computing for parallel bioinspired algorithms», Journal of Parallel and Distributed Computing (JDPC),Elsevier Science, Vol 66(8), pp 1052-1061, 2006.

The MAINLINE team of the I3S laboratory is specialized in Resource Annotation, Semantic Web and Collaborative Communicating Systems. They participe in many international and national projects in these areas such as the COLOR action 2007 "GRIWES", PCSI REFERECES (2006-2007). They deal with real applications in resource annotation for elearning and in collaborative organizational resources for industry. The following 4 references are examples of their contributions:

- T.D.T. Nguyen, N. Le Thanh. "Intergrating Indentification Contraints in Web Ontology" – 9th ACM-SIMIS-AAAI Int. Conf. ICEEIS (Conference on Enterprise Information Systems), June 12 - 16, 2007, Funchal, Madeia, Portugal
- T.A.L. Pham, N. Le Thanh "Decomposition-Based Reasoning for Large Knowledge Bases in Description Logics" - 13th Int. Conference ISPE, IOS Press, VOL 143 "Leading the Web in Concurrent Engineering", September 2006, pp 288-195
- C. Le Duc, N. Le Thanh, M-C. Rousset. "Compact Representation for Least Common Subsumer in Description Logic ALE ". – AICOM Journal (The European Journal on Artificial Intelligence) - Volume 19, Number 3, 2006, pp. 239 - 273
- O. Corby, R. Dieng-Kuntz, C. Faron-Zucker, F. Gandon. Searching the semantic web : Approximate query processing based on ontologies. IEEE Intelligent Systems Journal, 21(1), 2006

6 Organisation et management du projet / Project organization and management

2 pages au maximum. On décrira l'organisation mise en place pour le projet et la manière dont sera assurée la coordination de celui-ci. Le mode de pilotage du projet sera décrit en tenant compte des aléas susceptibles d'être rencontrés.

The members of the project will meet three times a year in order to review the project. During these meetings, we will hear a status report of each partner and discuss any important issues. Among these issues we will work on actual integration of the different tasks and software developed by each partner. The frequency of the meetings is designed to minimize the risks of the project. However, if certain risks should arise, we will decide on corrective action and update the planning.

7 Structure du projet – Description des sous-projets / Structure of the project – Work-packages

10 pages au maximum. On décrira le programme de travail en identifiant pour chaque étape, les objectifs poursuivis, le rôle de chaque partenaire et les moyens mis en œuvre. La valeur ajoutée des coopérations entre les différentes équipes sera argumentée. Si des doctorants sont présents dans le projet, on explicitera leur sujet de thèse et les conditions de leur encadrement.

We suggest the design and the development of a software prototype of the infrastructure which will allows bringing an innovative answer to the scientific bolts identified previously. We paid attention to the possibility of adding or updating data sources with moderate maintenance costs.

In figure 1 we present the general software architecture (the related components of the project are in bold). In this figure, the solid lines with solid arrow-heads represent the passage of a query through the system. The solid lines with hollow arrow-heads represent the use of information models during query processing. The doted lines with solid arrow-heads represent the passage of data from the data sources. As we can see, the parts related to the project are strongly interconnected: this will

obviously generate tight collaboration between the various partners. From this figure, we can see the components datamining, query reformulation, query optimization and caching which will be developed in this project. These four components are described in the following sections.

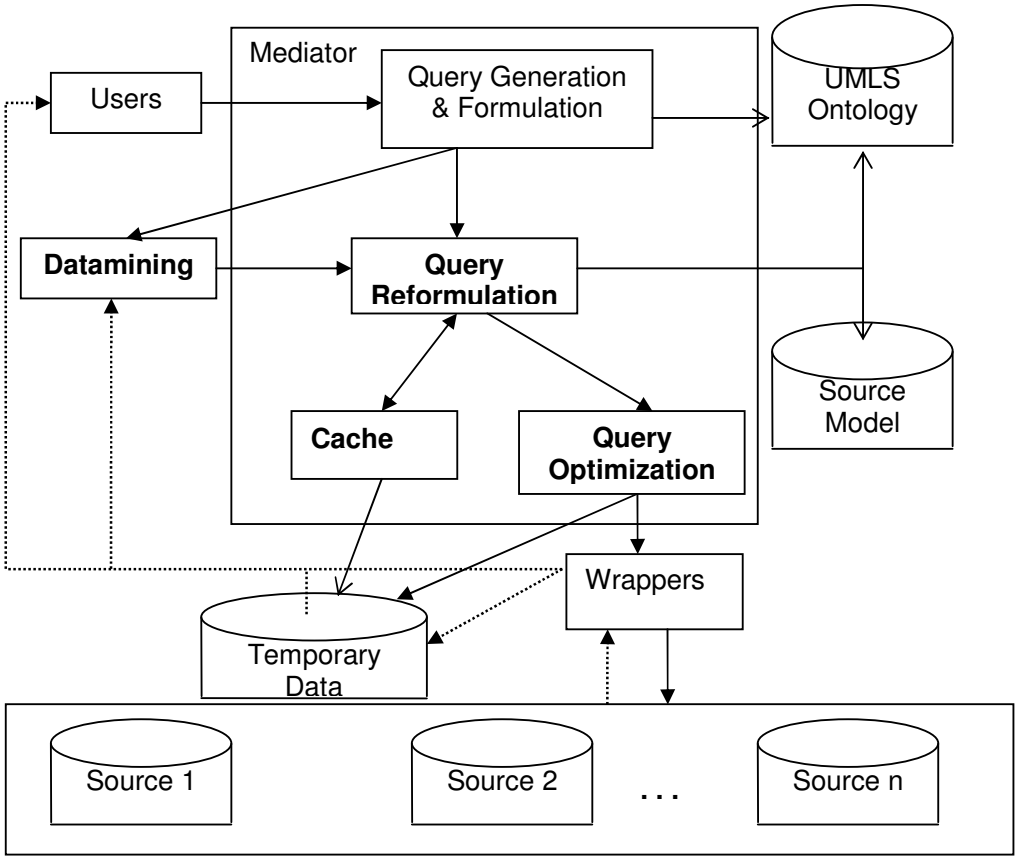


Figure 1: Software architecture of the MEDEOR project

In the mediation approach, an integrated global schema is used to model the data in the constituent sources. Users refer to such schema when querying the global information system. When a source is added or an existing source is updated the global schema will change. Hence, the, global schema must be independent of local schemas in order to handle scalability. Ontologies provide the semantic understanding of the relationships between different objects in different data sources. This feature is particularly important in biomedical informatics because of the complexity of this domain. In this project we use UMLS [Bod04] as ontology. UMLS is developed by the US National Library of Medicine. UMLS [Bod04,] as a global schema independent of the individual biomedical data sources and will only change when the biomedical domain evolves to include new discoveries and concepts. The stability of the concepts in the ontology is not affected by the addition and/or deletion of data sources; and/or updates in data sources of the integration system.

Query reformulation

A user generates a query using ontology for accessing data about objects (i.e., instances of concepts) in constituted data sources without knowing where and how the objects are organized in the sources. The query is defined by a set of selections and projection over the ontology terms satisfying certain conditions. There are possibilities that the terms used in a query are not the same as those that are used in a data source due to mismatch between the user's world view and the database designer's world view. In other words, there might be values in the data source that are syntactically different from user's terms but have the same meaning and express the same intention as the user. The user query needs to be reformulated in order to obtain the meaningful result from all the sources.

The query, generated by a user, needs to be translated into equivalent local queries of constituent data sources. In order to minimize the maintenance cost in query translation, a declarative mapping is adopted for query translation. Declarative mappings utilize the correspondence between global concepts and objects and local concepts and objects of the constituent data sources. The correspondence is (i) formally encoded such that a software process may inspect it, and (ii) stored independently of the software code that actually performs query translation. This is also known as modelling information sources for integration. The source model describes the structure and contents of the constituted data sources. The model of each source includes every fact that can influence the decision concerning when and how to utilize the source, such as data model used, query language, network location, cost model and local concepts. Modelling of information sources makes an integration system scalable because sources can be added and updated without changing the query translation software and consequently the cost of the query translation is minimized.

In query reformulation, the first step in rewriting a query expressed in the terms of the domain model is to select the appropriate data sources. This is done through declarative mappings from the concepts in the domain model to the concepts in the data source models that correspond directly to data source data (as discussed in the previous subsection). If the user requested terms (i.e., concepts) and a data source's terms (i.e., concepts) match, then the mapping is straightforward. However, in many cases, the mappings are not straightforward due to terminological mismatching. The original domain model query terms must be replaced by the terms that correspond to data source concepts. This replacement of terms is performed in the form of generalization, specialization and/or partition of domain concepts. After the selection of relevant data sources, the user query is then decomposed into the subqueries referring to data stored in the selected sources.

Moreover, some individual data sources have very limited query capabilities in biomedicine because they are not databases in the conventional sense. Consider for the example: *What is the 3D structure of all alcohol dehydro-genases that belong to the enzyme family EC:1.1.1.1 and is located within the human chromosome region 4q21-4q23?*

Some data sources may not accept *enzyme family* as an acceptable keyword and some may not allow the logical AND operator to combine two keywords, e.g., *enzyme family* and *human chromosome* region. However these data sources contain relevant data and therefore they cannot be ignored. In order to obtain meaningful results from such data sources, one approach is to select some intermediary data sources. The intermediary data sources are sources that cannot directly accept input or produce output constraints of the user query. However they can be used to translate the attribute of one data source into an attribute of another data source.

Query Optimization

Once the query has been reformulated to obtain a meaningful result, the next step is to generate the query execution plan for accessing and processing the data. The execution plan consists of precise operations that need to be performed for accessing and processing the data, as well as the order in which they are to be performed. The optimal execution plan is selected among the possible plans on the basis of the cost of accessing different information sources, the cost of retrieving intermediate results, and the cost of combining these intermediate results to produce the final result in a distributed environment.

In query processing systems for biomedical data sources, the mediator/wrapper architecture is adopted because the sources are autonomous and heterogeneous. In this context, the subqueries, reformulated from the user query, are sent to data source wrappers and then their results are integrated at the mediator level to produce the final result. The query optimization process is used to order the subqueries on the basis of their dependencies, select the location for processing the intermediate results, and determine which queries can be executed in parallel. Since biomedical sources are autonomous and heterogeneous, therefore, they are considered as black-boxes. As a consequence, the data retrieval rate from a particular source at the mediator is typically difficult to predict and control. It depends on the complexity of the subquery assigned to the sources, the load on the source and the characteristics of the network. Delays in data delivery may stall the query engine, leading to a dramatic increase in response time. Moreover, the characteristics of the subquery results are difficult to assess, due to the autonomous nature of the data sources. The sizes of intermediate results used to estimate the costs of the integration query execution plan are then likely to be inaccurate. The query execution plan in an integrated environment, prepared at compile time (i.e., static query optimization), produces poor

performance at runtime because the mediator has limited knowledge of the behavior of the remote data sources. In order to handle the performance issue, the query execution plan needs to be modified at runtime (i.e., dynamic query optimization) on basis of response time, availability and size of intermediate results.

In a large-scale biomedical mediation system, it is not possible to adapt the proposed dynamic optimization methods [IHW04, KD98, KMS00, BBD05]. Indeed, these methods are supervised by a master process that controls all the processes participating in the optimization and/or in the dynamic reoptimization to make all the necessary adaptations. The processes executing the running operations or operations waiting for resources are thus completely controlled by the optimizer. This control generates a bottleneck for the optimization process [AH+04, OMH05]. It thus becomes convenient to make the query execution autonomous and self-adaptable. In this perspective, an approach to be investigated consists of using a programming model based on mobile agents [FPV98], knowing that at present mobile agent platforms supply only migration mechanisms, but they do not offer migration decision policy. It is for that reason that we wish to design and to develop an execution model based on mobile agents and a proactive policy. This execution model will be coupled to a monitoring tools in order to take into account the variability of the environment (e.g. cpu load and network bandwidth).

Caching

The role of the cache in the architecture is to optimize the access and the handling of the queries. Queries may be complicated to reformulate (as seen before), thus it is important to keep the already computed queries in the cache in this context. Moreover, optimizing the access to the data with the use of query caching also decreases cost.

If the mediators are considered independent, one could imagine several cache management policies. For instance, where the query has the following shape: Q1 = “SELECT tableT.name FROM tableT WHERE tableT.id=3”, we can cache the exact result of the query with the query so that, if the next query is : Q2 = “SELECT tableT.name FROM tableT WHERE tableT.id=3 AND tableT.sum=50” then the first result is scanned and the result delivered immediately. On the other hand, if the cache system receives Q2, it may decide to transform the original query which would be too specific into something more general as Q1, or even more general, depending on the size of the cache and the related queries already received (somehow merging different queries into more general ones).

If we consider that there exist a number of caches collaboratively serving the queries, then the necessary information may be spread throughout the distributed system. Collaboration techniques based on the exchange of appropriate messages have to be constructed and sent from one cache to the other, in order to collectively define some semantic region of interests.

Data mining

The role of the data mining part of the project is to extract knowledge from distributed, dynamic and heterogeneous data. We will propose efficient and advanced distributed techniques combining multi-criteria optimization, metaheuristics such as evolutionary algorithms, grid computing and visual decision aid interaction. Indeed, many criteria must be taken into account to extract knowledge. The associated combinatorial problem is NP-Complete and efficient parallel metaheuristics deployed on Grids must be used. Finally, decision aid and visualisation techniques must be developed for biologists to make easier the biological validation of the results.

The knowledge model will be based on association rules, a well known model in the data mining community which is adapted to the application domain (geno-medical data). Such complex models and algorithms must be exploited by the end-user in a transparent way in order to free him/her from the burden of their development.

The proposed techniques will be validated on real applications from the network of genopoles in France such as the Genopole of Lille dealing with multifactorial diseases (http://www.genopole-lille.fr).

Common prototype

A common prototype which will integrate the various developments of partners will be developed. This prototype will use the platform compounded by several PC installed in the project “ACI masses de

données GGM" (LIRIS, LIFL and IRIT). In this project, we are thinking of increasing the number of PCs of this platform to begin our first experiments.

Currently, these three partner laboratories participate in the GRID5000 project aimed at creating an efficient national GRID of 5000+ processors. We intend to evaluate on a very large scale the solutions stemming from this project.

Organization of the project

We detail in this section the identified tasks required to achieve this project. First, we present for each task its decomposition into sub-tasks, for which a deadline and a deliverables are given. Each task is under the responsibility of one member of the project. Then, we summarize the schedule of the project by a Gantt diagram

	Query Reformulation (QR)	**Deliverables**	**Deadline**
Responsibility	*N. Le Thanh (I3S)*		
Patners involved	I3S, IRIT, LIFL		
Sub-tasks	- QR1: analysis of several biomedical data sources available on the Internet and concepts bound in UMLS. - QR2: Declarative mappings defining the correspondence between the global concepts and objects in UMLS and local concepts and objects of the constituent data sources. - QR 3: query reformulation expressed on the part of UMLS identified in QR1. The sub-task QR3 will lean on the task CM.	QR1 – Report QR2 – Software QR3 - Software	T0 + 6 T0 + 15 T0 + 30

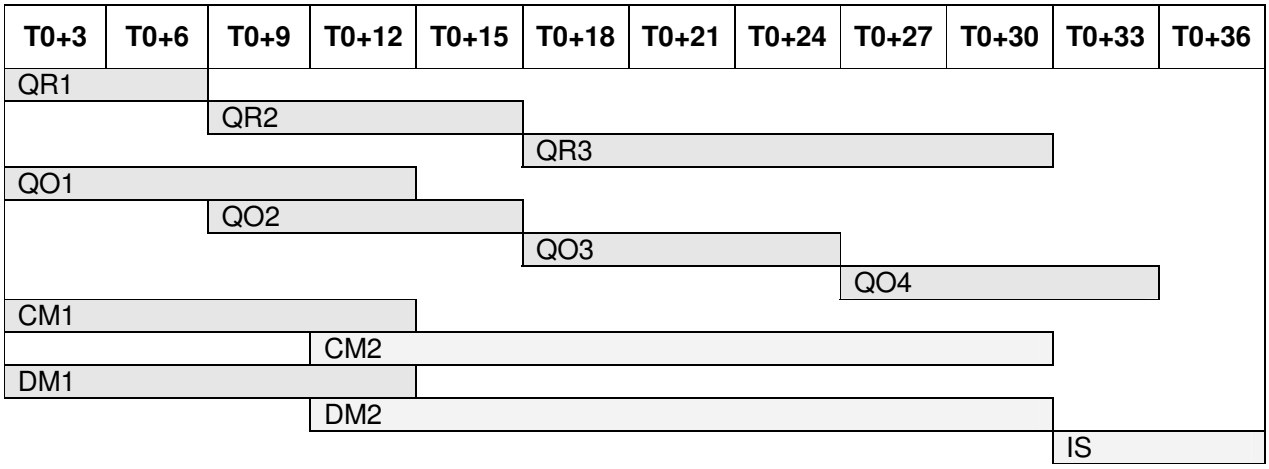
	Query Optimization (QO)	**Deliverables**	**Deadline**
Responsibility	*A. Hameurlain (IRIT)*		
Patners involved	IRIT, I3S, LIRIS		
Sub-tasks	- QO1: mobile agent based execution model. - QO2: scheduling of the sub-queries identified in QR3. - QO3: Code generation based on execution model developed in Q01. - QO4: initial placement of mobile agents generated in QO3. The sub-task QO2 depends on the task CM.	QO1 – Software QO2 – Software QO3 – Software QO4 – Software	T0 + 12 T0 + 15 T0 + 24 T0 + 33

	Cache Management (CM)	**Deliverables**	**Deadline**
Responsibility	*L. Brunie (LIRIS)*		
Patners involved	LIRIS, IRIT		
Sub-tasks	- CM1: local cache management for queries, identifying the relevant data structure to store the results of the queries. - CM2: collaborative cache management, identifying the patterns and information that has to be deployed among the participating mediators of the architecture.	CM1 – Software + article CM2 – Software + article	T0 + 12 T0 + 30

	Data mining (DM)	**Deliverables**	**Deadline**
Responsibility	*E. G. Talbi (LIFL)*		
Patners involved	LIFL		
Sub-tasks	- DM1: development of the data mining algorithms. - DM2: Validation of the algorithms on real applications.	DM1 – Software DM2 – Software + article	T0 + 12 T0 + 30

	Integrated Software (IS)	**Deliverables**	**Deadline**
Responsibility	*F. Morvan (IRIT)*		
Patners involved	All partners		
Sub-tasks	The prototype will integrate all the above defined developments.	IS - Software	T0 + 36

The following Gantt diagram shows the planning of the project.



Participation of teams to the project

The IRIT members will make a contribution to the project in Query optimization problems. The LIFL members will participate in the data mining part. The LIRIS and IRIT members will work on cache management. The I3S members will participate in the design and development of query reformulation. Each of the partners will have to cooperate strongly with the others, considering the interdependence of the tasks which appears in the detailled descriptions above.

Interest of the collaboration

This project offers a unique opportunity of collaboration between various French laboratories. The interest of this collaboration is based on the complementarities and the synergy between the partners. The scientific exchanges are situated at the levels:

- Query reformulation
- Query optimization
- Cache management
- Knowledge discovery.

Only these complementarities are able to supply all the skills indispensable to the design and to the development of the complex software architecture needed for biomedical mediation systems. The domains of competence covered by the consortium are data mining, distributed mediation systems on large scale networks and cache management. This collaboration was initiated in a previous GGM project of the “ACI Masses de données 2004” and has give several excellent results: several publications with one partner (Mobile Information Systems: An international Journal, Journal of Grid Computing, Journal

of Parallel and Distributed Computing) or several partners (example of joint publications: BioGrid 2005, IEEE Transactions on NanoBioscience), a prototype platform, 5 PhD thesis and 5 HDR (Habilitation à Diriger des Recherches). The main difference between the GGM and MEDEOR projects is that GGM project did not take into account the high dynamicity and heterogeneity of data. GGM project leans on a data warehouse approach which replicates the local data in shared sources. In the type of applications targeted by the MEDEOR project, this copy on a single site is simply not possible since the data volume is very large and the time to maintain the data warehouse up-to-date becomes prohibitive.

The complementarity of our 4 partners is also a guarantee of dissemination of the results of the project via:

- Co-authored publications in a wide editorial spectrum;
- The participation in several regional, national and international working groups:
  - The partners are members of several working groups “Impacts du Grid Computing, du Peer-to-Peer Computing, et du Mobile Computing sur les systèmes de bases de données et d'informations hétérogènes et distribuées à grande échelle” GDR I3 where A Hameurlain (IRIT), member of the present consortium is the leader.
- The joint organization of workshops and conferences. The IRIT (A. Hameurlain) and the LIRIS (L. Brunie) laboratories have five times organized the workshop PaDD ‘Parallel and Distributed Databases: Innovative Applications and New Architectures’ associated with the international conference on databases DEXA and three times the workshop DMG ‘Data Management in Grid’ associated with the international conference on VLDB. The IRIT (A. Hameurlain) laboratory have three time organized the workshop GLOBE ‘Grid and Peer-to-Peer Computing Impacts on Large Scale Heterogeneous Distributed Database Systems’ associated with the international conference on databases DEXA, with the cooperation of LIRIS (L. Brunie) and of LIFL (E.-G. Talbi) in the program committee.

PHD subject and management

IRIT Laboratory

- *Raddad Alking*: Cost Models for Large Scale Query Optimization, Advisor: Abdelkader Hameurlain, supervisor: Franck Morvan
- *Mahmoud El Samad*: Efficient biomedical query processing on data Grid, Advisor: Abdelkader Hameurlain, supervisor: Franck Morvan

LIRIS Laboratory

- *Yonny Cardenas*: Collaborative caching in Grids, advisor: Lionel Brunie, supervisor: Jean-Marc Pierson (IRIT)
- *Julien Gossa*: Grid user view, advisor: Lionel Brunie, supervisor: Jean-Marc Pierson (IRIT)

LIFL Laboratory

- *Alexandru Tantar:* Distributed docking on Grids, Advisor: El-Ghazali Talbi, supervisor: Nordine Melab

I3S Laboratory

- *Anastasiya Yurchyshyna:* Regulation modeling by an ontology approach, Advisor: Nhan Le Thanh, supervisor: Catherine Faron
- *Thi Dieu Thu Nguyen:* identification constraints in OWL-k and relational request transformation, Advisor: Nhan Le Thanh
- *Thi Anh Le Pham:* Ontology decomposition and request optimization, Advisor: Nhan Le Thanh

References

[AC+03] Y. Arens, C. Y. Chee, et al., “Retrieving and integrating data from multiple information sources”, Intl. Journal of Cooperative Information Systems, 2(2):127-158, june 2003.

[AH+04] J.P. Arcangeli, A. Hameurlain et al., “Mobile Agent Based Self-Adaptive Join for Wide-Area Distributed Query Processing”, Journal of Database Management, 15(4): 25-44, Oct-Dec 2004.

[AIS93] R. Agrawal, T. Imielinski and A. Swami, “Mining association rules between sets of items in large databases”, In proc. of the 1993 SIGMOD conf., pages 207-216, Washington, D.C, May 1993.

[AP+96] S. Adali, K. S. Candan, et al., “Query Caching and Optimization in Distributed Mediator Systems”, In proc. of the 1996 SIGMOD Conf., pages 137-148, June 1996.

[AKS96] Y. Arens, C. A. Knoblock, W. Shen, “Query Reformulation for Dynamic Information Integration”, Journal of Intelligent Information Systems, 6(2/3): 99-130, May 1996.

[BBD05] S. Babu, P. Bizarro, D. DeWitt, “Proactive Re-Optimization”, In proc. of the SIGMOD Conf., pages 107-118, June 2005.

[BBR03] J-F. Boulicaut, A. Bykowski and C. Rigotti, “Free-sets: a condensed representation of boolean data for the approximation of frequency queries”, Data Mining and Knowledge Discovery journal, 7(1): 5-22, 2003.

[BC+99] L. Bouganim, T. Chan-Sine-Ying et al., “Miro Web: Integrating Multiple Resources through Semistructured Data Types”, In proc. of the Twenty-Fifth International Conference on Very Large Data Bases, VLDB'99, pages 750-753, September 1999.

[BML+02] Z. Ben-Miled , N. Li et al., “Complex Life Science Multidatabase Queries”, proc. of the IEEE, 90(11):1754-1763, November 2002.

[BML+04] Z. Ben-Miled , N. Li et al., “On the Integration of a Large Number of Life Science Web Databases”, In proc. of the 1st International Workshop on Data Integration in the Life Sciences (DILS), pages 172-186. Springer, March 2004.

[BMS97] S Brin, R. Motwani and C. Silverstein, “Beyond market basket: Generalizing association rules to correlations”, In proc. of the 1997 SIGMOD conf., pages 265-276, Tucson, Arizona, June 1997.

[Bod04] O. Bodenreider, “The Unified Medical Language System (UMLS): integrating biomedical terminologie”, Nucleic Acids Research, 32(Database-Issue): 267-270, 2004.

[BR01] A. Bykowski and C. Rigotti, “A condensed representation to find frequent patterns”, In proc. of ACM PODS'01, pages 267-273, Santa Barbara, CA, USA, May 2001.

[CB+04] Christine Collet, Khalid Belhajjame, et al., “Towards a Mediation System Framework for Transparent Access to Largely Distributed Sources. The MediaGrid Project”, In First Intl IFIP Conf. on Semantics of a Networked World: ICSNW 2004, LNCS 3226, Paris, June, 2004

[Coq06] D. Coquil, “Conception et mise en oeuvre de proxies semantiques et coopératifs”, PhD thesis, march 2006.

[CP+05] D. Caragea, J. Pathak et al., “Information Integration and Knowledge Acquisition from Semantically Heterogeneous Biological Data Sources”, In proc. of 2nd International Workshop on Data Integration in the Life Sciences (DILS), pages 175-190. Springer, July 2005.

[CR94] C. M. Chen and N. Roussopoulos, “The Implementation and Performance Evaluation ofthe Adms Query Optimizer: Integrating Query Result Caching and Matching”, In proc. of EDBT'94, pages 323-336, Cambridge, UK, March 1994.

[CRS99] B. Chidlovski, C. Roncancio and M.-L. Schneider, “Semantic Cache Mechanism for Heterogeneous Web Querying”, In proc. of the Eighth International World-Wide-Web Conference, 1999.

[CV04] C. Collet and T.-T. Vu, “QBF: A Query Broker Framework for Adaptable Query Evaluation”, Proc. of 6th Intl. Conf. on Flexible Query Answering Systems, Springer Verlag Publishers, Lyon, France, pages 362-375, June 2004.

[DF+96] S. Dar, M. J.Franklin et al., “Semantic Data Caching and Replacement, 1996”, In proc. of the Twenty-First international Conference on Very Large Data Bases (VLDB'96), pages 330-341, September 1996.

[DKS92] W. Du, R. Krishnamurthy, M.-C. Shan, “Query Optimization in a Heterogeneous DBMS”, In proc. of the 18 Intl. Conf. on VLDB, pages 277-291, Aug. 1992.

[FPV98] A. Fuggetta, G. P. Picco, G. Vigna, “Understanding Code Mobility”, IEEE Trans. on Software Engineering, 24(5):342-361, 1998.

[GP+01] C. A. Goble, N. W. Paton et al., “Transparent Access to Multiple Bioinformatics Information Sources”, IBM System Journal, 40(2):532-551, 2001.

[GRM+04] M. Garcia-Remesal, V. Maojo et al., “ARMEDA II: Supporting Genomic Medicine through the Integration of Medical and Genetic Databases”, In proc. of IEEE International Conference on BioInformatics and BioEngineering (BIBE), pages 227-236. IEEE CS, March 2004.

[GST96] G. Gardarin, F. Sha, Z.-H. Tang, “Calibrating the Query Optimizer Cost Model of IRO-DB, an Object-Oriented Federated Database System”, In proc. of the 22nd Intl. Conf. on VLDB, pages 378-389, Sept.1996.

[HK00] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, August 2000.

[HKK00] E-H. Han, G. Karypis and V. Kumar, “*Scalable Parallel Data Mining for Association Rules(2000)”*, IEEE Transactions on Knowledge and Data Engineering, 12(3): 377-352, May/June 2000.

[HR+01] L. M. Haas, J. E. Rice et al., “Discovery Link: A System for Integrated Access to Life Sciences Data Sources”, IBM System Journal, 40(2):489-511, 2001.

[IF+99] Z.-G. Ives, D. Florescu et al., “An Adaptive Query Execution System for Data Integration”, Proc. of the ACM SIGMOD Intl. Conf. on Management of Data, ACM Press, Philadelphia, pages 299-310, June 1999.

[IHW04] Z. G. Ives, Al. Y. Halevy, and D. S. Weld, “Adapting to Source Properties in Processing Data Integration Queries”, In proc. of the ACM SIGMOD, pages 395-406, June 2004.

[JA+01] M. Joubert, S. Aymard, et al., “Ariane: A mediation framework with health information sources”, In proc. of the 10th World Congress on Medical Informatics, pages 343-347, ISO, 2001.

[KD98] N. Kabra and D. DeWitt, “Efficient mid-query re-optimization of sub-optimal query execution plans”, In proc. of ACM SIGMOD, pages 106-117, 1998.

[KM 06] S. Khan, F. Morvan, “Query Processing in Biomedical Informatics”, In proc of the 19th Intl Conf. on Parallel and Distributed Computing Systems, San Francisco, USA, ISCA, pages 165-170, 2006.

[KMS00] L. Khan, D.Mcleod, and C. Shahabi, “An Adaptive Probe-based Technique to Optimize Join Queries in Distributed Internet Databases”, Knowledge and Information Systems, 2: 373-385, 2000.

[KP+99] H. Kargupta, B. Park et al., “Collective data mining: A new perspective toward distributed data mining”, Advances in Distributed Data Mining, Eds: Hillol Kargupta and Philip Chan, AAAI/MIT Press, 1999.

[KPL03] J. Kohler, S.Philippi, M. Lange, “SEMEDA: Ontology-based Semantic Integration of Biological Databases”, Bioinformatics, 19(18): 2420-2427, 2003.

[LAS+04] V. Lopez-Alonso, J. P. Sanchez et al., “INBIOMED: a Platform for the Integration and Sharing of Genetic, Clinical and Epidemiological Data Oriented to Biomedical Research”, In proc. of the 4th Intl Symposium on BioInformatics and BioEngineering (BIBE), pages 222-226, IEEE CS. March 2004.

[LC99] D. Lee and W. W. Chu, “Towards Intelligent Semantic Caching for Web Sources”, Journal of Intelligent Information Systems, 17(1): 23-45, 2001.

[MD05] N. Messai, M. D. Devignes et al., “Querying a Bioinformatic Data Sources Registry with Concept Lattices”, In proc. of the 13th International Conference on Conceptual Structures, pages 323-336, Germany, Springer LNAI-3596, july 2005.

[MMB03] J.A. Mitchell, A.T. McCray and O. Bodenreider, “From Phenotype to Genotype: Issues in Navigating the available Information Resources”, Methods of Information in Medicine, 42(5):557-563, May 2003.

[MTV94] H. Mannila, H. Toivonen, and A. I. Verkamo, “Efficient algorithms for discovering association rules.”, In proc. KDD'94, pp. 181-192, Seattle, WA, July 1994.

[OMH05] B. Ozakar, F. Morvan, A. Hameurlain, “Query Optimization: Mobile Agents Versus Accuracy of Cost Estimation”, International Journal of Computer Systems Sciences and Engineering, 20(3): 161-168, May 2005.

[NF04] C.B. Necib, J.C. Freytag, “Using Ontologies for Database Query Reformulation”, In proc. of the 8th East European Conference on Advances in Databases and Information Systems (ADBIS), September 2004.

[PG+07] J.-M. Pierson, J. Gossa et al., “GGM : efficient navigation and minig in distributed geno-medical data”, IEEE Transactions on Nanobioscience, Special Issue on Computational NanoBioscience, to appear 2007.

[PRM+04] D. Perez-Rey, V. Maojo et al., “Biomedical Ontologies in Post-Genomic Information Systems”, In proc. of IEEE International Symposium on Bioinformatics and BioEngineering (BIBE), pages 207-214, Taiwan, IEEE CS. Mach 2004.

[RD98] Q. Ren and M. Dunham, “Semantic Caching and Query Processing”, Southern Methodist University, Technical Report 98-CSE-4, 1998.

[SA05] G. Soundarajaran and C. Amza, “Using Semantic Information to Improve Transparent Query Caching for Dynamic Content Web Sites”, In proc. of the International Workshop on Data Engineering Issues in E-Commerce, pages 132-138, April 2005.

[Suj01] W. Sujansky, “Heterogeneous Database Integration in Biomedicine; Methodolgical Review”, Journal of Biomedical Informatics, 34:285-298, 2001.

[UF00] T. Urhan and M. J. Franklin, “XJoin : A reactively-scheduled pipelined join operator”, IEEE Data Engineering Bulletin, IEEE CS, 23(2): 27-33, September 2000.

[Zak00] M. J. Zaki, “Scalable algorithms for association mining”, IEEE Trans. On Knowledge and Data Engineering, 12(3): 372-390, May/June 2000.

[ZMS03] Q. Zhu, S. Montheramgari, Yu Sun, “Cost Estimation for Queries Experiencing Multiple Contention States in Dynamic Multidatabase Environments”, Knowledge and Information Systems, 5(1):26-49, 2003.

8 Liste des livrables / List of deliverable

Un tableau de l'ensemble des livrables du projet sera inclus sous la forme indiquée ci-après. Les dates sont à exprimées sous forme T0+x [mois].

	Libellé du livrable	Type[1]	Responsable	Partenaires participants	Date
0	Site web du projet – Mise en place au plus tard 6 mois après le démarrage du projet et mise à jour au moins semestrielle	web	Coordonnateur	All partners	T0+6
1	Minutes of the kick-off meeting	report	F. Morvan	All partners	T0
2	Biomedical data sources and concepts bound in UMLS (QR1)	report	N. Le Thanh	Nice, Lille and Toulouse partners	T0+6
3	Short project report	report	F. Morvan	All partners	T0+6
4	Mobile execution model (QO1)	software	A. Hameurlain	Toulouse, Lyon and Nice partners	T0 +12
5	Project report	report	F. Morvan	All partners	T0+12
6	Local Cache management (CM1)	Software+article	L. Brunie	Lyon and Toulouse partners	T0+12
7	Data mining software (DM1)	Software	E-G. Talbi	Lille partner	T0+12
8	Declarative mapping (QR2)	Software	N. Le Thanh	Nice, Lille and Toulouse partners	T0+15
9	Sub-queries scheduling (QO2)	Software	A. Hameurlain	Toulouse, Lyon and Nice partners	T0+15

[1] Logiciel, Publication, Site web, Communication, …

10	Short project report	Report	F. Morvan	All partners	T0+18
11	Project report	Report	F. Morvan	All partners	T0+24
12	Code generation (QO3)	Software	A. Hameurlain	Toulouse, Lyon and Nice partners	T0+24
13	Short project report	Report	F. Morvan	All partners	T0+30
14	Collaborative cache management (CM2)	Software+article	L. Brunie	Lyon and Toulouse partners	T0+30
15	Query reformulation (QR3)	Software	N. Le Thanh	Nice, Lille and Toulouse partners	T0+30
16	Application (DM2)	Software + article	E-G. Talbi	Lille partner	T0+30
17	Initial agent placement (QO4)	Software	A. Hameurlain	Toulouse partner	T0+33
18	Final report	Report	F. Morvan	All partners	T0+36
19	Integrated software	Software+article	F. Morvan	All partners	T0+36

9 Résultats escomptés – perspectives / Expected results and perspectives

2 pages au maximum.

9.1 Retombées scientifiques et techniques

On résumera les objectifs du projet et les résultats escomptés, en proposant des critères de réussite et d'évaluation. On décrira également les perspectives scientifiques et/techniques ouvertes au-delà de la durée du projet. Préciser les impacts escomptés concernant les retombées scientifiques et techniques directes et expliquer comment la pérennité des retombées scientifiques et techniques sera assurée.
Dans le cas de projets prévoyant l'exploitation des outils, codes ou méthodes dans la communauté scientifique, expliciter la communauté concernée, les modalités prévues pour l'impliquer et/ou lui permettre d'exploiter les résultats. Présenter les objectifs par rapport aux projets similaires ou concurrents.
D'une manière plus générale, expliquer comment les retombées scientifiques et techniques seront diffusées au sein de la communauté scientifique.

The purpose of this project is to propose a software infrastructure to query the heterogeneous set of biomedical data sources in order to allow scientists to extract new knowledge.
The expected scientific consequences are methodological and applicative. For the methodological aspect, the expected results are the design and the development of:

- Query reformulation methods. The query reformulation methods select the appropriate data sources and decompose the query into subqueries taking into account the differences of naming, polysemy, and semantic differences.
- Dynamic query optimization methods which allow efficient access to heterogeneous data distributed on a large scale network.
- Caching policies which store the data of the most frequent subquery on the end-user site.
- Generic mining methods hiding from the user the complexity of the methods.

Beyond the methodological aspects, we believe that the obtained results will allow the development of platforms allowing biologists to be able to query heterogeneous biomedical data sources distributed on a large scale network with moderate maintenance costs. This kind of platform will especially allow the analysis of diseases which can be explained by several factors as for example cardiovascular and Alzheimer diseases.

The criteria to evaluate the success of the project are two-fold:

- Publications: Each partner must publish in conferences of the domain the obtained results. Furthermore, some publications including all partners must be written showing the cooperation of the teams.
- Software: At the end of the project, a prototype software infrastructure will allow querying several biomedical data sources.

In order to access data from multiple data sources in an integrated system, a query is generated. This process needs to have familiarity with constituted data sources (i.e., source relevance and interfaces) and information about their data. However, in the biomedical domain, it is very difficult, if not impossible; to memorize terms or to be aware of the relevant data sources due to the domain complexity and the large number of data sources. Moreover, end-users are not prepared to handle query languages such as SQL. A possibility is to utilize domain ontologies for query generation because ontologies conceptualize metadata and domain expert knowledge about the data set. Consequently this helps in acquiring information about the data, which can be used in query generation. Using ontologies in graphical user interfaces for query generation provides data source transparency, i.e., obtaining data from various sources without knowing their query capabilities. In other words, users browse the ontology to find out what they can retrieve and how they can ask questions without having to memorize terms or be aware of the relevant data sources. Moreover, the specialization and generalization of a concept in the ontology helps in incremental building and manipulation of query expressions through interaction with a visual representation of the ontology.

9.2 Retombées industrielles et économiques escomptées (le cas échéant)

On présentera les retombées industrielles et économiques liées au projet. Si la mise au point d'un nouveau produit, procédé ou service est visée, on traitera également le problème des réglementations et des normes, existantes ou à venir. Présenter la situation actuelle du marché qui pourrait bénéficier des retombées du projet en termes de pertinence et portée possible par rapport à la demande économique et situer la place du projet dans la stratégie industrielle de (ou des) l'entreprise(s) impliquée(s) dans le projet et notamment l'évaluation du risque et de la faisabilité industrielle.

10 Propriété intellectuelle / Intellectual property

On présentera une analyse des questions de propriété intellectuelle et industrielle identifiés ou susceptibles de se poser, en termes de brevets existants, de licences à obtenir. Les principes de l'accord de propriété intellectuelle qui sera mis en œuvre entre les partenaires du consortium doivent être explicités. En cas de publication de logiciel libre, des indications sur les types de licences utilisées devront être fournies.

Il est rappelé que le règlement relatif à l'attribution des aides de l'ANR prévoit que : "A la demande du chef de projet, la confidentialité des résultats est de droit. La propriété de ces résultats appartient aux bénéficiaires de l'aide, qui en disposent selon les modalités convenues à leur niveau et sous réserve des droits à intéressement des inventeurs. Sous réserve de la nécessité de prévoir une période de confidentialité, dans les cas où des résultats sont à protéger, le bénéficiaire doit s'assurer par toute mesure appropriée de la diffusion publique des comptes rendus scientifiques ou de leurs résumés."

The partners of the project make a commitment to sign an agreement within 6 months following the announcement of the financing. This agreement is concerned with the exploitation of the results, the software property as well as the publication policy. This agreement will be published on the private part of the project website.

11 Moyens financiers demandés / Financial resources

On précisera les moyens mis en œuvre par chacune des équipes tels que décrits lors de soumission en ligne (équipement, fonctionnement, main d'œuvre, déplacements, prestations) et on en présentera ici brièvement une justification. On précisera également si certains de ces postes feront ou pourraient faire l'objet de cofinancements.

Financial justification for the Toulouse partner:

1- Renewal of a part of the computer system for the team members 12 000 euros (6+4+2) for 6 researchers (5 members and 1 PHD requested) is 2 000 euros per person for the duration of the project.
2- Participation in conferences of the domain. A participation for each member of the team per year: 5+1 researchers * 2 000 euros = 12 000
3- The travel of the team members to participate in the meetings of the project: 3 meetings in foreseen per year, with on average 3 persons at a cost of 500 euros = 3*3*500 = 4 500
4- Access to specialized documentation and consumables (cartridge printer, paper) : 1 000 per year
5- Employment of trainees: 2 000 per years.

Financial justification for the Lyon partner:

1- Renewal of a part of the computer system for the team members 10 000 euros (4+4+2) for 5 researchers (4 members and 1 engineer requested) is 2 000 euros per person for the duration of the project.
2- Participation in conferences of the domain. A participation for each member of the team per year : 4 researchers * 2 000 euros = 8 000
3- The travel of the team members to participate in the meetings of the project: 3 meetings foreseen per year, with on average 3 persons at a cost of 300 euros = 3*3*400 = 2 700 euros. The town of Lyon being more central, the travel expenses are less compared to the other cities. Starting from Lyon, it is possible to go easily by train to Toulouse, Lille and Nice.
4- Access to specialized documentation and consumables (cartridge printer, paper) : 1 000 per year
5- Employment of trainees: 2 000 per years.

Financial justification for the Lille partner:

1- Renewal of a part of the computer system for the team members 12 000 euros (4+4+4) for 6 researchers is 2 000 euros per person for the duration of the project.
2- Participation in conferences of the domain. A participation for each member of the team per year : 4 researchers * 2 000 euros = 8 000
3- Travel of the team members to participate in the meetings of the project: 3 meetings foreseen per year, with on average 3 persons and a cost of 500 euros = 3*3*500 = 4 500 euros.
4- Access to specialized documentation and consumables (cartridge printer, paper) : 1 000 per year
5- Employment of trainees: 2 000 per year.
6- A postdoctal position for 12 months.

Financial justification for the Nice partner:

1- Renewal of a part of the computer system for the team members 12 000 euros (6+4+2) for 5 researchers + 1 PHD is 2 000 euros per person for the duration of the project.
2- Participation in conferences of the domain. A participation for each member of the team per year : 6 researchers * 2 000 euros = 12 000
3- Travel of the team members to participate in the meeting of the project: 3 meetings foreseen per year, with on average 3 persons and a cost of 500 euros = 3*3*500 = 4 500 euros.
4- Access to specialized documentation and consumables (cartridge printer, paper): 1 000 per year
5- Employment of trainees: 2 000 per year.

12 Experts / Experts

Le projet pourra indiquer dans le tableau ci-dessous une liste d'expert susceptibles d'expertiser le projet et n'étant pas, à la connaissance du coordonnateur du projet, en situation de conflit d'intérêt par rapport au projet. L'Unité Support CEA et l'ANR se réservent le droit de solliciter ces experts pour ce projet ou pour tout autre projet. Le projet pourra aussi indiquer sous le tableau une liste d'expert ou d'entités qui ne doivent pas être sollicités pour expertiser ce projet, en indiquant, le cas échéant, le motif.

Suggestion d'expert pour l'évaluation[2]				
Prénom	Nom	Courriel	Affiliation (labo/entreprise/..)	Domaine(s) d'expertise
Michel	Scholl	scholl@cnam.fr	CNAM Paris	Systèmes d'intégration de données à large échelle
Claudia	Roncancio	Claudia.Roncancio@imag.fr	LSR/IMAG	Gestion de données en environnements mobiles et largement distribués

[2] *Si possible prévoir des experts étrangers.*