SISELS: a mediation system for giving access to Biology resources

Gabriela Montiel-Moreno[1*], José Luis Zechinelli-Martini[1], and Genoveva Vargas-Solar[2]

[1] Research Center of Information and Automation Technologies
Universidad de las Américas, Puebla,
Sta. Catarina Mártir s/n, 72820, San Andrés Cholula, México
[2] French National Council of Scientific Research
Laboratory of Informatics of Grenoble
681 rue de la Passerelle, BP 72, Saint Martin d'Hères, France

Contact: gabriela.montielmo, joseluis.zechinelli@udlap.mx,
Genoveva.Vargas-Solar@imag.fr

(Paper received on February 29, 2008, accepted on April 15, 2008)

Abstract. This paper describes SISELS a mediation system that enables the configuration of virtual laboratories that support transparent access to biomedical data. The main objective of SISELS is to provide users with transparent access to distributed resources satisfying certain semantic requirements for contributing to the solution of a problem. Biomedical information is seen as distributed resources that can contribute to solve a biological problem. SISELS manages biological information using views that provide different perceptions of the same resources. A view represents the semantic requirements of a group of experts to study a specific problem. Given a problem expressed in terms of concepts, SISELS analyzes subscribed resources that provide related concepts and generates a view that represents an answer. Queries and their associated results are used to maintain a problem catalog. The problem catalog provides an easy access to frequent information and promotes information sharing and collaboration between researchers from different communities of the same knowledge domain. SISELS uses three ontologies defined in $\mathcal{SHIQ}(\mathcal{D})$ [4]. It implements them using OWL [11], and uses an inference service for mediating resources and managing generated knowledge.

1 Introduction

Medicine is a science that produces vast amounts of information, useful in the development of new treatments against diseases. This information is contained in different resources such as images, genome data, documents, and Web resources, which can be located in distributed geographic zones. This information can be manipulated in order to execute Bio-informatic processes. For this reason, medical area requires tools for integrating and manipulating existing biological information in order to generate new knowledge.

Nowadays, given a problem scientists have to manually analyze each resource with respect to the problem to verify its utility. This process is long and complex when the

Special Issue in Electronics and Biomedical Informatics,
Computer Science and Informatics
Research in Computing Science 35, 2008, pp. 187-198

number of resources increases. In some cases, scientists ignore resources useful to the problem because they are not defined explicitly in terms of the specified problem. In other cases, scientists from different communities explore the same problem from different perspectives and lack of a problem catalog to visualize different approaches.

A *virtual laboratory* provides transparent access to heterogeneous and distributed data providers and mechanisms to execute queries over resources according to concepts used in a specific knowledge domain. A virtual laboratory allows to share and manage great amounts of data in a coordinated and controlled way. Besides, it provides integrated views of resources (data, systems, documents) belonging to different organizations. These views are exploited by researchers in order to solve scientific problems.

In order to build a *virtual laboratory*, it is necessary to represent its knowledge that associates resources to concepts of a specific knowledge domain (e.g., Biology) [10]. From this knowledge, customized views of resources can be generated and adapted to the requirements of a group of experts. Retrieved information must be relevant and consistent with respect to a specific context of study.

This paper presents SISELS a knowledge based mediation system used to build virtual laboratories adapted to a knowledge domain (Biology). The rest of this article is organized as follows. Section 2 describes how to build a virtual laboratory. Section 3 describes metadata associated to a resource through the bio-resource ontology. Section 4 describes the mapping ontology that defines the semantic correspondence between concepts in Biology. Section 5 presents the view ontology used to store integrated views of information generated by SISELS. Section 6 describes our approach for generating views over resources based on semantics. Section 7 describes implementation issues concerning a prototype virtual laboratory. Section 8 describes related work and compares it with SISELS. Finally, Section 9 concludes the paper and discusses current results and future work.

2 Building a virtual laboratory

SISELS (Semantic Integration System for Explotaition of biomedicaL resourceS) is a mediation system that enables the configuration of virtual laboratories to support transparent access to biomedical data (cf. Figure 1). In order to provide data according to given biological problems, SISELS uses the notion of view. A view expresses the relationship between a problem and resources that provide information about it. Both the problem and the resources are expressed under a normalized vocabulary shared by a set of experts (users). For example, a biologist studying cells from Basidiomycota fungi to determine a new treatment specifies his/her requirements with a set of concepts like: Cell, FungalCell, and Basidiocarp.

2.1 Resource

Resources are characterized in an structural and semantical way. In SISELS the semantical representation of a resource is defined by an ontology. An ontology is represented through concepts, attributes and properties within the biological area. Resources are represented in an homogeneous way through a single ontology that represents the knowledge domain of SISELS and classifies biological concepts.

2.2 Views

Information in SISELS is organized in views. A view represents the semantical requirements of a researcher and allows to have different perspectives from resources. A view is adapted according to a requirement expression. The requirement expression identifies those concepts used to access resources.

2.3 Problem

A problem is defined by concepts of the knowledge domain associated to resources that can be used for solving it. The problem catalog provides an easy access to frequent information and promotes information sharing and collaboration between researchers from different communities.

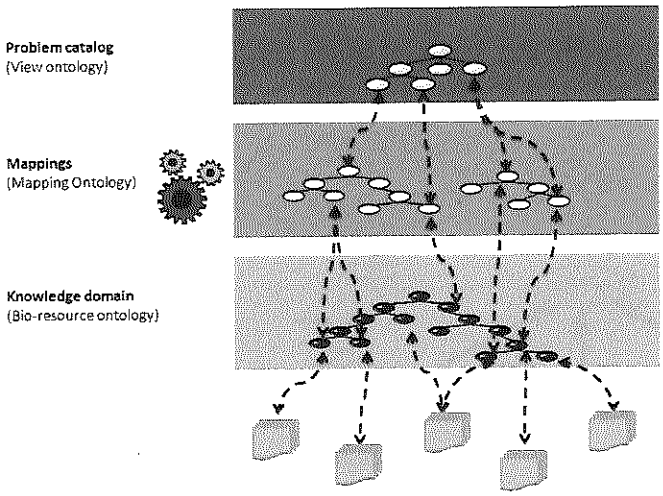

Fig. 1. SISELS general approach.

2.2 Integration approach in SISELS

The knowledge domain of SISELS is defined through the fusion of the ontologies describing the content of resources. The bio-resource ontology represents resources and associates them with biological concepts. The domain of SISELS is enriched when a new resource is subscribed to the system, when new knowledge is generated or when a biologist specifies his semantical requirements.

Researchers in the biological area specify their cases of study by using an ontology, composed by their own terms. When a biologist expresses a scientific problem in terms of biological concepts, she/he searches a set of resources that can contribute to solve a problem.

Given a problem expressed in terms of concepts of a knowledge domain, SISELS specifies the appropriate semantic mappings required to identify the resources satisfy-

ing the semantic defined by a researcher. We build the mapping ontology to represent the semantic correspondence between terms.

By using the defined mappings, SISELS analyzes subscribed resources that provide related concepts and generates a view that represents an answer.

SISELS uses knowledge representation models to achieve the semantic exploitation of biological resources. SISELS uses three ontologies based in the description logic $\mathcal{SHIQ}(\mathcal{D})$ [4] and implemented using OWL [11]: *bio-resource*, *mapping* and *view*. An ontology is composed by at least one class which represents a set of individuals sharing certain characteristics. A class is characterized by attributes that manage a data type (integer, string, and real) and has an extension defined by the concept of individual [2]. Normally, an ontology is composed also by one or more properties that define a binary relation between two concepts.

3 Bio-resources ontology

The **bio-resources ontology** models structural and semantic content of resources subscribed to SISELS (cf. Figure 2).

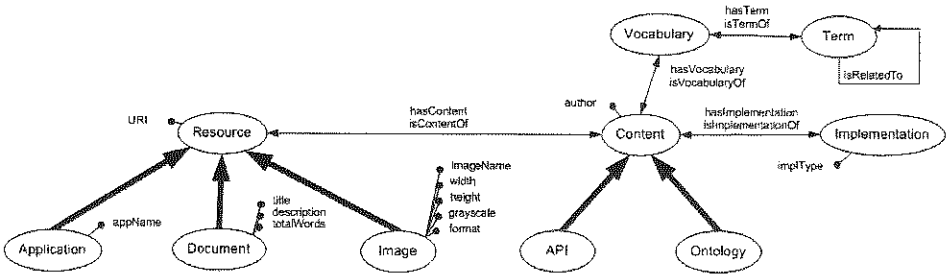

Fig. 2. Bio-resources ontology.

The main concept is Resource which is the general representation of a resource of the system. A resource is represented by an *URI* (*Universal Resource Identifier*) which is composed by a name and access protocol. In addition, a resource is characterized according to its type: Application, Document or Image. Associated to a resource, there is a Content that can be structured according to different models (e.g., an API or Ontology).

Formally, a resource is defined using $\mathcal{SHIQ}(\mathcal{D})$ in the following way:

$$\text{Resource} \equiv \exists\ \textit{hasContent}.\text{Content} \wedge =1\ \textit{URI(String)}$$

According to its format, a resource has associated metadata and characteristics describing its structure. For instance, the concept Document represents a specialization of a resource characterized by the attributes: *title, description, totalWords*. Once metadata

associated to a resource is defined, it is necessary to specify the nature of its content by the concept Content.

Information associated to a resource is managed as individuals under the ontology described. To illustrate, consider a resource denominated geneontology in PDF document format whose subject is genetics and that is related to a semantical content described in OWL format. This resource is represented as an individual geneontology of the concept PDF. The geneontology is related to contentgeneontology, which is an instance of Ontology, and with genetics which is an instance of Topic by using *hasContent* and *hasTopic* properties.

4 Mapping ontology

The **mapping ontology** represents semantic correspondences, named mappings, between concepts of different sources and the bio-resources ontology (cf. Figure 3).

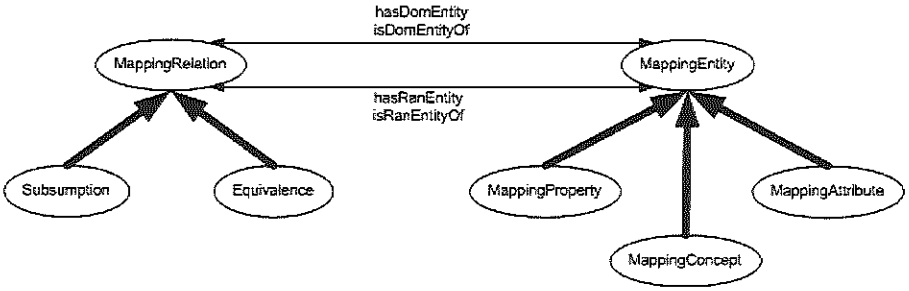

Fig. 3. Mapping ontology.

The main concept is MappingRelation which is classified in Equivalence and Subsumption. A MappingRelation is associated to the class MappingEntity by the properties *hasDomEntity* and *hasRanEntity*. A mapping is represented in $\mathcal{SHIQ}(\mathcal{D})$ language as follows.

$$\text{MappingRelation} \equiv \exists\ \textit{hasDomEntity}.\text{MappingEntity} \wedge \exists\ \textit{hasRanEntity}.\text{MappingEntity}$$

The mapping ontology is composed by instances of biological concepts, properties and attributes which own at least one semantic correspondence with another entity of a resource schema. SISELS allows to define three types of mappings as in [6]:

Equivalent mapping. Two concepts used by different resources are semantically equivalent i.e., ConceptA ≡ ConceptB, if an *equivalent mapping* is defined in the mapping ontology. For example, given the concept Cell, equivalent concepts in the bio-resources are: Microorganism, Ectoplasm, Embryo, and Unit.

Sound mapping. A *sound mapping* establishes that individuals of a concept from the ontology A are a subset of individuals in a concept of the ontology B (ConceptA ⊆

ConceptB). For example, given Fungi recovers all its existing specializations within the bio-resources ontology: Chytrids, Zygomycetes, Ascomycota, and Basidiomycota.

Complete mapping. A *complete mapping* states that a concept of resource A is a superset of a concept in resource B (ConceptA $\supseteq$ ConceptB). For example, given the concept Arachnid, a supertype selection at second level, includes concepts which are its ancestors within the bio-resources ontology: Carnivorous, Arthropod, Chelicerata.

Queries are reformulated into subqueries for accessing information within different resources. In order to achieve this task, SISELS uses mappings between the bio-resources ontology and concepts used by resources. Given a query defined as a domain ontology, SISELS proposes three types of selection of mappings: selection of equivalent concepts, selection of subtypes at *n* levels and selection of supertypes at *n* levels.

5 View ontology

The **view ontology** represents concepts related to a problem according to a knowledge domain. Each concept is associated to all resources which make reference to it (cf. Figure 5).

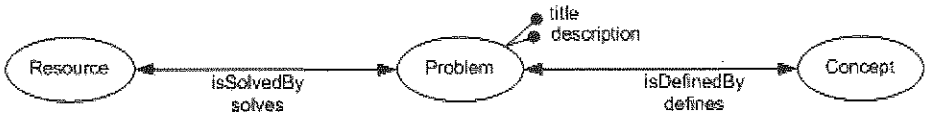

Fig. 4. View ontology.

A view is a set of concepts which define the semantics of a biological problem and is defined under $\mathcal{SHIQ}(\mathcal{D})$ language as follows.

$$\begin{aligned} \text{Problem} \equiv\ & \exists\ \textit{isDefinedBy}.\text{Concept} \wedge \\ & \exists\ \textit{isSolvedBy}.\text{Resource} \wedge \\ & =1\ \textit{title(String)} \wedge \\ & =1\ \textit{description(String)} \end{aligned}$$

The main concept in the view ontology is Problem which represents a biological problem. Each problem is associated with at least one Concept using the property *isDefinedBy*. A concept defines the semantics of the problem and must belong to the SISELS knowledge domain. Also, a Problem is associated to a set of instances of Resource for identifying those resources that provide content associated to the problem.

6 Generating views in SISELS

Given a query that defines the requirements of a biologist using a set of concepts, SISELS obtains an integrated view of resources. A view is specified as a query on the bio-resources ontology. Recall that the view represents the semantical requirements of a biologist.

Queries are expressed in terms of SISELS domain. For example, consider a biologist who wants to study the behavior of the antimiotic agent Vinorelbine in the treatment of breast cancer. The biologist must specify to SISELS the semantic related to his problem through the definition of concepts like: Vinorelbine, Antimiotic_agent, Breast, and Cancer.

Views generation in SISELS is done in four phases: analysis, reformulation, assignment, and generation. The rest of this section focuses on the description of each phase.

6.1 Analysis

This task verifies if a query is well formed. A query is well formed if it is model of the ontology: its concepts are either a class, a property or an attribute; concepts belong to the SISELS domain, and if they verify the axioms and constraints associated to the bio-resources ontology. This task executes a query by verifying whether each concept, representing the semantics of a problem, is defined as an instance by the bio-resources ontology.

6.2 Query rewriting

The objective of this phase is to prove whether a query can be rewritten with respect to the concepts representing resources content. Hence, a query is rewritten by expressing the concepts of the bio-resources ontology in terms of those used in the content of the resources. This is done by selecting equivalent concepts, supertypes or subtypes for each concept defined in the query over the mapping ontology using a reasoning service [21]. For example, consider the term: Cancer. In the rewriting phase, the biologist identifies that Cancer is a synonym of Malignant_tumor and a specialization of Tumor.

6.3 Resources filtering and assignment

From the lists of concepts retrieved by query rewriting, a biologist can filter the concepts that define his/her query domain so that SISELS looks for all resources that collaborate to the solution of a problem. Reselection allows users to access a larger amount of resources relevant to the solution of their problem.

Given a list of concepts defining the semantics of a query, a list of resources associated to these concepts must be generated (assignment process). The resource list is obtained by querying the bio-resources ontology. For example, suppose that we want to identify the resources described by the concept Vinorelbine. In first instance, it is necessary to recover the contents of type Ontology which are related to the concept

through the property *isClassOf*. Then, it is necessary to find the resources related to each retrieved content by the property *isContentOf*. This way, we would recover a Breast Cancer web page and a Vinorelbine document.

Retrieved resources are described partially or totally by the concepts of the query. A biologist can filter the resources relevant to the solution of his/her problem.

6.4 Generation

After assignment, an integrated view with all the information required is generated. A view is defined as an instance of the view ontology and is described by the concepts of a query and their derivates (Section 5).

In our example, the system defines the problem over the breast cancer treatment breast_cancer_treatment as an instance of the concept Problem. The semantic of the problem is defined by expressing a relation between the problem and each query concept Vinorelbine, BreastCancer and MalignantTumor through the property *isDefinedBy*. The solution of the problem is defined by expressing a relation between the problem and each relevant resource used by the biologist through the property *isSolvedBy*.

7 Implementation and experimental validation

In order to validate our approach we conducted an implementation of a Biology virtual laboratory oriented to biomedicine resources integration. Therefore we implemented mechanisms to characterize, manipulate, and enrich the knowledge located in heterogeneous resources. The objective was to help researchers in the detection of new knowledge and resources with content related to problems.

The prototype was implemented with the JAVA platform version 1.5.0, which offers libraries for implementation of graphical interfaces, generation of graphical trees and handling of data structures. Metadata ontologies were built using Protégé ontology editor that allows designing ontologies in OWL-DL language [11]. Query processing and knowledge inferences are achieved using a Racer inference engine which uses a set of mechanisms for querying, creating, and managing knowledge bases. Our ontologies are visualized using the JUNG framework.

7.1 Resource subscription

Our system allows to subscribe new resources into the system and to characterize them. By using a graphical interface, a user specifies the URI of the resource, information related to its format and the URL of the ontology representing its semantic. Given these data, the system creates a **bio-resource** as an instance of the class Resource in the bio-resources ontology and relates it to its semantical content. Content from a resource is represented through the extraction of concepts, properties and attributes defined in the ontology associated to a resource.

The prototype implements a set of schemas representing the content of different sources. The knowledge domain of the system is defined through existing bio-ontologies: *Cell ontology, Amino-Acid ontology, Fungal anatomy ontology*, and *Clini-*

cal ontology for breast cancer. The *Cell ontology* [13] is designed to identify an structured vocabulary to characterize different types of cells, and describe approximately 680 cell types related to Plantae, Fungi, Animal and Prokaryota. The *Amino-acid ontology* [12] represents the structure and properties of different types of amino-acids, around 50 concepts. The *Fungal anatomy ontology* [9] defines a controlled vocabulary to describe the anatomy of several fungi and other microorganisms and is composed by approximatively 100 concepts. Additionally, we built a *Breast cancer ontology* characterizing the existing types of cancer, their related symptoms, stages and their possible treatments.

7.2 Explicit mapping constructor

Our prototype provides a graphical interface to the users with two lists of concepts, attributes, and properties defined within the system. Based on these lists, a researcher can specify a set of semantic correspondences between two entities specifying their type (equivalence, subtype, supertype). The system generates each one of the specified mappings as instances of MappingRelation, whose domain and range are defined by the instances of the entities involved.

For example, a biologist can define the following mappings into the system: Vinorelbine@*Source1* ≡ Navelbine@*Source1*, Cancers@*Source1* ≡ MalignantTumor@*Source3*, BreastCancer@*Source3* ⊆ Cancers@*Source1* and Cancers @*Source1* ≡ MalignantTumor@*Source3*.

7.3 Query processing

Our prototype provides the list of biological concepts stored in the laboratory as a hierarchical tree or a graph. From this list, the researcher selects the concepts representing the semantics of a biological problem and the type of search required (subclass, superclass, equivalence, predetermined). Consider a query Q defined by the concepts: Vinorelbine, Breast and Cancer and the presence of three resources defined by the following ontologies: *Clinical ontology for breast cancer* (Source1), *Amino-Acid ontology* (Source2) and *Breast Cancer ontology* (Source3).

The prototype uses SISELS to verify the consistency of the query and to recover the concepts related to the query based on existing mappings. By selecting or rejecting the retrieved concepts, the user can redefine his/her query. In our example, the query can be redefined by the concepts: Vinorelbine, Navelbine, BreastCancer and MalignantTumor. Then, the system generates a bidimensional list, where each element is composed by a biological concept and a set of resources. Our prototype counts with a virtual interface that allows biologists to filter the resources retrieved by the system and store his/her study case into the problem catalog. The problem catalog of our prototype is represented by an ontology implemented in the OWL language [11].

7.4 Experimental validation

We validate our virtual laboratory through the construction and evaluation of a set of queries composed by different number of concepts and requiring different kind of selection (equivalence and subtype). We analyze the efficiency of query processing in the prototype according to the execution time, the number of concepts retrieved after rewriting phase and the number of resources retrieved.

It is important to consider that the number of retrieved concepts after assignment phase depends strictly on the number of mappings defined in the system. For this reason, the queries used to evaluate the systems consider biological concepts defined as a generalization of one to twenty concepts. Our knowledge base was composed by 750 biological terms.

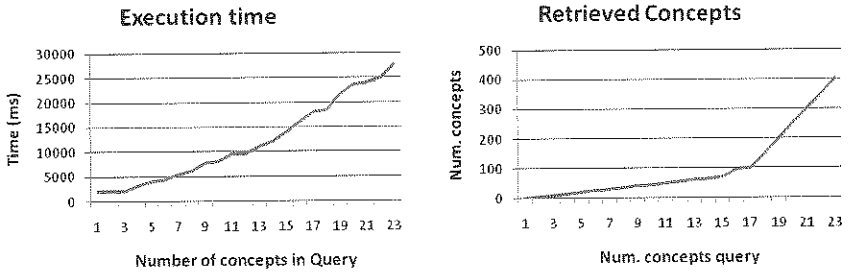

Fig. 5. Experimental results.

Figure 5a presents the execution time of query processing with respect to the analysis and rewriting phases for a set of queries defined by a set of one to twenty concepts. Figure 5b presents the number of retrieved concepts for the defined queries. The query processing execution time of in our system is reasonable when the system retrieves around 400 biological concepts. We use this result for estimating the execution time of queries over a bigger knowledge. It is of course influenced by the number of concepts in the query and the defined mappings among ontologies and it is polynomial. Finally, we consider that biologists must conduct a quantitative evaluation. We are currently working with Biology communities for validating SISELS.

8 Related Works

Centralized mediation systems like Carnot [8], SIMS [3], and TAMBIS [20] use a single ontology to model real world entities and properties. Other centralized systems like DWQ [7] and Picsel [16] use the hybrid ontology approach in order to achieve the resources integration and a conjunctive query model.

Mediation distributed systems like Observer [18] and Somewhere [1] manage multiple ontologies to integrate distributed resources. Peers are represented through an ontology which describes the content of a set of resources. Queries are expressed in terms of the ontology describing the node where it belongs and are executed locally in the

peer to retrieve data from its underlying resources. It is necessary to define the semantic correspondences between nodes for retrieving data and information stored in other nodes.

SISELS incorporates the exploitation and structural and semantical integration of resources through the use of a global schema. Inspired on description logics based knowledge representation, we express information associated to resources and queries using description logics.

9 Conclusions

SISELS is a mediation system that integrates resources (documents, applications and data bases). This system provides scientists transparent access to distributed information satisfying their requirements to study specific problems. As a result of structural and semantical characterization of resources, SISELS offers tools for classifying and semantically integrating resources to exploit biological information. SISELS maintains a knowledge representation of resources that describes their content structurally and semantically.

The knowledge domain in SISELS is defined by a set of biological concepts defined by experts and is enriched when a new resource is added, new knowledge is generated or a new problem is defined. Once the knowledge domain is defined, it is necessary to establish a relation between concepts through mappings. SISELS proposes techniques for query processing techniques bases on views. These views adapt to the requirements of a group of experts in order to study a Biology problem. Based on views, SISELS provides transparent access to biological resources and promotes information sharing between communities.

Future work relays on the definition of strategies and knowledge rules for view management and query processing by using inference mechanisms. This would reduce the computational cost related to the management of large amounts of data. We are currently defining techniques to prove the consistency of biological terms. This can be achieved through the construction of a knowledge base that contains a set of rules describing the principles of Biology. Finally, it is important to consider that reasoning is computational expensive especially when we deal with vast amounts of information. For this reason, it would be interesting to explore possible optimization strategies to reduce these computational costs.

Perspectives related to the implementation of our prototype include to stabilize and to validate the virtual laboratory with research groups in the biological area. It is necessary to formalize strategies for view management in order to satisfy in an optimal way the requirements of a group of experts and to define existing relations between different views. Consequently, we could use SISELS to define applications in other scientific areas like astronomy. We are currently collaborating in e-Grov project to give access to astronomical resources.

References

1. P. Adjiman, P. Chatalic, F. Goasdoué, M.C. Rousset and L.Simon (2004). Distributed reasoning in a peer-to-peer setting European Conference on Artifical Intelligence.

2. G. Antoniou and F. van Harmelen (2004). A Semantic Web Primer, chapter The Semantic Web vision. MIT Press.
3. Y. Arens, C.N. Hsu, C.A. Knoblock (1997). Query processing in the SIMS information mediator, Readings in Agents.
4. F. Baader, D. McGuinness, D. Nardi, and P. Patel-Schneider (2003). The description logics handbook: Theory, implementation and applications, chapter An introduction to Description Logics. Cambridge University Press.
5. G. Bruno (2006). "ADEMS" a knowledge-based service for intelligent mediator configuration. PhD thesis, Institut National Polytechnique de Grenoble.
6. D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi and R. Rosati (1998). Information integration: Conceptual modeling and reasoning support Conference on Cooperative Information Systems.
7. C. Collet, M.N. Huhns and W.M. Shen (1991). Resource integration using a large knowledge based in carnot, Building large knowledge based Computer IEEE, 24-12. Chs91
8. M. Constanzo (2005). Fungal anatomy ontology. Available at: http://www.yeastgenome.org/fungi/fungalanatomyontology/.
9. R. Davis, H. Shrobe, and P.Szolovits (1993). What is knowledge representation? AI Magazine, 14(1):17–33.
10. L. Deborah and H.Frank (2004). OWL web ontology language overview. Available at: http://www.w3.org/TR/owl-features/.
11. N. Drummond, G. Moulton, R. Stevens, and P. Lord (2005). Amino-acid ontology. http://www.co-ode.org/ontologies/amino-acid/2005/10/11/.
12. O. Hoffman. Cell ontology (2007). http://www.berkeleybop.org/ontologies/.
13. T. Kirk, A.Y. Levy, Y. Sagiv, D. Srivastava (1995). The Information Manifold Information Gathering from Heterogeneous, Distributed Enviroments.
14. D.B. Lenat and R.V. Guna (1990). Building large knowledge based systems Representation and Reasoning in the Cyc Project, Addison-Wesley, Reading, Massachusetts, United States.
15. A.Y. Levy (2000). Queries using views: A Survey Technical Report: Computer Science Dept., Washington, Univ.
16. A.Y. Levy and M.C. Rousset (1998). Combining horn rules and description logics in carin, Artificial Intelligence 104-12.
17. E. Mena, V. Kashyap, A.P. Sheth, A. Illarramendi (1996). OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies Conference on Cooperative Information Systems.mksi96
18. M.T. Ozsu and P. Valduriez (1999). Principles of distributed database systems. Prentice Hall, second edition.
19. N. Paton, R. Stevens, P. Baker, C. Goble, S. Bechoofer, A. Brass (1999). Query processing in the Tambis bioinformatics source integration system SSDBM '99: Proceedings of the 11th International Conference on Scientific and Statistical Database Managment, IEEE Computer Society.
20. Racer Systems (2004). Racer reference manual version 1.7.19.