

Query Answering Using Ontologies in Agent-based Resource Sharing Environment for Biological Web Information Integrating

Jiann-Jyh Lu and Chun-Nan Hsu

Institute of Information Science, Academia Sinica

Nankang, Taipei, Taiwan

{jjlu, chunnan}@iis.sinica.edu.tw

Abstract

A variety of biological data is transferred and exchanged in overwhelming volumes on the World Wide Web. How to rapidly capture, utilize and integrate hundreds of biological databases on the Web is one of the most critical issues in bioinformatics. In this paper, we describe our preliminary study to address this issue in an agent-based resource-sharing environment, where we apply the Web agent technology and DAML+OIL ontologies of biological knowledge to enable biologists to make *concept-to-concept* query. Given a query, the agent coordination service in this environment will search for a set of suitable paths in the ontologies that links the query concept to the goal concept. The Web agents will be launched to extract the data from the target Web resources and transfer the data according to the found path to answer the query. This approach makes the details of querying Web resources transparent to the biologists and allows them to query biological resources on the Web in the terms of their domain knowledge.

Keywords

Software agents, ontology, DAML+OIL, Semantic Web, bioinformatics

1 Introduction

With the recent advance in bio-technologies, numerous genome and proteome databases from various biological communities have been developed to assist in genomic research. There are hundreds of genomic and biological databases open for public access throughout the World Wide Web. In 2002, journal *Nucleic Acids Research* [Baxevanis, 2002] reports 335 important web-based public domain biological Web databases, jumping from 281 in 2001. Their figure in 2003 is about 386 [Baxevanis, 2003]. The total number of available Web biological databases is estimated at about 600+. Their data size is

also increasing at a tremendous rate. Data management and data integration are keys to successfully maximize the utility of these biological databases.

An urgent need in bioinformatics is the discovery of suitable resources and the marshalling of those resources to work together to perform a task. The first one of the two key challenging issues is how to integrate the data from the wide-spread resources for the use of biotech researchers for a variety of purposes. The other issue is how to apply computational techniques to analyze data and knowledge resources in order to answer complex biological queries. The information integration system described in this paper can provide researchers a solution to address these two issues. Usually, a nontrivial query in biological research is so complex that we have to decompose the query into several less complex subqueries to answer the entire query. To handle these subqueries systematically, technologies are ready for biologists around the world to delegate all of these subqueries to the agents with different functions, including accessing online databases, making a data query, submitting request to analysis tools and organizing analysis results into a report.

In this paper, we present our preliminary study of answering biological queries using biological Web resources. We combine the wrapper agent technologies and ontologies of molecular biology to enable biologists to issue *concept-to-concept* query. For example, biologists can issue a query to find a set of gene sequences given an organism. "Gene sequences" and "organism" are concepts in biology. We refer to the former as a *goal concept* and the latter as a *query concept*. The system will search for a set of suitable paths that links the query concept to the goal concept. The Web agents will be launched to extract the data from the target Web resources and transfer the data between agents according to the found path. Since the knowledge base makes the details of querying a Web resource transparent to the biologists and allows them to query the knowledge base in the terms of their domain knowledge, biologists will be able to utilize the Web resources more easily and efficiently.

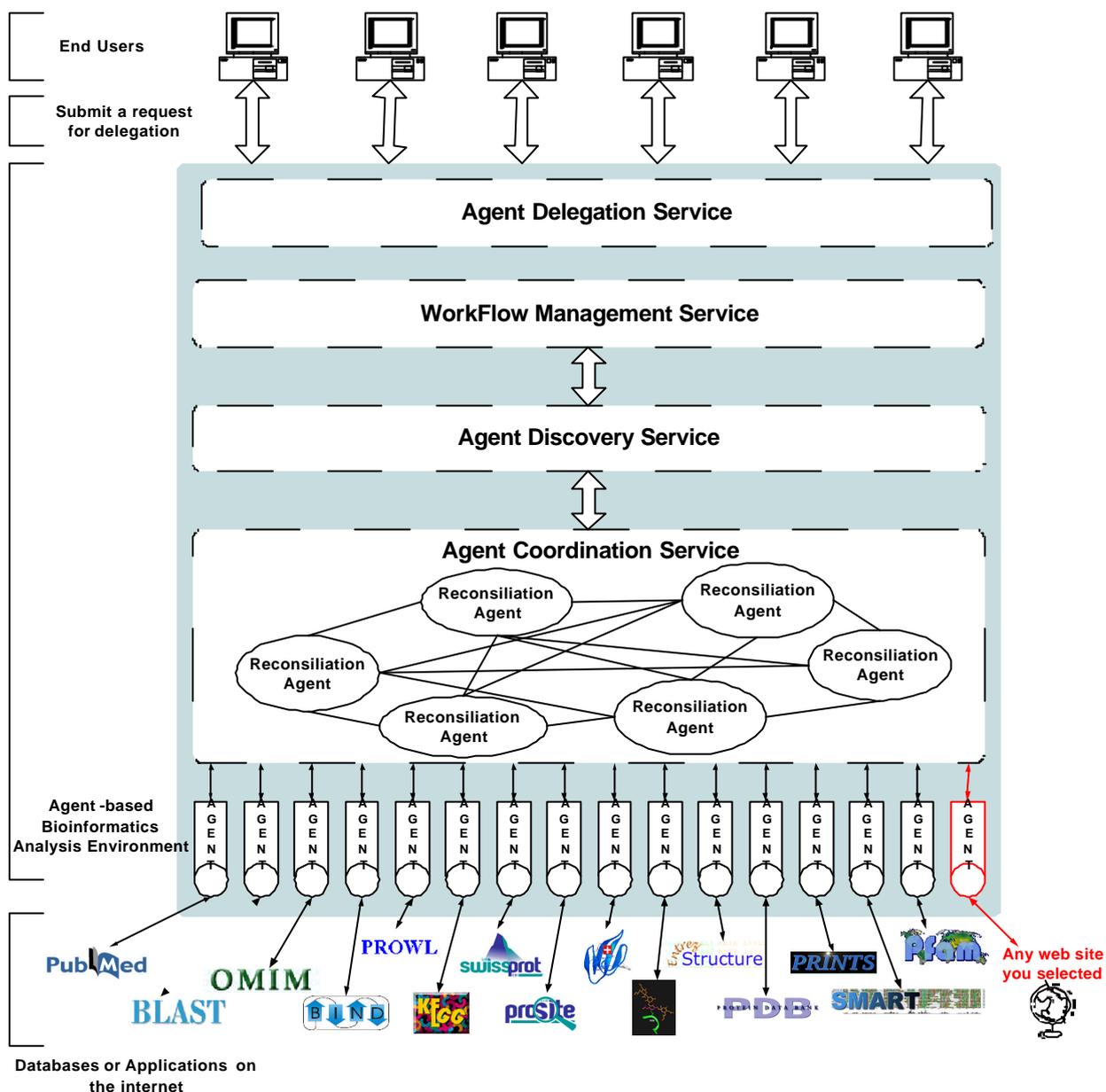


Figure 1. Architecture of agent-based resource environment for protein knowledge services

The proposed idea in this paper is part of the *Protein knowledge services* project currently conducted at Academia Sinica, Taiwan. This project is aimed at building an *agent-based resource sharing environment* with a scalable and extensible open platform (see Figure 1), built using the technologies from information integration agents, Web services, the distributed computing techniques and metadata service via ontology for data and application services interoperability. Because the data generated by one agent could be the

input to another agent in the workflow process, we have to ensure that the semantic type of the data (e.g., a protein sequence) matches the agent's input type. To serve this purpose, we are designing a set of ontologies to describe the metadata of the agents, which covers the data type of inputs and outputs and the information about which agent should be delegated, where to allocate agents, how to compose the workflow and how to interoperate data generated by agents in a bioinformatics setting.

The original contributions of the idea in this work include that we use ontologies to model metadata of the agents for query answering rather than modeling the data sources, as in previous work in information integration for structured databases. This is critical in integrating Web resources because it is relatively easy and frequent that a Web resource is changed. When that happens, in previous work, wrapper agents as well as the model of data sources need to be updated to keep the system working. By contrast, in our proposed agent-based resource-sharing environment, we only need to maintain the Web agents without the need to change the ontologies.

2 Related Work

Heterogeneity of databases and software resources continues to impede the integration of biological information. Researchers in computer science, in particular, those in databases and artificial intelligence communities, have studied this problem for decades and achieved some solid results. In recent years, the problem of information integration has received considerable attention due to the growing number of information sources available on the Internet. Among different approaches, one sees various levels of homogenization and centralization. At one extreme is by hyper-linking separately developed Web pages, but this approach is limited in terms of user interface capability and query power. At the other extreme is by data warehousing. However, since a data warehouse requires a standard data definition schema of all integrated data sources, it may not be appropriate for integrating biological databases where databases are maintained and developed independently and it is practically impossible to provide such a schema. A middle road is to apply the notion of information mediator [Wiederhold, 1992] by defining a “*domain model*” that maps user queries to underlying heterogeneous sources. The domain model is designed for each application domain. For each data source, there is a “*source model*” as well as a “*wrapper agent*” that wraps the data source so that details of how to access the data source, such as its location, data definition schema, query languages, etc., can be encapsulated from users. Important results include SIMS [Knoblock et al., 1994], Information Manifold [Kirk et al., 1995], TSIMMIS [Chawathe et al. 1995], Infomaster [Duschka and Genesereth, 1997], etc. Their efforts are mainly focused on query planning, query reformulation, and query optimization. These approaches work well for integrating structured and semi-structured databases, and have been deployed in many real-world applications, such as in e-commerce. However, to integrate Web-based data sources, including those for genetics and molecular biology, we still need to program wrapper agents for each of the Web data sources, which now can be generated from user-labeled extraction examples (see, e.g., [Hsu and Dong, 1998].)

Meanwhile, in bioinformatics, researchers have adopted a variety of information integration techniques and specialized these techniques for different applications in genetics and molecular biology [Letovsky, 1999]. Important examples include SRS, ISYS, iProClass, IBM DiscoveryLink and MGI etc. Among them, one of the most widely applied solutions is SRS of Lion Bioscience. Initially developed at European Bioinformatics Institute (EBI), SRS [Carter et al., 1998] supports and integrates many famous and useful databases. But when a biologist wants to integrate their a new database (e.g., their own experiment data) into SRS, they must program in a specialized language called “Icarus” to parse their own data source. Other biological databases “integrate” external Web-based databases by hyper-linking. More advanced data integration systems depend on “server proxy” or “mirror sites” to access external data. They also depend on hand-coded specialized wrapper software to each connected data sources. This limitation has several drawbacks, in particular they are difficult to extend, migrate, maintain, or scale up.

One of the distinguished project aimed at fulfilling the biologists’ need in Web data integration is myGrid [MyGrid, 2002]. myGrid is a multi-organisational project funded by the EPSRC as part of the UK Research Councils e-Science programme and aims to develop the necessary infrastructural middleware that operates over an existing Web services, Grid infrastructure and the Semantic Web to support scientists in making use of complex distributed resources. However, the myGrid will succeed only if the community Grid-enables its applications and databases. Although the pace of constructing Grid infrastructure is accelerating and the impetus is overwhelming, it is still premature now.

In our project, we propose an alternative solution that is workable but much simpler than myGrid. We plan to design the agent delegation service, workflow management service, the agent coordination service and agent discovery service to find the appropriate task-specific distributed agents over the Internet and then and coordinate them as a workflow pipeline by using the same standard of Web services. We are also designing a set of ontologies, like myGrid, to describe the metadata of the agents, including the definition, the scope of the task assigned, the semantic type of the input and output data, the condition for delegation, and where and how to discover, etc. This paper will focus on our preliminary study on the query-answering algorithm in this project.

3 Background

In this section, we review the technologies applied in this work.

3.1 Web Agent

A web agent is a software agent that delegates its user to browse the Web and return the extracted data in structured, machine-interpretable formats. Previously, we

have developed technologies required to practically generate and maintain a large number (up to thousands) of Web agents for the existing Web, that is, the Web that mainly constitutes of HTML/XML Web pages.

The technologies of Web agents can be divided into two parts: controlling Web browsing sessions and content extraction from Web pages. State-of-the-art Web agent can be trained to simulate any possible Web browsing session but data extraction can only be applied to Web pages with structured layout. For more details see [Hsu and Dong 1998; Hsu et al. 2003].

3.2 DAML+OIL

[DAML+OIL, 2001] is an ontology language specifically designed for the Web, and is the basis of the W3C Semantic Web Activity WebOnt standardization process. The notion of the Semantic Web [Berners-Lee et al., 2001; Hendler, 2002], which aims to move from syntactic interoperability to semantic interoperability, relies on machine interpretable semantic descriptions. Thus, DAML+OIL builds upon existing Web standards, such as XML and RDF, and is underpinned by an expressive Description Logic (DL). Formal semantics enables machine interpretability and reasoning support as well as human communication --- an aim of ontological description.

DAML+OIL takes an object-oriented approach, with the structure of the domain being described in terms of *classes* and *properties*. An ontology consists of a set of *axioms* that assert, e.g., subsumption relationships between classes or properties. DAML+OIL also supports the full range of XML Schema datatypes. Overall, it has a model theoretic semantics, and a rich set of constraints available for class descriptions.

In this paper, we use an ontology-building tool called OilEd [Sean et al., 2001] to construct our ontologies of the metadata of the agents and biological knowledge in DAML+OIL.

3.3 Ontologies in Bioinformatics

Ontologies are useful in intelligent information integration. In AI, ontologies were developed for the following purposes: knowledge sharing and reuse, data exchange among programs, unification of disparate data, knowledge representations, and knowledge-based services etc. They provide a shared and common structure of a domain thus giving a common understanding of this domain, and may be used for overcoming semantic heterogeneity. Recently, many ontologies for knowledge representation in the biological domains have been proposed [Robert *et al.*, 2000; Nataliya, 2001]. Well-known bio-ontologies include:

- The TAMBIS Ontology (TaO)
- The EcoCyc Ontology
- The Gene Ontology (GO)
- The RiboWeb Ontology

- The Schulze-Kremer ontology for molecular biology (MBO)

All these ontologies are very different and specific to their intended use. Among them, TaO is an ontology designed especially for data integration. EcoCyc is an ontology of *ecoli*. GO is an ontology of gene product function and RiboWeb represents knowledge of Ribosomal subunit structures, data and methodologies. They are, however, not quite reusable because of the diversity of their representation forms, the explicitness of their semantics and the range of the applications they address. In this work, we design our own ontologies based on these ontologies.

4 Agent-based Resource Sharing Environment

In this section, we briefly describe the agent-based resource-sharing environment as illustrated in Figure 1. In this environment, every user can set one's own personal preference configuration, such as data mining tools, databases and parameters for specific analysis, in the system in order to assign delegation to the agents and prepare the workflow. After delegation and workflow preparation, the system allocates the suitable agents over the Internet through the agent discovery service. Then, the agents (the last row in Figure 1) are deployed to perform their tasks by accessing their designated Web resources. Web agents can cooperate through ontologies and the reconciliation agents assigned in the user-specific workflow process. Details will be described in Section 5.

The agent-based resource-sharing environment is an extensible open platform for data and application services interoperability. The agents can talk to each other as well as other applications in the same Web service infrastructure. Web services [Web services, 2002] is a collection of XML-based technologies and has emerged as a set of open standards, including SOAP (Simple Object Access Protocol), WSDL (Web Services Description Language) and UDDI (Universal Description, Discovery, and Integration), to address agent communication issues such as service discovery, interoperability, data exchange and business processes. Besides Web services, there is a set of ontologies as Semantic services to describe the metadata of the agents for agent delegation, agent discovery, workflow organization and integrating data generated by agents.

5 Query Answering

In this section, we discuss how to answer a query in the agent-based resource-sharing environment. Initially, we need ontologies to provide semantic services required for query answering:

- Domain knowledge in molecular biology;
- Agent classifications and relationship;

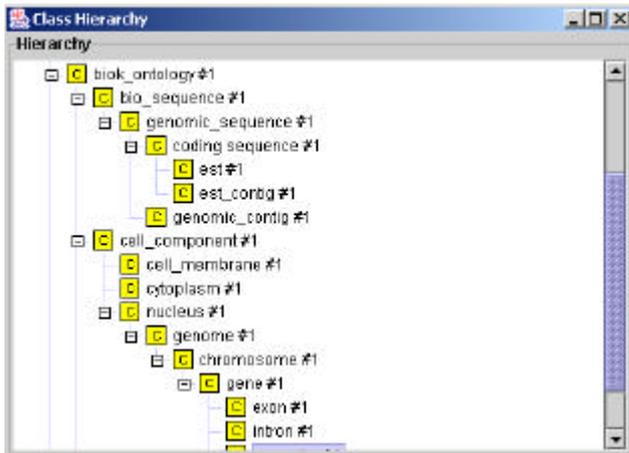


Figure 2. A fragment of our biological knowledge ontology implemented by OilEd.

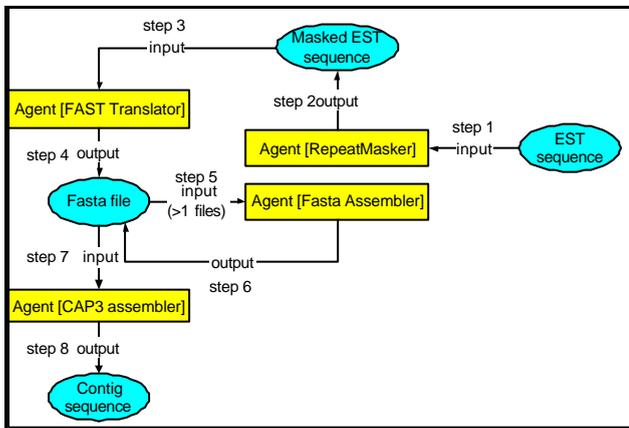


Figure 3. An example of agent coordination in the simplified EST Clustering analysis.

- Agent descriptions, definitions of the task assigned and semantic type of the input and output data;
- The metadata for agent discovery, agent delegation, agent coordination and workflow preparation.

Figure 2 shows a fragment of our ontology for molecular biology.

Now that the ontology for biological knowledge is given, users can issue a *concept-to-concept query* in biology terms. Consider the following example. This example is about a simplified EST clustering analysis, illustrated in Figure 3. Expressed sequence tags (ESTs) are short nucleotide sequences that are considered as a shortcut to the alternative spliced, expressed forms of the genes. Clustering related ESTs may provide invaluable hints for interpreting genome sequences. Conceptually, this can be accomplished by issuing a query to obtain the

“EST contig” given a set of “EST.” In this case, the query concept is “EST” and the goal concept “EST contig.” Both terms can be found in the ontology.

To answer this query using Web agents requires a workflow where the data generated by one agent serve as the input to another agent. The ontologies of the agents allow us to ensure that the semantic type of the data (e.g., a protein sequence) matches the data receiving agent’s input type. The problem of query answering can be cast as the problem of search for a path from the query concept to the goal concept. The path consists of a sequence of agents starting from an agent whose input type is the query concept and ending with an agent whose output type is the goal concept. The agents in between must match their input types and output types to form an executable workflow path.

Back to our EST clustering example. Four agents have to coordinate each other as a workflow to answer this query. At the first, “EST sequence” is submitted into “Agent [RepeatMasker]” as the input. Then the output of the “Agent [RepeatMasker]”, “Masked EST sequence”, is submitted into “Agent [FAST Translator]” for changing data format into fasta. Then, several Fasta files as input are submitted into “Agent [Fasta Assembler]” to combine into a big fasta file, which contains several EST sequences. Finally this big fasta file as input is submitted into “Agent [CAP3 assembler]” to assemble those EST sequences into a EST contig, and thus the query answering workflow is completed.

In addition to the data types of the input and output, there are other constraints for applying a Web agent and path searching algorithms may not be applicable in those cases. We may apply other query planning algorithms (e.g., [Kirk et al., 1995; Knoblock et al., 1994]) that allow for more expressive definitions of the goal and operators in the agent coordination service.

6 Discussion and Future Work

Capability to quickly collect and integrate biological data available on the Web is critical for highly competitive biotech research. In this paper, we present our idea of query answering service in an agent-based resource-sharing environment. We are currently working on implementing the prototype of this environment. In this environment, administrators can manage hundreds or more software agents that regularly collect data from remote external biological databases. In this environment, agents can also be invoked on demand to perform on-line, exploratory data collection. During the execution of the agent-based environment, loads of databases will be collected from the Web resources and are up to the application software systems to utilize. A set of biological-task oriented ontologies should describe the metadata of the agents to guide the delegation service, workflow management service, agent discovery service, and agent coordination service. We will apply the Web

services standards as the infrastructure for agent communication so that we can scale up and maximize the utility of our Web agents for widely distributed users. In this environment, users can perform the following six tasks:

- Locate an agent service, through Web Service protocol UDDI.
- Configure (personalize) an agent service
- Invoke an agent service and execute it either remotely (through SOAP protocol) or locally (through JAVA applet or other protocols)
- Save a configured agent service locally in a personal repository
- Register a new agent service to a directory of service so that other users can share that new agent service
- Provide “provenance” service, i.e., the environment will inform the users with the latest updates or changes to the registered agent services

We also plan to build more example bioinformatics applications using the resulting prototype to actually assist biologists to perform real research.

Acknowledgments

We wish to thank Rita Huang and biomedical scientists at the Institute of Biomedical Science, Academia Sinica for their helpful comments on this work.

References

- [Baxevanis, 2002] Andreas D. Baxevanis. The Molecular Biology Database Collection: 2002 update. *Nucleic Acids Research*, Vol. 30, No. 1 pp. 1-12, 2002.
- [Baxevanis, 2003] Andreas D. Baxevanis. The Molecular Biology Database Collection: 2003 update. *Nucleic Acids Research*, Vol. 31, No. 1 pp. 1-12, 2003.
- [Berners-Lee et al., 2001] Tim Berners-Lee, James Hendler, O. Lassila, The Semantic Web, *Scientific American*, May, 2001.
- [Carter et al. 1998] P. Carter, C. Coupaye, D.P. Kreil, T. Etzold. Analysing and Using Data from heterogeneous Textual Databanks with SRS. In S. Letovsky (ed.), *Molecular Biology Databases*. Kluwer Academic Press, 1998. Also SRS web site: <http://srs.ebi.ac.uk>.
- [Chawathe et al. 1995] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS project: Integration of heterogeneous information sources. In *Proceedings of IPSJ Conference, Tokyo, Japan*, pages 7--18, October 1995.
- [DAML+OIL, 2001] DAML+OIL Reference Description, 2001. (<http://www.w3.org/TR/daml+oil-reference>)
- [Duschka and Genesereth, 1997] Oliver M. Duschka and Michael R. Genesereth. Querying planning in infomaster. In *Proceedings of the 1997 ACM Symposium on Applied Computing*, San Jose, CA, February 1997.
- [Hendler, 2002] James Hendler. Agents and the Semantic Web. *IEEE Intelligent Systems*. pp.30-37, 2002.
- [Hsu and Dong, 1998] Chun-Nan Hsu and Ming-Tsong Dong, Generating finite-state transducers for semi-structured data extraction from the web, *Journal of Information Systems*, Special Issue on Semi-structured Data. Volume 2, Number 8, 1998.
- [Hsu et al. 2003] Chun-Nan Hsu, Chia-Hui Chang, Harianto Siek, Jiann-Jyh Lu and Jen-Jie Chiou, Reconfigurable Web Wrapper Agents for Web Information Integration. In *Proceedings of IJCAI-2003 Workshop on Web Information Integration*, 2003.
- [Kirk et al. 1995] T. Kirk, A. Y. Levy, Y. Sagiv, and D. Srivastava. The information manifold. In *Proceedings of the AAAI Spring Symposium on Information Gathering in Distributed Heterogeneous Environments*, Stanford, California, March 1995.
- [Knoblock et al. 1994] C. A. Knoblock, Y. Arens, and C. N. Hsu. Cooperating agents for information retrieval. In *Proceedings of the Second International Conference on Cooperative Information Systems*, Toronto, Ontario, Canada. University of Toronto Press, 1994.
- [Letovsky, 1999] S. Letovsky, editor. *Bioinformatics: Databases and Systems*. Kluwer, Norwell, MA, USA, 1999.
- [MyGrid, 2002] The myGrid project, 2002. (<http://www.mygrid.info>)
- [Nataliya, 2001] Nataliya Sklyar. Survey of existing bio-ontologies. Institute for informatik, 2001.
- [Robert et al., 2000] Robert Stevens, Carole Goble and Sean Bechhofer. Ontology-based Knowledge Representation for Bioinformatics. *Briefings in Bioinformatics*, 2000.
- [Sean et al., 2001] Sean Bechhofer, Ian Horrocks, Carole Goble and Robert Stevens. OilEd: a Reason-able Ontology Editor for the Semantic Web. To appear in *24th German / 9th Austiran Conference on Artificial Intelligence*, 2001.
- [Web Services, 2002] Web Services Activities. 2002. (<http://www.w3.org/2002/ws/>)
- [Wiederhold 1992] Gio Wiederhold. Mediators in the Architecture of Future Information Systems. *IEEE Computer*, March, 1992.