A Service-Oriented Information Sources Database for the Biological Sciences

Gordon K. Springer
Department of Computer Science
University of Missouri-Columbia
108 Mathematical Sciences Building
Columbia, MO 65211 USA
springer@csdeca.cs.missouri.edu

Timothy B. Patrick
Medical Informatics Group, School of Medicine
University of Missouri-Columbia
605 Lewis Hall
Columbia, MO 65211 USA
patrick@csdeca.cs.missouri.edu

Abstract

Researchers in the biological sciences require access to a variety of information sources located in various places on different computer networks. In order to satisfy the information needs of a researcher, appropriate information sources must be selected and access to these information sources and the computing services supporting them must be provided in a way that does not distract the researcher from problems of real interest. At the University of Missouri-Columbia a service-oriented information sources database is being developed as a key component of a layered-model design of an intelligent system which will provide a research environment appropriate to the needs of researchers in the biological sciences.

Introduction

One of the difficulties for today's researchers in the biological sciences is gaining access to a variety of information sources that are appropriate for carrying out their research activities.[1] There is an ever increasing number of public and private information sources that are available on a variety of accessible computer networks. However, while these sources are available, it is difficult for a researcher to keep up with the knowledge of what sources are available, where they are located and how to access them to produce fruitful results.

The problem, in general, is that a researcher's information needs may be satisfied by selecting and accessing the information sources relevant to these needs. However, handling details of accessing these information sources and the computing services supporting them may constitute a significant distraction to the researcher. What is needed is a way of mapping the researcher's information needs into requests for access to relevant information sources and supporting computing services. At the same time the researcher should be shielded from the details of accessing these information sources and supporting computing services. This can be accomplished by using a layered model as shown in Figure 1.

In the model shown in Figure 1, the researcher's information need (e.g., DNA sequence analysis) is logically satisfied when it is mapped to an appropriate information source (e.g., FASTA (Pearson & Lipman 1988) to GenBank) and access is provided to that information source. That information source is, however, supported by some computing service or services. Thus, in order to access that information source, the researcher's

[1]This publication was supported in part by grant numbers LM05513 and LM07089 from the National Library of Medicine. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Library of Medicine.

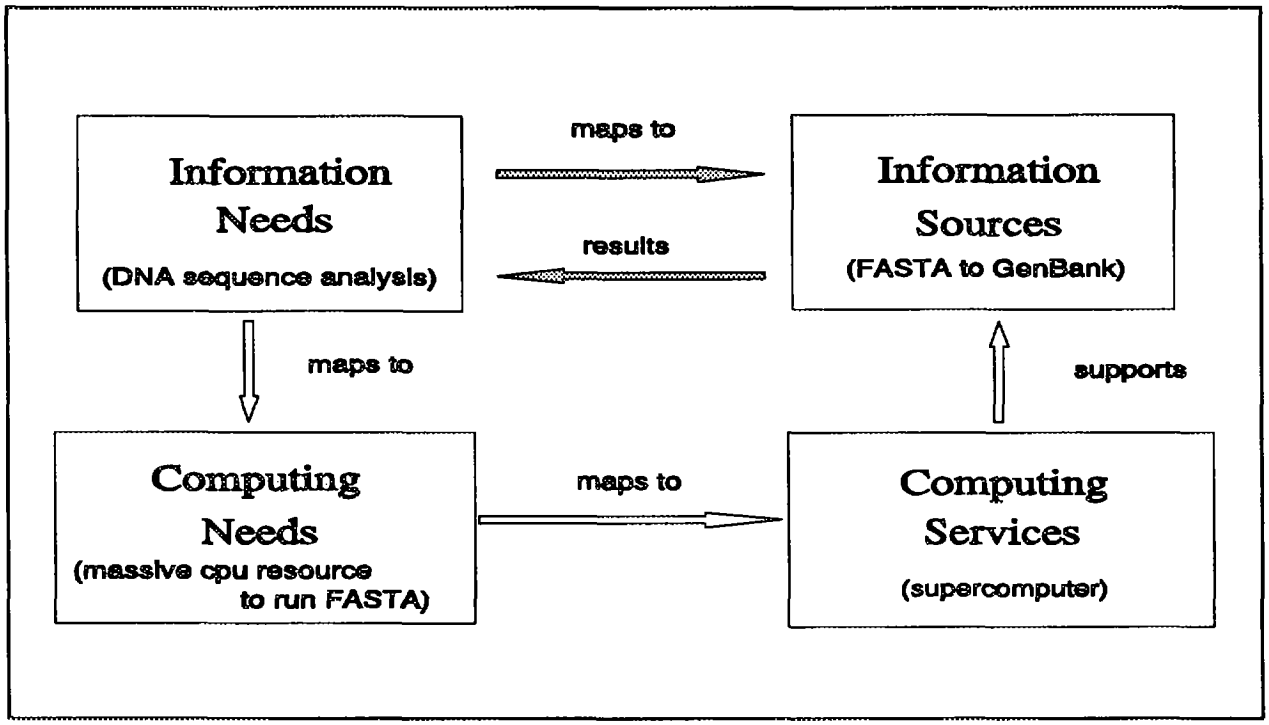

Figure 1 Layered-Model of Information Source Access

information need must be mapped to an appropriate computing need (e.g., use of a massive cpu resource to run the FASTA program against the GenBank database) and a physical connection must be established between the researcher's local computing service (e.g., the researcher's pc or workstation) and a network accessible computing service (e.g., a supercomputer) supporting the required information source. That is, the researcher's information need is turned into a request that is conveyed to the relevant information source across a computer network and the results of that request are returned to the researcher without the researcher having to explicitly know where the computing services supporting the information source are located or how to negotiate the network to access them. Clearly not all uses of FASTA to search GenBank require supercomputer resources, but some requests would be appropriately serviced in this way in order to provide timely and responsive results to the researcher. An essential element of this approach includes determining what service is appropriate to satisfy a particular request.

A project has been initiated at the University of Missouri-Columbia to develop an intelligent system based on the layered-model depicted in Figure 1 which will provide a research environment for the biological sciences. Key to this effort is the design of an information sources database. This database contains the information on what various information sources contain, how to properly request information from these information sources, and sufficient data to be able to locate the computing services supporting these information sources in the network. In addition, if a particular information source is replicated and accessible in the network, the system will dynamically determine which currently available service is best prepared to service the request.

Four Problems

The general problem is that a researcher's information needs may be satisfied by selecting and accessing the information sources relevant to these needs. In addition, the researcher must be shielded from the details of accessing these information sources and the computing services that support them. This general problem may be broken down into at least four sub-problems. The first three problems- the *relevance/quality filtering*

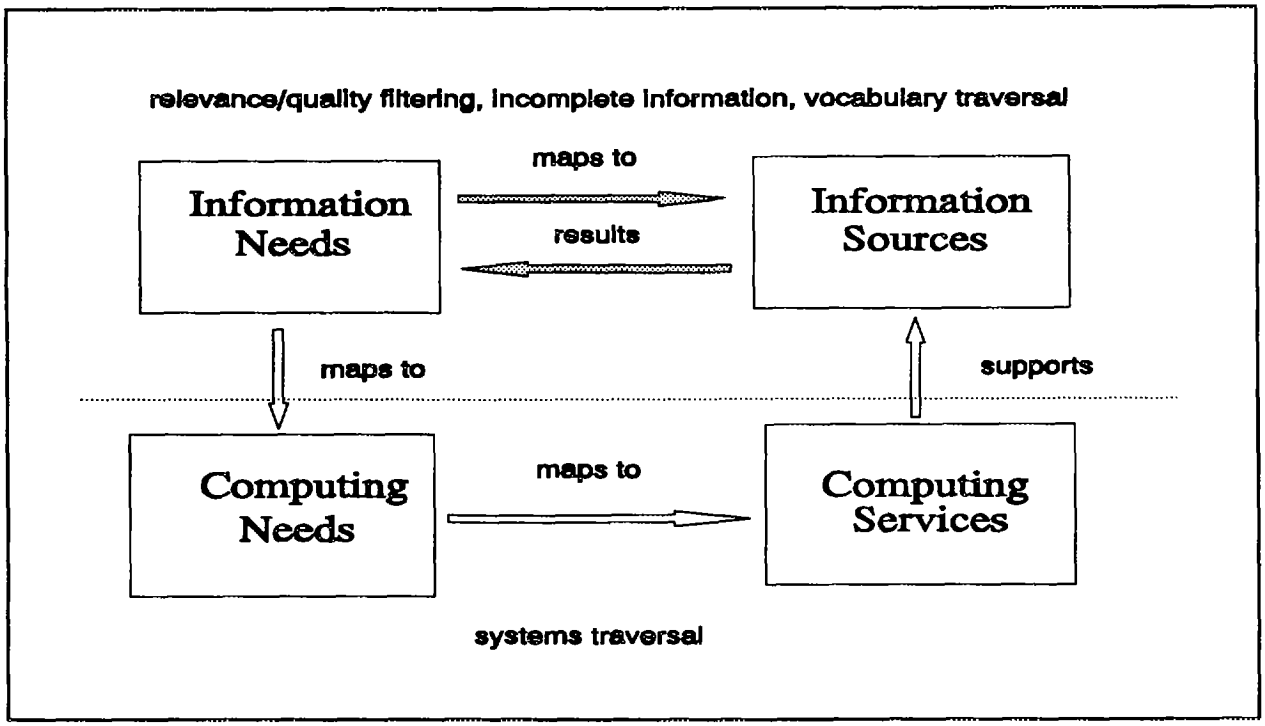

Figure 2 Four Problems and the Layered-Model

problem, the *incomplete information problem*, and the *vocabulary traversal problem*- are information level problems. They concern the intelligent selection of and the providing of logical access to the required information sources, as depicted in the upper portion of Figure 2. The fourth problem- the *systems traversal problem*- depicted in the lower half of Figure 2, is a computing and networking level problem. It concerns the intelligent selection of and the providing of transparent access to the computing services which support the required information sources.

The relevance/quality filtering problem arises from the difficulty of filtering out information most relevant to a given query from an abundance of information and an abundance of information sources available. The incomplete information problem arises due to researchers being unaware of information sources or which may be hidden from them. The vocabulary traversal problem arises due to a lack of standardized vocabularies between and among various information sources. Different specialized vocabularies may have different syntactic, hierarchical, syndetic and semantic properties, and translating between the vocabularies is a non-trivial task. The systems traversal problem stems from information sources being scattered around heterogenous computer networks which requires accessibility across multiple computer systems and networks. Any one of these problems alone may constitute a significant impediment to fruitful research. Certainly in combination they may be expected to divert a researcher's attention away from the problems of real interest.

The MMBIR Information Sources Database

The Missouri Molecular Biology Information Resource (MMBIR) (Springer, Loch, & Patrick 1992) was designed to address the systems traversal problem. Current research (National Library Of Medicine grant LM05513, Gordon K. Springer PI) is focused on developing the MMBIR into an intelligent system capable of addressing not only the systems traversal problem but also the problems of relevance/quality filtering, incomplete information, and vocabulary traversal. The MMBIR Information Sources Database

Information Source	Contents of Results	Location Annotation	Method of Access	Input/Output Vocabulary	Related Sources
FASTA to GenBank	evolutionary distance between an unknown sequence and known sequences	uuid of server offering the service	technical characteristics of server	(1) GenBank Taxonomy; (2) GenBank Keywords	(1) FASTA to EMBL; (2) Grateful Med to Medline
Quickscan to GenBank	sequences associated with author, title, or organism type	uuid of server offering the service	technical characteristics of server	(1) full text; (2) GenBank Taxonomy; (3) GenBank Keywords	Grateful Med to Medline
GCG Pileup	evolutionary distance between selected sequences	uuid of server offering the service	technical characteristics of server	*	(1) FASTA to GenBank; (2) Grateful Med to Medline

Figure 3 Sample MMBIR/ISDB Entries

(MMBIR/ISDB) is being developed as a key component of the evolving MMBIR system.

In order to most effectively address the information needs of researchers, the MMBIR/ISDB will be a service-oriented database. The database attempts to map a researcher's need for information to services that can address these needs. This mapping not only identifies relevant programs and databases, but also includes the locating of and the accessing of supporting computing services on the network. In respect to its emphasis on services, the MMBIR/ISDB differs from other information sources databases such as the Unified Medical Language System (UMLS) Information Sources Map (Humphreys & Lindberg 1992) and the LiMB (Listing of Molecular Biology Databases) database (Keen et al. 1992), which have tended to index factual or bibliographic databases themselves, rather than the programs and servers that provide access to those databases.

Figure 3 depicts a portion of three sample entries in the MMBIR/ISDB. The database entries include fields which identify the services which are available and additional information which can be utilized to map a researcher's information needs to an information source that can service these needs. Directly or indirectly these fields support the solution to the four sub-problems discussed earlier.

The relevance/quality filtering problem is addressed by the *contents of results* field (based upon some suitable standard vocabulary such as UMLS meta-concepts (Humphreys & Lindberg 1992)) and the *related sources* field. Using a knowledge base and inference engine located at the user's local workstation, or directly accessible from it on the network, a special expert system or Knowbot (Anthes 1991, Daviss 1991, Markoff 1990) will both (1) search the MMBIR/ISDB for the information sources deemed to be most relevant to the user's information needs and (2) examine the contents of any results returned to determine the further relevance of other information sources.

The tabular design of the MMBIR/ISDB addresses the incomplete information problem. The database is volatile and extensible. It supports either manual or automatic addition of entries for newly available or newly discovered information sources. Similarly, it provides for the elimination of entries for information sources that disappear or

cease to exist. Although the initial MMBIR/ISDB will contain entries for well-established information sources, the design will accommodate the activities of MMBIR explorer-Knowbots which search for and identify new information sources in the network.

The *input/output vocabulary* field of the MMBIR/ISDB, and the use of a suitable standard vocabulary in the *contents of results* field, will address the vocabulary traversal problem. It should be noted that these fields merely indicate the vocabularies that must be intertranslated and do not themselves directly support such translation. Thus, in order to most effectively address the vocabulary traversal problem, the MMBIR system will require some additional translation aid such as the UMLS Metathesaurus (Humphreys & Lindberg 1992).

The *location annotation* field and the *method of access* field of the MMBIR/ISDB will provide the MMBIR sufficient information to address the systems traversal problem. The MMBIR is a client/server system based on the remote procedure call (RPC) model. Currently, the MMBIR utilizes HP/Apollo Network Computing System (Kong et al. 1990) clients and servers (Springer, Loch, & Patrick 1992). Future development of the MMBIR calls for migration of the system to the Open Software Foundation's Distributed Computing Environment (OSF DCE) clients and servers (OSF 1992). Since the RPC component of DCE is based upon the HP/Apollo RPC model, this migration can be readily accomplished (Rosenberry, Kenney, & Fisher 1992). Each entry in the MMBIR/ISDB will be associated, either directly or indirectly, with an NCS or DCE server. In Figure 3, the location annotations are universal unique identifiers (uuids) used by broker daemons (NCS) or directory service daemons (DCE) to determine the instantaneous network location of the servers providing access to the desired information source. The *method of access* field contains information about the server and the way the server is accessed (e.g., whether it is a primary or an intermediary server, whether or not use of the server requires special authorization, etc.). This model provides for information sources to be freely moved around in the network without requiring any modifications to the MMBIR/ISDB.

When multiple servers supporting the same information source are available, the MMBIR system will intelligently and dynamically determine which underlying computing service (e.g., the researcher's local pc or a remote supercomputer) will best serve the researcher's specific needs and may make appropriate recommendations to the researcher. For example, suppose that a researcher requires access to FASTA to GenBank service and the MMBIR system locates three FASTA to GenBank servers, two of which are on supercomputers and one on a small workstation. Based on negotiations between the available servers and the MMBIR system, it will be determined which server is best suited *at that particular time* to service the researcher's particular query in the most timely fashion. The MMBIR system may then automatically transmit the researcher's query to the chosen server, or may recommend to the researcher that the query be processed at that server. In neither case, however, will the researcher have to locate the optimal service or access it manually.

Overview of the MMBIR

Figure 4 shows a general outline of the envisioned network layout for which the MMBIR/ISDB will provide support. A researcher will make requests for service from his/her local computing service (i.e., a PC, MacIntosh, workstation, etc.). Using the MMBIR/ISDB, the MMBIR system will attempt to select and access the relevant information sources wherever they might exist in the network. Servers for these information sources will either be directly accessible or indirectly accessible through intermediary servers. Regardless of the specific implementation, the results of the service will be returned to the researcher's local computing service for further processing. If deemed appropriate, with the help of the MMBIR/ISDB and the researcher, the MMBIR system will locate and access additional information sources as required.

Servers that support information sources can *advertise* their existence and availability by *registering* with a network directory service (CDS in DCE or GLBD in NCS). This permits the MMBIR, with the help of the MMBIR/ISDB, to locate an appropriate server and to access it without the researcher being overtly aware of the exact network location of that server. It is beyond

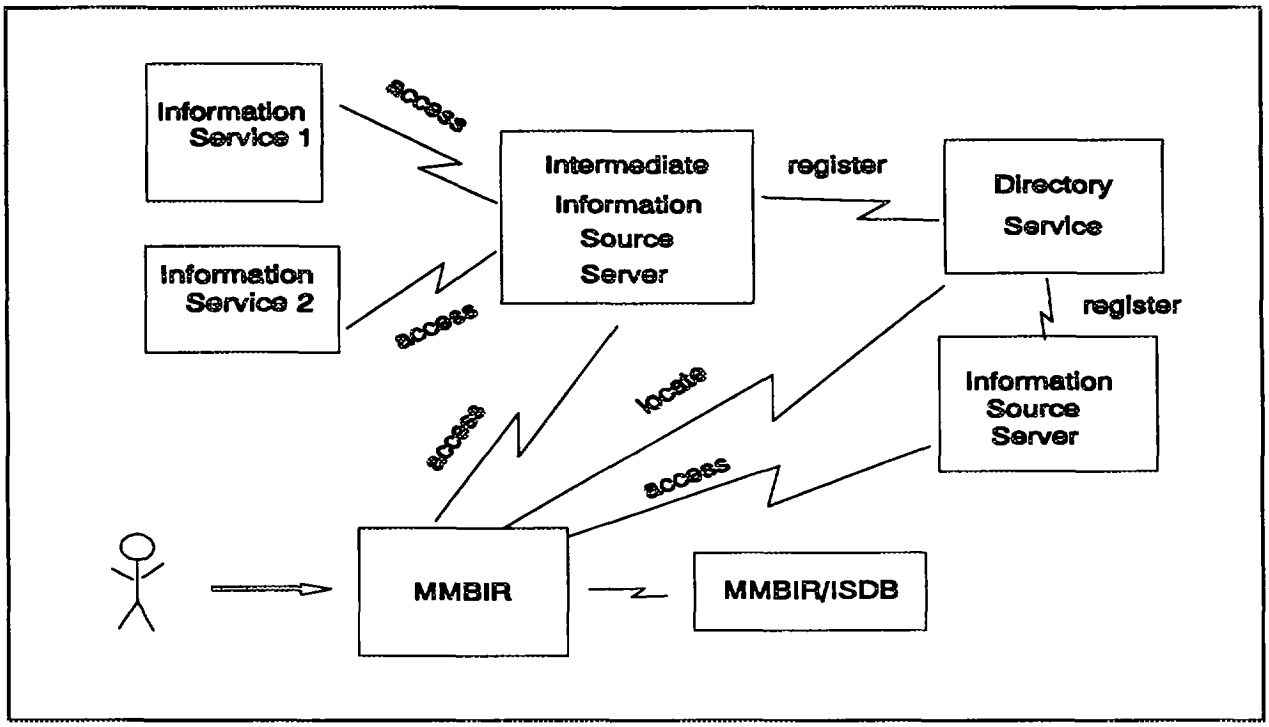

Figure 4 Network Layout of the MMBIR

the scope of this paper to discuss the details of this solution to the systems traversal problem. These details will be presented in another paper (Springer, in preparation).

Taking a simple example, assume a researcher wants to perform a sequence analysis using the FastA program in the GCG package (Devereux, Haeberli, & Smithies 1984). This need is satisfied by the system locating an entry in the MMBIR/ISDB for the GCG FastA service. Using the information in this entry, an accessible GCG server will be located and the researcher's request will be transmitted to the identified server to run the FastA program. Upon completion of the program, the GCG server will return any results to the researcher's local computing service. These results will be available to both the researcher and the MMBIR system for further use.

The MMBIR system will peruse the returned results and using the MMBIR/ISDB will identify that bibliographic indexes to the Medline database are present in the results. The MMBIR/ISDB *contents of results* and *input/output vocabulary* fields will be consulted. Once these bibliographic indexes have been located, the MMBIR/ISDB will be again consulted to identify what the Medline service is and where it can be located. A query will be constructed to retrieve the bibliographic entries from Medline and the Medline access will occur utilizing the server that supports the Medline service. Upon return of the results from the Medline service, these results will again be made available to the researcher and the MMBIR system to determine if any additional information sources should be accessed to acquire additional pertinent information for the researcher. This process continues until a researcher's information needs have been satisfied or it is determined by the researcher that no additional information is available or needed. In no case, however, does the researcher have to locate a service or access a service manually.

Conclusion

The MMBIR/ISDB provides a basis for programs which can function as intelligent research aids. It provides data for a knowledge base which can be used to make intelligent and timely decisions regarding information sources that may satisfy a researcher's information needs. It also provides

data for a knowledge base which can be used in an intelligent and timely selection of the computing services best suited to the researcher's needs. As such, it will be an invaluable tool to provide researchers with the knowledge and ability to carry on their research endeavors relatively unimpeded in a heterogenous computing environment. In addition, a researcher's possible inexperience with computers and networks as well as his/her limited knowledge of possibly useful information sources will not be an impediment to fruitful research. The researcher only needs to decide, guided by recommendations from the MMBIR system, what information source and supporting computing service he/she wants to use, not how to access these information sources and supporting computing services.

References

Anthes G. H. 1991. Let Your "Knowbots" Do the Walking. *Computerworld* 25: 17.

Daviss B. 1991. Computer Gremlins: The New Batch. *Marketing Computers* April 1991: 31.

Devereux, Haeberli, and Smithies 1984. A Comprehensive Set of Sequence Analysis Programs for the VAX. *Nucleic Acids Research* 12(1): 387-395.

Humphreys B. L. and Lindberg D. A. B. 1992. The Unified Medical language System Project: A Distributed Experiment in Improving Access to Biomedical Information. In Proceedings of the Seventh World Congress on Medical Informatics, 1496-1500. Amsterdam: North Holland.

Keen G., et al. 1992. Access to Molecular Biology Databases. *Mathematical and Computer Modelling* 16: 93-101.

Kong M., et al. 1990. *Network Computing System Reference Manual.* Englewood Cliffs, NJ.: Prentice-Hall, Inc.

Markoff J. 1990. Creating a giant computer highway; Robert Kahn's vision of a national network of information begins to take hold. *New York Times* v139, Sept. 2, 1990.

Open Software Foundation (OSF) 1992. *Introduction to OSF™DCE.* Englewood Cliffs, NJ.: Prentice-Hall, Inc.

Pearson w. R. and Lipman D. J. 1988. Improved Tools for Biological Sequence Comparison. In Proceedings of the National Academy of Sciences of the United States of America, 2444-2448. Washington: National Academy of Sciences.

Rosenberry W., Kenney D., and Fisher G. 1992. *Understanding DCE.* Sebastopol, CA.: O'Reilly and Associates, Inc.

Springer G. K. A National Scientific Computing Environment for the Biological Sciences. *in preparation.*

Springer G. K., Loch J. L., and Patrick T. B. 1992. An Open System Network for the Biological Sciences. In Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care, 535-539. New York, NY.: McGraw-Hill, Inc.