BioSYNTHESIS: bridging the information gap*

By Naomi C. Broering, M.L.S., M.A., Director

H. Richard Gault, Systems Analyst

Georgetown University Medical Center Library
3900 Reservoir Road, N.W.
Washington, D.C. 20007

Howard Epstein, Technical Manager

AT&T Federal Systems
Washington, D.C. 20036

BioSYNTHESIS is a prototype intelligent retrieval system under development as part of the IAIMS project at Georgetown University. The aim is to create an integrated system that can retrieve information located on disparate computer systems. The project work has been divided in two phases: BioSYNTHESIS I, development of a single menu to access various databases which reside on different computers; and BioSYNTHESIS II, development of a search component that facilitates complex searching for the user. BioSYNTHESIS II will accept a user's query and conduct a search for appropriate information in the IAIMS databases at Georgetown. For information not available at Georgetown, such as full text, it will access selected remote systems and translate the search query as appropriate for the target system. The search through various computer systems and different databases with unique storage and retrieval structures will be transparent to the user.

BioSYNTHESIS I is complete and available to users. The design work for BioSYNTHESIS II is under development and will continue as a multiyear technical research effort of the proposed Georgetown IAIMS implementation project.

The BioSYNTHESIS system, a prototype intelligent retrieval system, is an outgrowth of the Georgetown project to develop an Integrated Academic Information Management System (IAIMS). It is an experimental software system that retrieves information stored in disparate databases and hardware systems. It is an attempt to facilitate multiple database searching and to develop a means of attaining system integration and cohesiveness from a varying group of databases. BioSYNTHESIS is being developed in two phases: BioSYNTHESIS I, the gateway system, an interface which provides transparent access to selected databases; and BioSYNTHESIS II, a more complex system designed to function as an information finder responding to user queries for specific information. BioSYNTHESIS II is a multiyear endeavor currently under development as part of the IAIMS implementation project at Georgetown.

* The project is supported in part by grants from the National Library of Medicine (NIH 5R01LM04450) and the AT&T Foundation.

BIOSYNTHESIS I

Developed in 1987 as part of an IAIMS research-related project, BioSYNTHESIS I is being tested at the Dahlgren Memorial Library by users who search the medical center's IAIMS databases. It is a gateway system with a single menu, which displays the IAIMS databases, consisting of bibliographic, information, diagnostic, and communication systems. These databases serve as core resources available to Georgetown users. The IAIMS databases (Figure 1) include:

Bibliographic databases

1. The Library Information System (LIS), an online catalog of the library's books, journals, audiovisuals, and computer software holdings [1].
2. The miniMEDLINE SYSTEM™, an in-house subset of the National Library of Medicine (NLM) MEDLARS file, which contains references to articles (including abstracts) from

™ The miniMEDLINE SYSTEM is a trademark of Georgetown University.

Figure 1
Existing BioSYNTHESIS I menu

IAIMS AND LIS

BIBLIOGRAPHIC DATABASES
1 Online Catalog
2 miniMEDLINE
3 ALERTS/CURRENT CONTENTS
INFORMATION DATABASES
4 Drug and Poison Info
5 Physician Data Query (PDQ)
DIAGNOSTIC SYSTEMS
6 RECONSIDER
COMMUNICATIONS
7 Electronic News
8 Mail Box

SELECT one or Q to quit:

over 460 journal titles published over the latest three years. These references and abstracts are updated monthly, and the MeSH vocabulary is updated annually [2].
3. ALERTS™/Current Contents® Search System, the latest article references listed in the tables of contents of clinical and life sciences journals in the Institute of Scientific Information (ISI) system. This file is updated bimonthly [3].

Information databases
4. Drug Information System developed by MicroMedex, which includes special files such as DrugDex, PoisonDex, and EmerginDex. This file is updated monthly [4].
5. The Physicians Data Query (PDQ), a cancer treatment protocol system developed and maintained by the National Cancer Institute (NCI). It is updated monthly [5].

Diagnostic system
6. RECONSIDER, a diagnostic prompting system developed at the University of California, San Francisco [6]. This system is a full-text version of the *AMA's Current Medical Information Terminology* (CMIT), Fifth edition, 1981. It is being used as an experimental teaching tool for clinical problem solving by the School of Medicine to expose students to the sort of system they may be using in future medical practice.

Communications
7. Electronic News, an automatic news system, supported by the library, that transmits a variety of special news items, such as changes in library hours, searching tips, special events, and conferences.
8. Mail box, an electronic mail system for message exchange among students and faculty.

BioSYNTHESIS I provides a direct connection from the user's terminal to the target database. As a result, users must know which databases they wish to search and how to conduct the search routines for appropriate retrieval. The system allows the user to highlight text on the screen, for example, a title or medical term, and transfer it to a search in another database. The eight systems in BioSYNTHESIS I reside on several different computers, which are linked to each other via TCP/IP Ethernet and the library's ISN local area network (LAN). LIS, miniMEDLINE, and Electronic News are housed in a DEC PDP 11/44. ALERTS/Current Contents, the Drug Information System, and PDQ are maintained on a DEC VAX 785. Mail box and RECONSIDER are located on an AT&T 3B15.

The BioSYNTHESIS software resides on an AT&T 3B2/600 system. Prior to development of BioSYNTHESIS II, remote users had to know which machine to access for their desired application. In addition, Mail box was not available. While simple user interfaces existed on the databases, there was no integration among the disparate computers or the databases to provide a consistent user interface. The single menu (as shown in Figure 1) has greatly improved ease of use. The system also currently establishes links to remote computers via dial-up, direct RS-232, Ethernet (TCP/IP), etc.

The BioSYNTHESIS software is being written in UNIX™. The general control scheme was written in UNIX Shell command language 'C.' The communications script language and user interface language are two packages (Sparrow and PROTO) in the public domain for UNIX systems. After the initial version was tested, many of the programming routines were rewritten in 'C' to improve performance. This coding method—develop and test using Shell, then move stabilized programs to 'C' for speed—is being used for BioSYNTHESIS II and various aspects of the entire project.

THE CONCEPT OF BIOSYNTHESIS II

To begin to understand the concept of BioSYNTHESIS II, one should consider the nature of an information system. It is a collection of well-defined items of data, related and organized in some meaningful manner. Compare this with a knowledge system, which consists of a core information system along with rules defining the scope, context, and use of that information. In general, knowledge systems deal only with their own embedded information. They have little or no ability to dynamically access information from other knowledge systems.

A more powerful type of knowledge system, which can be called a "composite knowledge system," may be conceived by decoupling the knowledge rules from an underlying information system and providing a mechanism whereby the new system can learn to dynamically extract information from several dispa-

™ ALERTS is a trademark of Georgetown University.
® Current Contents is a registered trademark of the Institute for Scientific Information, Inc. (ISI).

™ UNIX is a trademark of AT&T Bell Laboratories.

rate collections of information. In order to provide maximum flexibility, this definition should be independent of the organization of the various information databases, the hardware on which they reside, and the methods used to access them.

There exist numerous sources for bibliographic, diagnostic, and clinical information in formats available on personal computers; on in-house, multiuser systems; and on dial-up, vendor systems. Although well managed as individual medical information systems, they do not function as a composite system. Generally, the system is able to access only its own information base. Often there is widespread duplication of information across the various databases. In addition, indexing methods and user interfaces are inconsistent from system to system.

Today, individuals must use their own set of searching rules when attempting to access and reconcile information from two or three knowledge systems. As the number of knowledge systems grows, it becomes increasingly difficult for any one person to manage all the rules that must be used to move between systems. In addition to knowing the access method for each system, the user must be familiar with differences in organization and representation of data. A complete literature/information search on one closely defined medical topic may take the researcher into a dozen information systems. Long before a search reaches this degree of complexity, the automated knowledge system becomes the means of choice for accessing and organizing the necessary information.

BIOSYNTHESIS II DESIGN GOALS

The initial phase of BioSYNTHESIS II, currently under development, will be a composite system which draws information from several databases, each of which has its own unique data structures and user interface. Similar to BioSYNTHESIS I, it will be designed to retrieve information from different databases and computers. Once a query is made, the system will search for a response, and the user will interact with the system only at key points to refine or broaden a query.

A knowledge system such as BioSYNTHESIS II is one way of blending various information systems. Although each information system is built differently, indexed differently, and organized differently, the searching mechanism embodied in BioSYNTHESIS II can "bridge" these information systems, forming them into an integrated whole and thereby making it necessary for the user to learn only one interface.

The intent is to tie together bibliographic, information, and diagnostic systems into one vast, integrated information network. Specifically, bibliographic sources include library online catalog, circulation, and journal article references with abstracts, and full-text systems. Information sources include factual data such as drug information and treatment modalities. The diagnostic systems are expert systems that prompt the user for input and can serve as decision support systems for physicians. In concept, a user might request information on one particular medical topic and BioSYNTHESIS II would determine whether to answer that request with a list of citations, an online article, a direct information retrieval, or a "conference" with another knowledge system. The user would conduct an interactive dialogue with BioSYNTHESIS to fine tune the search criteria, to expand the retrieval criteria in one direction, or to limit it in another direction. Finally, the user could store the research results in a personal information system and acquire materials outside the information network's domain via a gateway system for document retrieval.

These activities imply development of the following functional subsystems within BioSYNTHESIS: a user interface, an access manager, a context selector, an information analyzer, and support services.

1. **User interface.** Provides the user with a consistent interface for exchanging data with the selected databases. It is important that the user be able to enter data in a consistent format, regardless of the database to which that data is being transferred. It is just as important that the data being returned to the user be returned in a consistent format, regardless of the data's source.
2. **Access manager.** Responsible for the actual physical communications with the selected database. This includes automatic logon and logoff, as well as the ability to perform a "conversation" with the selected database, simulating a complete user session. This is done via a series of communication "scripts," which contain detailed knowledge of how to interact with a given database. The commands available for these scripts form a very high-level programming language, which allows the communication process to be tailored to each database and modified quickly and easily.
3. **Information analyzer.** The format of the output from the supported databases varies considerably from one system to another. This functional subsystem analyzes the data, reformatting it into a consistent internal form. In addition, it may "interpret" incoming data, for example, to determine if the data is a command request or textual information. This module is integrated into the access manager and provides a measure of intelligence for the communications process.
4. **Context selection.** Determines the appropriate database(s) necessary to fulfill the purpose of the user's request. It uses a combination of table lookups and explicit rules to select the target database(s). It may also query the user if more detailed information is needed to narrow the target database(s).
5. **Support services.** BioSYNTHESIS II will link to library cataloging and circulation, to external document delivery services, to photocopy services, and to personal filing services for research, storage, and data manipulation through a series of special functions.

Figure 2
Prototype BioSYNTHESIS II Menu 1988–89

```
IAIMS AND LIS

BIBLIOGRAPHIC DATABASES
    1 Online Catalog
    2 miniMEDLINE
    3 ALERTS/CURRENT CONTENTS
    4 Full Text
INFORMATION DATABASES
    5 Drug and Poison Info
    6 Physician Data Query (PDQ)
DIAGNOSTIC SYSTEMS
    7 RECONSIDER
    8 DXplain
COMMUNICATIONS
    9 Electronic News
   10 Mail Box
   11 External Access

SELECT one or Q to quit:
```

Figure 4
Bibliographic module: BioSYNTHESIS II

```
BIOSYNTHESIS

Step 1: ENTER term(s) - Use TAB or ARROW for field
Step 2: Select a Screen Command, press RETURN or ENTER
KEYWORDS: 1
          2
          3
AUTHOR:
TITLE:
JOURNAL:
YEAR:          MONTH:          Screen Commands
                                  S  Save
                                  Q  Quit
                               ENTER a Choice:
```

THE BIOSYNTHESIS II PROTOTYPE

Georgetown University has developed a research prototype of BioSYNTHESIS II which embodies initial versions of the functional subsystems and acts as an experimental platform for testing new ideas and processes. The prototype targets integration of the same eight IAIMS databases included in BioSYNTHESIS I, but in addition includes access to a full-text retrieval component using BRS, which has full-text medical journal articles, and access to DXplain, a diagnostic system based at the Massachusetts General Hospital, Harvard Medical School. This portion of BioSYNTHESIS was released for testing in early 1989. (Figure 2)

Primary technical design work to-date has been with the user interface, access manager, and information analyzer functions. These subsystems deal with the way BioSYNTHESIS I and II physically interact with the user and with other information systems. This physical level of intersystem connection must be in place before development of the context selector can be considered. The user interface to the BioSYNTHESIS II prototype is a combination menu and a "fill-in-the-blanks" subsystem (see Figure 3). It guides the user through the various system functions by allowing the user to conduct bibliographic searches, use information systems, manage a personal bibliographic database, retrieve documents, and send messages to a librarian.

An ideal component that is not yet designed and is reserved for research in future years is a context selection component. It is a second system layer which would provide much of the intelligence needed for a full BioSYNTHESIS II. It is complex and requires thesaurus building. Without context selection, the user will have to make explicit decisions regarding the nature of the needed information (i.e., whether articles or books are needed, whether specific information frames are desired, or whether the retrieval results indicate a need to research a secondary information source). The information analyzer, functioning in conjunction with the context selector, will also represent a major technical programming issue to be addressed in the research effort.

Figure 3
Future BioSYNTHESIS II menu

```
IAIMS AND LIS

1 BIOSYNTHESIS
2 Bibliographic Databases
3 Information Systems
4 Diagnostic Systems
5 Communications Systems

SELECT one or Q to quit:
```

BIBLIOGRAPHIC SEARCH

Nevertheless, there are major sections of BioSYNTHESIS II that can be developed and released in the next few years. One such component is the bibliographic component, the core module of BioSYNTHESIS II. It is this component that makes the first attempt to understand a research request. In the bibliographic module, the user may provide keywords, author, title, and date ranges to retrieve citations (Figure 4).

A basic function of the bibliographic module will be to translate terms entered by the user into a standard medical vocabulary. The use of such a controlled vocabulary is attractive as a coherent mechanism for presenting information requests across systems. Unfortunately, a universally standard, controlled medical vocabulary is not yet available.

The most suitable vocabulary available currently is MeSH, the Medical Subjects Headings system supported by NLM. However, the presentation of information in MeSH is hierarchical (tree structured) rather than cross-referenced (network structured), and MeSH does not widely support abbreviations and synonyms. In addition, several medical information systems do not use MeSH as an indexing vocabulary. The Unified Medical Language System (UMLS) proposed by NLM may address and solve these problems. Because of widespread diversity in medical language, cross-indexing systems will probably be needed. MeSH, being the most suitable vocabulary available, serves as the base for BioSYNTHESIS II's bibliographic module.

A controlled vocabulary system will be a key to the context selector, as it will provide a way of mapping terms from one system to another. For example, assume that for a given term there is an associated major heading within the vocabulary. The decision-making rules could conveniently reside at the major heading and secondary heading levels. Therefore, the basis of context selection in this case would be to follow the vocabulary term up to a main or secondary subject heading and look for one or more decision rules based upon that heading. A simple example of such a rule would be a term which indexes to "Neoplasms" (a MeSH major heading). This would alert the context selector to look in Georgetown's miniMEDLINE bibliographic system, or PDQ, a cancer protocol knowledge system. Such rules can be based on tables or decision trees to provide entry points into secondary systems from a starting term or search criteria.

In the bibliographic subsystem, the user provides keywords, and BioSYNTHESIS II matches them to MeSH terms. It then presents true MeSH terms, resembling the original request, to the user and asks for a refined selection of them. The refined term list then forms the primary search criteria. Once the search criteria have been established, the search is begun using the appropriate database, and the desired information is retrieved. The user can review the retrieved information and further refine the search by specifying MeSH terms to drop, to expand (go up to a broader subject heading), to restrict (drop to a narrower MeSH heading), or to add. In addition, the user can cross-reference terms to take the search request in a new direction and rerun it. This process would be continued until the search results were satisfactory, at which point the bibliographic references could be permanently saved.

THE INFORMATION SUBSYSTEMS

The information subsystems provide users with a "window" into a variety of factual, expert, or knowledge systems. The user will interact normally with a system, but at any desired point can switch into a "context search" mode, highlight words or phrases of research interest on the display screen, and specify how they should be used. BioSYNTHESIS II would accept these new terms and could conduct a bibliographic search on them, as outlined above, concurrent with an active expert system session. This would provide a preliminary research mechanism.

Searching information subsystems will be complex because they must have varying file structures and no standard vocabulary system. A means of mapping user terms into these "textual"-based knowledge systems will have to be developed to provide searching capabilities beyond simple bibliographic searches.

PERSONAL FILING SUBSYSTEM

In addition to being able to file results, the user will be able to recall previous search results, or continue the initial search. The information analyzer will be a key part of the filing system. Whenever a citation is received, either from a bibliographic system or embedded in text in an expert system, the information analyzer will decipher the citation, separating it into author, title, index term, and several other components. This will provide citation compatibility across different bibliographic systems. For instance, in order to search a cross-reference, BioSYNTHESIS II must know how to locate the cross-references within the citation. It will, because it will analyze the citations and separate them into their component parts. Analyzing citations guarantees that information output from BioSYNTHESIS II will be reusable as input for continued research.

DOCUMENT RETRIEVAL SUBSYSTEM

Any citation could be marked for retrieval in BioSYNTHESIS II. Retrieval could be direct (via online text systems) or indirect (via photocopy requests, circulation system requests, or interlibrary loan requests). All interfaces to these systems would be managed by BioSYNTHESIS II.

HOW BIOSYNTHESIS II WILL WORK

BioSYNTHESIS II may conceivably represent a new, hybrid, front-end system, which collects information from various components, interacts with them, and compares information in real time. It may present an image of a single, vast collection of data and interrelated information appearing to the user as a single logical database. In addition to the information content and relation problems inherent in designing such a system, there are some technical problems. Ideally, such a system would work with other knowledge and information systems without user consideration of machine architecture, location, implementation language, or user interface design. Ideally, at the programmer level, the tools would be independent of implementation language and of machine type and location. None of these will be available for BioSYNTHESIS II. Additionally, BioSYNTHESIS II will have to contend with the "given" nature of the target information system it is searching. The target system or database may be under the control of other parties and, therefore, unable to be modified. In any case, the modification of existing information systems would introduce a level of complexity too great to consider for a system of this magnitude.

The key technical design concept in BioSYNTHESIS II will be the virtual "expert user." The BioSYNTHESIS II access manager will be the basic user for all knowledge and information systems with which BioSYNTHESIS II communicates. As the "expert user," BioSYNTHESIS II will know how to log onto these systems, how to format requests for information, and where to look for the responses to those requests. The virtual user will be implemented in communication dialogue scripts, where BioSYNTHESIS II is told what to say to the target information system, how the information system is likely to respond, and what to do in various cases based on the target system's response. There will be a different virtual user (system of dialogue scripts) for each supported system.

The main BioSYNTHESIS program module will manage overall decision making and flow control for the BioSYNTHESIS II session, transport user requests to the access manager, interpret incoming data via the information analyzer, and blend outputs together for presentation to the user through the user interface.

BENEFITS OF BIOSYNTHESIS II

Dealing with the various information systems as a set of users, and using appropriate high-level design tools, BioSYNTHESIS II will make it possible to implement a knowledge-based, integrated system very rapidly, without modifying existing database systems and requiring highly complex programming approaches. In addition, it will resolve the problem of inconsistent searching functions, another problem inherent in combining information systems. BioSYNTHESIS II, for example, may retrieve a set which is the logical AND of all search terms, including keywords. In such a case, all criteria must be met for the match to be successful. Some knowledge systems cannot retrieve information in that manner. In these cases, BioSYNTHESIS II augments the functionality of those systems by retrieving all relevant information, logically ANDing it internally, and then presenting it to the user. Therefore, when the designer conceives a basic set of system-level functions which the composite knowledge system should provide, certain virtual expert users can be augmented to provide this extended functionality. For example, a user might extract all information meeting a given search criteria (say 30,000 records) and then sift through all this information manually to select only those records which satisfy the criteria. This clearly is not practical to do manually, but the components of the composite system run at high computer speed, which makes this a feasible task. Thinking of functional augmentation this way has a big advantage; the analysis proceeds as though the task were being manually implemented and considerations of a target system's specific implementation are not necessary.

VALUE OF THE PROTOTYPE

To test the concept, a basic BioSYNTHESIS II prototype was designed in Shell language using several custom designed, programming tools. A scripting language was necessary to drive the dialogues with the target knowledge systems. The attributes of the scripting language required conditional execution, machine/communications, media independence, multiple session support, and a simple syntax. In addition, a screen designer was needed to provide a consistent interface for the user.

In designing the basic prototype, information was gathered about the complexities involved in creating a working system. The prototype functions slowly and requires considerable operation before it can become functional. While the ability to conduct several simultaneous background searches is appealing, it is impractical because the process ties up too many entry ports to the computers, thereby shutting off other users. However, ideas are being formulated on ways to proceed with the final design, which will be conducted during the IAIMS implementation phase.

SUMMARY AND CONCLUSION

BioSYNTHESIS II is the most promising, but also the most complex, part of the system. The prototype represents an initial stage of possible knowledge system approaches to ease information access for users. It is important to proceed in a coordinated manner to make retrieval easy for health professionals and to provide them with a variety of useful databases to ease medical decision making for the delivery of patient care. While there may be other approaches to bridging information systems on the horizon, Georgetown has selected the BioSYNTHESIS approach as a possible solution for achieving system integration. Achieving a full BioSYNTHESIS will be a gradual process. Plans are to release portions of BioSYNTHESIS II in phases over the next few years, as modules are tested and properly modified. This way users can begin to use the system without waiting for a full BioSYNTHESIS II and can provide input for future modifications.

REFERENCES

1. BROERING NC. The Georgetown University Library Information System (LIS): a minicomputer-based integrated library system. Bull Med Libr Assoc 1983 July;71(3):317–23.

2. BROERING NC. The miniMEDLINE SYSTEM®: a library-based end-user search system. Bull Med Libr Assoc 1985 April;73(2):138–45.

3. BROERING NC, CANNARD BC. Building bridges: LIS-IAIMS-BioSYNTHESIS. Spec Libr 1988 Fall;79(4):302–13.

4. MACNEAL HOSPITAL INFORMATION SERVICES. New MicroMedex information system provides poison, drug and emergency information via PCS/ADS system. (Publicity Release) 1984 November 6.

5. PDQ: a new database on cancer therapy. NLM News 1982 Sept;37(9):4.

6. BLOIS MS, ET AL. RECONSIDER: an experimental diagnostic prompting program. UCSF ACP Computer Workshop 1983. Med Info Sci 1983;7–24.

Received February 1988; accepted April 1988

FROM THE *BULLETIN*—75 YEARS AGO

On the Use of Medical Libraries

Why more physicians are not readers in libraries is probably explained by the arduousness of their labors and the long, uncertain hours but there are other things which should be considered in this connection. Among these, we may mention the attractiveness of the reading room, the ease with which books may be consulted and the presence of new books and new journals in a convenient place so that they may be had without asking. There are not many large libraries so constructed that they can use the open shelf system and there are probably not many that, even if they were so fortunately situated to make this possible, would care to have the books open to all their readers. But there is nothing so conducive to consulting numerous authorities as having them just under one's hand and the open shelf system has very much that can be said in its favor but in the absence of it a clever librarian can do much to overcome its shortcomings by placing in the reading room various groups of books relating to topics of present interest or sets of books that are new to the library or even small collections of books on given topics choosing some of the newer things and some of the classics.

Bull Med Libr Assoc 1914 Jan;3(3):41