Biological Databases: A New Scientific Literature

Robert J. Robbins

Biology is entering a new era in which data are being generated that cannot be published in the traditional literature. Databases are taking the role of scientific literature in distributing this information to the community. The success of some major biological undertakings, such as the Human Genome Project, will depend upon the development of a system for electronic data publishing. Many biological databases began as secondary literature—reviews in which certain kinds of data were collected from the primary literature. Now these databases are becoming a new kind of primary literature with findings being submitted directly to the database and never being published in print form. Some databases are offering publishing on demand services, where users can identify subsets of the data that are of interest, then subscribe to periodic distributions of the requested data. New systems, such as the Internet Gopher, make building electronic information resources easy and affordable while offering a powerful search tool to the scientific community. Although many questions remain regarding the ultimate interactions between electronic and traditional data publishing and about their respective roles in the scientific process, electronic data publishing is here now, changing the way biology is done. The technical problems associated with mounting cost-effective electronic data publishing are either solved, or solutions seem in reach. What is needed now, to take us all the way into electronic data publishing as a new, formal literature, is the development of more high-quality, professionally operated EDP sites. The key to transforming these into a new scientific literature is the establishment of appropriate editorial and review policies for electronic data publishing sites. Editors have the opportunity and the responsibility to work in the vanguard of a revolution in scientific publishing.

Introduction

Biology is entering a new era, in which every year megabytes of archival-quality data are generated in each of hundreds of laboratories around the world. Although this information is important, it cannot be published or used in the traditional form of journal articles. A new kind of computer-based scien-

Robert J. Robbins is program director for Bio-informatics Infrastructure in the Office of Health and Environmental Research of the U.S. Department of Energy. He is also associate professor of medical information at Johns Hopkins University, where he served as director of the Applied Research Laboratory, William H. Welch Medical Library, and director of the Informatics Core of the Genome Data Base before going to DOE. Before joining the Hopkins faculty in 1991, he served as program director for Database Activities in the Biological, Behavioral, and Social Sciences at the National Science Foundation. He currently serves on the advisory boards for several biological databases. He received his Ph.D. in zoology from Michigan State University in 1977. He also holds an A.B. in Chinese and Japanese history from Stanford University. Address for correspondence: Welch Applied Research Laboratory, Johns Hopkins University, 2024 East Monument Street, Baltimore, MD 21205.

tific literature is emerging in which data are distributed in electronic databases. This *electronic data publishing* (EDP) raises many questions. What is EDP? Is EDP really needed and useful? Are databases a form of publishing? How does the process of EDP resemble that of traditional publishing? What does it mean to edit or review data? Is EDP just a new kind of vanity press? Can electronic publishing be trusted? What is the role of editors in EDP?

The technical and legal problems associated with electronic publishing have been discussed at great length elsewhere and they will not be treated here. Instead, I will discuss EDP and its role in biology, showing how in some areas of biology EDP has evolved from an electronic version of a traditional review into a new kind of primary literature. Two specific examples of EDP relevant to biology will be presented to illustrate the power of current systems and the challenges they pose. Finally, I will consider several of the publishing and editing issues raised by EDP.

What Is Electronic Data Publishing?

In any research area there is a continuum that begins with raw data and continues through derived information and finally culminates in the refined knowledge that constitutes our understanding of the field. Traditional biological publishing has emphasized information and knowledge, not data. Now biological research is generating many findings that are close to the data end of the spectrum, yet need to be shared with the research community. To accomplish this, several large databases have arisen to support EDP for molecular biology. For example, in the United States, GenBank® and GSDB (genome sequence data base) collect nucleotide sequences and PIR (the protein identification resource) does the same for proteins. PDB (protein data bank) stores protein structures. Genetic map information is managed at several organism-specific centers—GDB (genome data base) for humans, FlyBASE for Drosophila, GBASE for the mouse, AAtDB for Arabidopsis, and so on. Small projects collect and manage information on restriction enzymes, molecules of immunological interest, etc.

EDP projects occur in all areas of biology, not just genomic research. The Flora North America project at the Missouri Botanical Gardens is an effort to identify and catalog all known species of plants on the continent. The National Science Foundation's Long Term Ecological Research projects collect data on ecological phenomena that occur on decade or longer time scales. Recently the Department of Interior announced plans to establish a National Biological Survey that will be charged with identifying, cataloging, and studying the complete biota of the nation.

The Human Genome Project

The international Human Genome Project (HGP)—the first really big science project in biology (Cantor, 1990; DeLisi, 1988; Watson, 1990)—provides a com-

pelling argument for EDP. The official goals of the project are: (1) construction of a high-resolution genetic map of the human genome; (2) production of a variety of physical maps of all human chromosomes and of the DNA of selected model organisms; (3) determination of the complete sequence of human DNA and of the DNA of selected model organisms; (4) development of capabilities for collecting, storing, distributing, and analyzing the data produced; and (5) creation of appropriate technologies necessary to achieve these objectives (USDOE, 1990). The first three goals involve biological bench research, whereas the fourth calls for the development of an adequate EDP infrastructure to manage the resulting data. The fifth is a frank admission that none of the other goals can be met with current technology.

With DNA sequences, the need for EDP is acute. Molecules of deoxyribonucleic acid (DNA) are the code script of life. DNA is a polymer, consisting of a linear sequence of four different subunits called nucleotides. The nucleotides are often abbreviated as A, T, C, or G, after their fuller names of adenine, thymine, cytosine, and guanine. Thus, a particular DNA molecule can be specified with a string of these four letters. For example, GAATTCTAA . . . is the beginning of the DNA string that codes for the protein beta-hemoglobin. At conception, each human parent contributes one set of DNA instructions (a haploid genome) as twenty-three chromosomes containing more than three billion nucleotides. The two parental contributions combine to produce the redundant diploid genome of the child.

Appreciating the amount of information in a human haploid genome is best done with an analogy. Imagine the DNA sequence for one sperm cell typed in 10-pitch type, as a linear sequence of 3.3 billion A's, T's, C's, and G's, on a continuous ribbon. This ribbon could be stretched from San Francisco to Chicago, then on to Baltimore, Houston, and Los Angeles, and finally back to San Francisco, with about 60 miles of ribbon left over.[1] This is just the amount of information in one sperm cell. Since, on average, any two haploid genomes differ in about one out of a thousand nucleotides, cataloging human diversity would demand enormous information resources. Data in this quantity cannot be used unless stored in a computer system.

Obtaining and cataloging strings of A's, T's, C's, and G's is not enough. Additional information is required to describe completed sequences. For example, commentary on the gene that codes for human beta-hemoglobin can be found in GSDB, GenBank, PIR, GDB, and OMIM (On-line Mendelian Inheritance in Man). At present, these databases contain more than 500,000 bytes of information about this gene, which itself is less than 5,000 nucleotides in length. Although such a hundred-fold multiplier will not occur for all sequences, the total amount of data and information produced by the HGP will be vast. Simply put, advances in electronic data publishing will accompany the genome project or the genome project will fail.

Biology Requires Electronic Information Management

A few years ago, Harold Morowitz made the visionary claim that information-management technology will do for biology what calculus has done for physics. In physics, it is held that one proton is the same as another, that any two hydrogen molecules are interchangeable, etc. In contrast, no two living things are exactly alike. Therefore, when studying living things collectively, we must have ways to store and manipulate the information that describes them as individuals. We cannot just deal with the standard properties of *the* mouse, *the* rat, or *the* human being. The essential individuality of living things and the requirement of maintaining and accessing data about individuals is what underlies Morowitz's claim.

It is not just molecular biology that requires access to powerful information-management technology. Ecosystem-level analysis requires acquiring and manipulating huge amounts of data. Tracking the millions of names and hundreds of millions of specimens of the world's biota is better done with an automated system. High-resolution global climate modeling requires more computer power than is currently available. When the Earth Observing System is fully operational, it is expected to generate terabytes of remote-sensing data per day.

Databases As Publishing

Are databases publishing? Can database development and distribution ever play the same kind of role as traditional publishing in the communication of scientific findings? Before addressing these questions, let us consider traditional publishing and the distinction between primary and secondary literature.

Figure 1 shows, from the perspective of the research community, the flow of information in primary literature and in reviews. This simplistic figure makes two points: (1) the publication of reviews follows that of the primary literature, and (2) researchers need not be (and generally are not) aware of the technical and professional infrastructure necessary for the production of printed literature.

Ignoring the role of editors and the publishing infrastructure, primary literature is a direct communication from the originator of the findings to the community. Review literature involves an intermediary (who must be expert in the field) who extracts related findings from the primary literature, then presents the summarized work along with a professional analysis and commentary. Both primary and review literature involve discrete acts of individual scholarship, with distinct beginnings and ends. We do not expect authors to commit themselves to a lifetime of continuous review of a particular field.

Although editors must be familiar with the details of actual journal production, many biologists see the publishing of primary literature as involving the straightforward preparation of a manuscript which is then sent to an editor

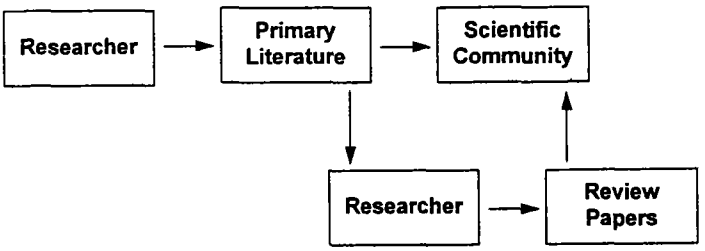

FIGURE 1. The flow of information in traditional publishing from the point of view of scientific researchers. Of necessity, review papers must lag behind the primary literature.

and which, after some arbitrary and unconscionable delay, finally appears in print.

That biologists ignore publishing infrastructure is a mark of its maturity and success. Good infrastructure is always nearly invisible to those who use it. We drop envelopes into metal boxes, expecting them to be delivered within a few days, without worrying about how it will happen. We place telephone calls casually, unmindful of the tremendous infrastructure necessary to ensure that dialing a particular number will connect us precisely to the one telephone we seek. We regularly take advantage of indoor plumbing without marveling at the social and structural infrastructure required to deliver safe, fresh water into every occupied building in the civilized world.

One of the reasons that databases seem different from traditional publishing is that database development is an immature field with a highly visible infrastructure. They are expensive to build and difficult to maintain. With our attention attracted to their awkward infrastructure, we can miss some fundamental similarities between early databases and traditional reviews and between current databases and primary literature.

Early Database Development

Early on, many of the important biological databases, such as GenBank or PIR, were similar to review articles: Important findings were extracted from the literature by a single researcher who then compiled and published them in a form that supported further use and analysis.[2] Such efforts had no explicit support and instead were operated as bootleg activities. With time, their value was recognized, leading to a call to "speed it up" and to make the data available in electronic form. Specific funding became available, either as grants or contracts, and database staff were hired to accelerate the process of data acquisition and distribution (Figure 2).

Here, the database is still functioning as a traditional review, albeit one with a staff of assistants and considerable technical resources. However, the key operational part of traditional reviews—that the responsibility for their assembly lies with the reviewer, not with the producers of the primary literature—is still in place.

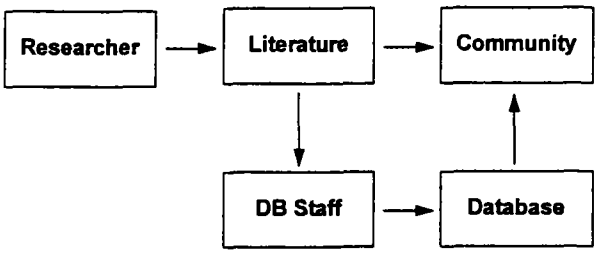

FIGURE 2. The flow of information in early biological databases. Several full-time staff were assigned to scanning the literature in an effort to find and extract relevant data in a timely manner. Although essentially an infrastructure activity, the work of database staff has a prominent role. Such a visible infrastructure often indicates an immature system and suggests that many changes will occur before a stable and reliable state is achieved.

Despite the best of intentions, the increasing rate at which the data were being generated caused the databases to fall further and further behind. A great backlog of data began to accumulate, producing cries for a change in the way the databases were operated. The notion that there was something terribly wrong with the operation of the databases became widespread (Kabat, 1989; Lewin, 1986).

The Database Scaling Problem

A GenBank problem did exist, but it was with the apportionment of responsibility between researchers and database staff, not with the database organization itself. Under the initial contract plan, the database staff were charged with collecting *all* sequences that satisfied certain criteria. No equivalent responsibility to assist in the process devolved to those generating the data. Instead, the database was seen as a service fully responsible for acquiring and publishing the sequences. This created a database scaling problem so that every increase in data volume resulted in an equal, and impossible to meet, increase in effort required of the database staff.

A moment's thought shows that, in rapidly growing fields, this operational model will result in a data-backlog problem no matter how competent or dedicated the database staff. If the rate of data generation is increasing and if every entry in the database must be extracted from the primary literature by database staff, then either the number of staff must increase proportionally to the data flow, or the database will fall behind.

Since unlimited staff growth is impossible, resolving the backlog must involve uncoupling the growth of database staff from the growth in data flow. One solution is to allow the researchers who generate the data to transfer their data directly into the database (Figure 3). The idea of direct submission was articulated by Fickett in May 1987 at the First CODATA Workshop on Nucleic Acid and Protein Sequencing Data (Fickett, 1989): "I envision . . . major changes in the data input process. First, the data path will change, with experimentalists

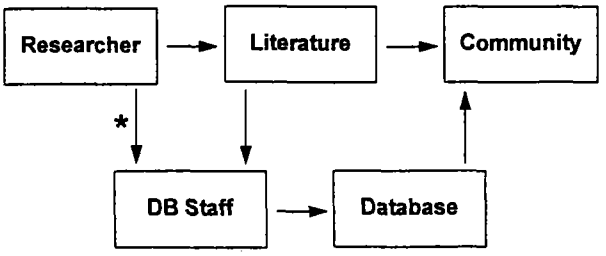

FIGURE 3. The effect of data submission on the flow of information in electronic data publishing. The arrow marked with the asterisk represents the first step toward making the database a form of primary literature.

sending data they now publish mostly in journals directly to the databases. Second, the process of reorganizing the data and entering it into databases will shift from databank staff to experimentalists."

At the same meeting, Fickett also explicitly recognized the similarity of database development to the production of review literature: "Rather than a set of closely similar records made by a databank staff, we should think of each database as a richly structured review article, continuously updated and revised by the community at large." This notion of continuous revision anticipated some challenging aspects of databases as publishing. The prophecy that a database might become a single giant review, in a perpetual work-in-progress state, with collective authorship, real-time updates, and continuous editing has, in many ways, come to pass.

Direct Data Submission

The backlog was solved with a software package called *AuthorIn*, which researchers now use to prepare files for direct submission to the database. Because these files can be read by database-loading software at Los Alamos, accession of the entry can be accomplished in days, instead of the months previously required. The effect has been enormous (Cinkosky et al., 1991):

> In 1984, it took on average over 1 year to get nucleotide sequences from journals to the users. This year, even though we processed ten times as much data (14.1 million nucleotides in 1990, as opposed to 1.38 million in 1984), the average delay between the time that an article appears and the time that the data are available in the database is 2 weeks.

Direct data submission also facilitated a step toward primary-literature status—editorial involvement. For the first time, editorial quality control could be applied to the sequence information itself.[3] With paper publications, neither reviewers nor editors can easily review sequences themselves, and most sequences sent to paper journals have been published as submitted. Now, over

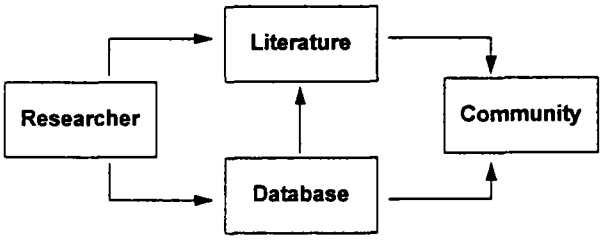

FIGURE 4. Mature electronic data publishing will be an alternative form of publishing. Databases will parallel, or even precede, the print literature. From the researcher's perspective, there will be no more difficulty in preparing an EDP submission than in preparing a journal submission, and the technical staff at the database will be largely unnoticed. Authors will be responsible for submitting their findings, where their authorship will be clearly recognized, and volunteer editors and reviewers will help ensure the quality and trustworthiness of the resulting system.

90 percent of nucleotide sequence data have been submitted directly to one of the collaborating databases (DNA DataBase of Japan, European Molecular Biology Laboratory Data Library, and GSDB GenBank) and all of these data must pass many software checks before they enter the database. The test results are sent to the originators of the data, who frequently use them to revise and improve their data. Consequently, directly submitted sequences are generally of much higher quality than those that appear in print form and are captured into the database. In effect, direct data submission replaces three error-generating steps (manuscript preparation, typesetting, and journal scanning) with one error-correcting step.

Direct submission also gives database editors time to examine the entire database and to make regular improvements in its form and substance. Currently, the daily workload for GSDB is evenly divided between updates to existing entries and the acquisition of new entries. Many databases now are naming external "curators" over portions of the data in the database and charging the curators with the responsibility of providing overall guidance and quality assurance for that portion of the database in their care.

Mature Electronic Data Publishing

Although a great success, direct data submission still has some distance to go. Many biologists are still uncomfortable using computers, and software packages like *AuthorIn* will never be as easy to use as commercial packages that cost tens or even hundreds of millions of dollars to develop. However, as computer systems improve, biologists should begin to consider the preparation of direct submissions to be as routine as manuscript preparation. More important, few will think about the technical steps required for publication that will occur after the manuscript has been sent to the database. The infrastructure for electronic data publishing will have become invisible (Figure 4).

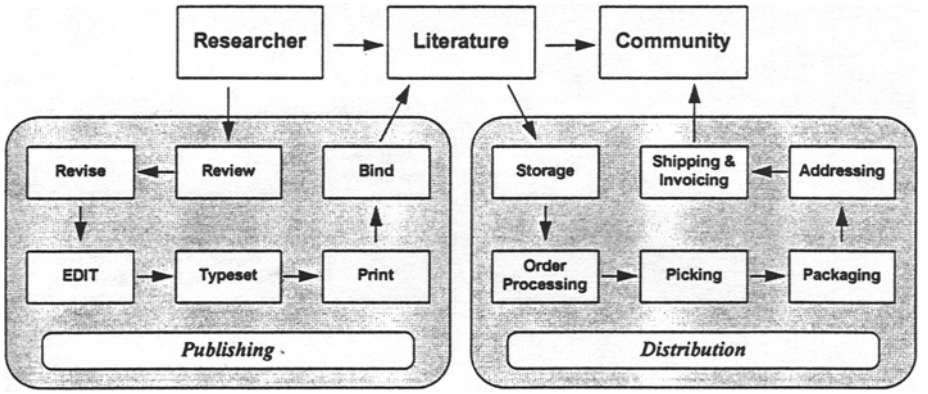

FIGURE 5. Traditional publishing requires an infrastructure to support both the publishing and the distribution processes.

With the maturation of EDP, databases and standard literature will become parallel processes in scientific publishing, with databases leading, not following, the printed page. In some areas, this is already beginning to happen. For example, a fully editable copy of the Genome Data Base was on site, at the Eleventh Human Genome Mapping meeting, where findings were entered into the database as they were presented. When the meeting ended, the updated database was transferred to its home in Baltimore and put on-line for the world to use. Within 48 hours of their presentation, abstracts from the meeting, and the underlying data, were available to researchers around the world.

Publishing Infrastructure

Although the infrastructure of traditional publishing may be invisible to researchers, editors know that considerable work is required for both the publishing and the distribution of printed communications (Figure 5).

Almost by definition, the role of the traditional editor is confined to the publishing side of the operation. Once the bound journals are ready for shipment, editing is finished and only distribution remains. This is, as we shall see, in distinction to the situation found with electronic data publishing, where the need for editing can be continuous because the processes of authorship and review are also continuous.

Building and distributing databases also require infrastructure. The process of designing and populating the database is analogous to publishing a journal, and providing public access to the database is a form of distribution (Figure 6).

The relative emphasis on particular infrastructure components varies for different databases. Initially, GenBank was just a structured text file, assembled using plain hard work. Distributing the data meant providing a copy of the file on tape. Using the file was difficult, and a third-party industry emerged to

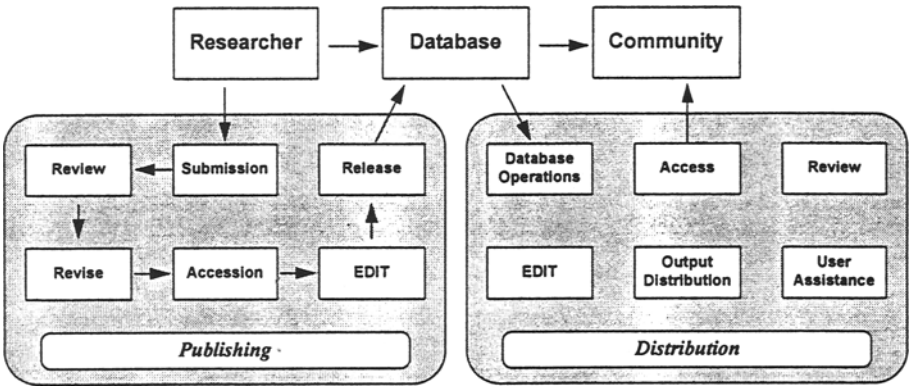

FIGURE 6. Like traditional publishing, EDP requires an infrastructure to support publishing and distribution. However, distribution for EDP is not a serial process, but rather a number of activities carried out in parallel and to different degrees for different users of the system.

provide software tools for manipulating the data. Support for editing the data was limited to a few software tools that helped staff on site at Los Alamos modify the file.

About the time that *AuthorIn* was developed, GenBank at Los Alamos was redesigned and the database was moved into a relational database management system (RDBMS). This allowed the development of software tools for interactive editing. Now, the flow of data into the database is a fairly smooth linear process that permits the staff to direct their attention to editorial issues, not technical ones.

The use of a commercially available RDBMS greatly increased the options for data distribution. Structured text files, created laboriously in the past, can now be generated routinely as simple reports. More importantly, the RDBMS is a "client-server" system, which allows the data to reside on one computer (the data server) while the user-interface software resides on another (the client), which can be far distant. Customized client software can run on a local system but access the main database in real time. The GSDB team at Los Alamos has also made a server directly accessible by third-party developers so that sites with special needs can develop their own custom software to manipulate the data on the GSDB server.

Examples of Electronic Data Publishing

Here we discuss two examples of present electronic data publishing: the Genome Data Base, a single project that is a key part of the international effort to map and sequence the human genome, and Gopher, a loose collaboration of thousands of sites worldwide that collectively provide information on topics from accounting to zoology. These two EDP activities illustrate the diversity present in electronic publishing. Both now provide valuable information to the

scientific community and both pose interesting editorial and publishing challenges.

The Genome Data Base

The GDB™ Genome Data Base is located at the Johns Hopkins University and funded by the U.S. Department of Energy and by NIH's National Center for Human Genome Research. GDB collects, manages, and disseminates the nonsequence data generated by the HGP. GDB contains information on genetic loci and probes (reagents used to identify regions of the genome), genetic maps, citations (bibliographic data, including abstracts), mutations and polymorphisms, populations, contacts (people—GDB is in part a giant Rolodex), genetic libraries (collections of cloned material), and cell lines. OMIM™ is also part of GDB.

Until 1991, GDB data entry occurred only during annual meetings. However, the HGP has increased the rate of data generation and the need for real-time access to genome data. Now data entry is continuous and increasing rapidly, with more new genes added in 1992 than were added at the previous three annual meetings combined. Fuller discussions of GDB are available elsewhere (Cuticchia et al., 1993; Jacobson, 1991; Pearson, 1991a, 1991b; Pearson et al., 1992).

Most users access GDB by connecting to the computer system at Johns Hopkins, either by Internet or by telephone. However, because GDB data are of interest to researchers worldwide, the database has been "published" through the establishment of remote, read-only nodes at many sites. Currently there are several in Europe and one each in Japan and Australia. Editing takes place only on the computer in Baltimore, but editors can work from anywhere in the world, provided they can establish a connection to the GDB computer. Editing routinely occurs from sites around the world.

Although GDB contains hundreds of relational tables, users work at a higher conceptual level through a few forms-based managers, with a separate manager being available for each data type. When the user selects the data manager for the data type of interest, the first data retrieval screen appears (Figure 7).

The bottom panel shows the screen that appears when locus manager is selected. By filling in appropriate values, the user can request the return of data about any subset of genetic loci in the database. Here, the screen is filled out to ask for all of the loci and anonymous DNA segments on chromosome 21.

When a retrieve is executed, the data are first returned in "table view," which means that the user is shown a table consisting of one summary line for each retrieved object. By highlighting a record and then toggling into "detail view" the user can see all of the information about a particular entry (Figure 8).

Relevant information from other managers may be gathered through "intelligent chaining." That is, calling a second manager from the first automatically

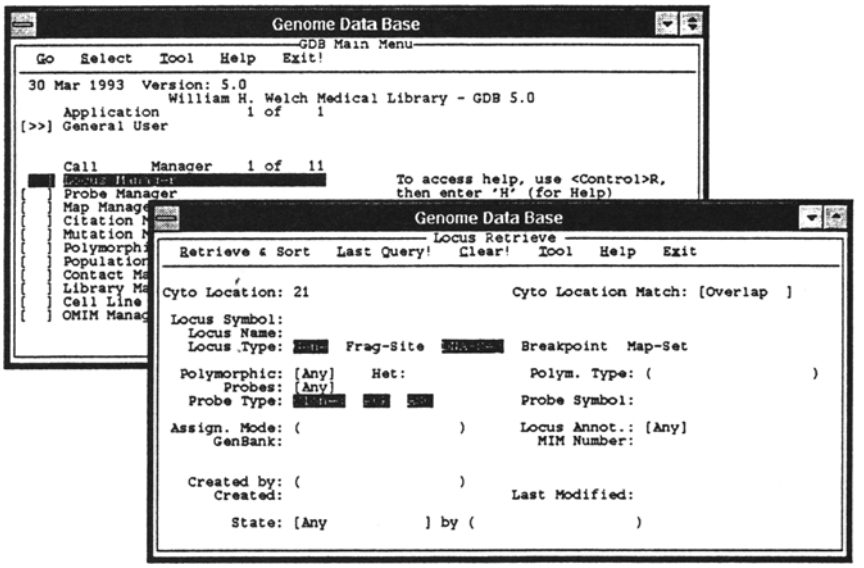

FIGURE 7. GDB uses a mouse-controlled, forms-based interface. The top panel shows the main GDB screen from which the user can select the data manager of choice. The lower panel shows the query screen that appears when the locus manager is chosen.

conditions the call based upon the retrieved information in the present manager. Thus, a call to the probe manager from this screen would be a request to retrieve all of the known probes that interact with the currently selected loci. By appropriately cascading many such conditioned calls together, the user can develop subtle and precise queries of the database.

Building such queries can be a powerful way to interrogate the system, but it can also be tedious, especially if the user needs to repeat a query regularly to detect updates and additions. To reduce the tedium, the current version of GDB supports *publication on demand.* After constructing a query, the user can ask the system to send the results via e-mail and then to *save the query and schedule it for regular re-execution and automatic distribution of the results at specified intervals.* This became available in March 1993. In the first two months of service, more than 1,500 requests were made, and new requests are occurring at the rate of 20–30 per day. At least 40 standing subscriptions are already in place.

Publication-on-demand changes the role of the database from publication to publisher. The user interacts with the interface to determine what information is available, then decides what to "buy"[4] and places an on-line order, either for a one-time publication, or for a subscription to a specified review, effectively designed by the user but carried out by the database staff and the research community as they populate the full database from which the review is extracted.

As the data volume in GDB grows, the database will increasingly become a

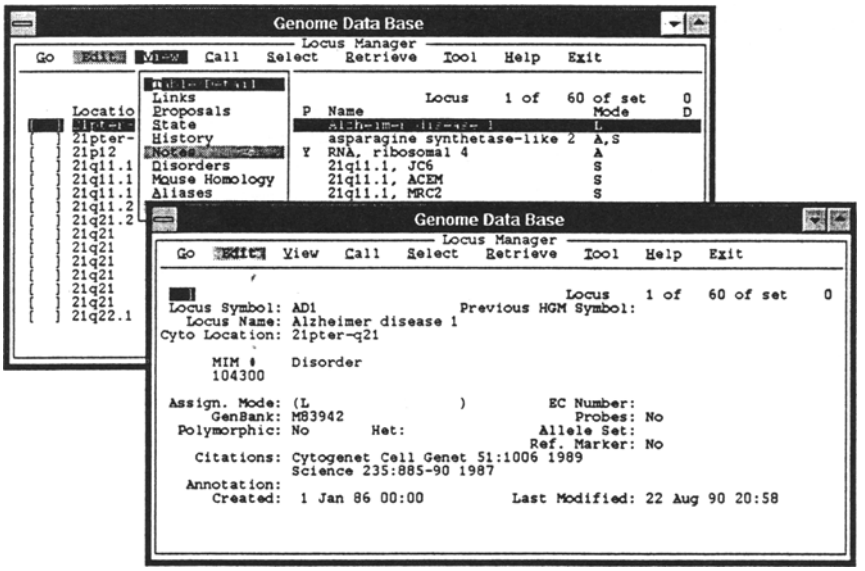

FIGURE 8. The response to the query illustrated in Figure 7. The top panel shows the first screen of the "table view" format. The pull-down menu is used to switch to the "detail view" format. The bottom panel shows the entry for the Alzheimer's gene in "detail view" format.

resource from which desirable extracts are published, either on the demand of individual users, or as periodic standard reports. Even now, GDB regularly provides thousands of pages of typeset extracts for reviews of the state of the genome, such as those undertaken at annual Human Gene Mapping workshops or Chromosome Coordinating Meetings. These extracts, and others, are available on-line as PostScript files or as parsable data files.

Gopher

Gopher is an international collaboration hosted by voluntary contributors worldwide. Gopher servers use readily available software and reside on thousands of computers around the world. Gopher is a distributed client-server system that is accessed through a copy of the client software, which must be installed on a networked local computer. Servers all over the world can be interrogated from the same client. Transfer from one server to another is trivial, and data retrieval is point and click.

Gopher clients are available in a number of formats, ranging from a simple, menu-driven system that can support an old VT-100 terminal, to graphical systems that can run on PCs, Macintoshes, or X-windows systems. Whatever its style, a Gopher client presents the user with a uniform view into (literally) a world of data. The basic Gopher interface is the simple menu. Every menu choice in Gopher either (1) retrieves another menu, (2) retrieves text, data, graphics, software, or other files, (3) initiates a query directed to a specific

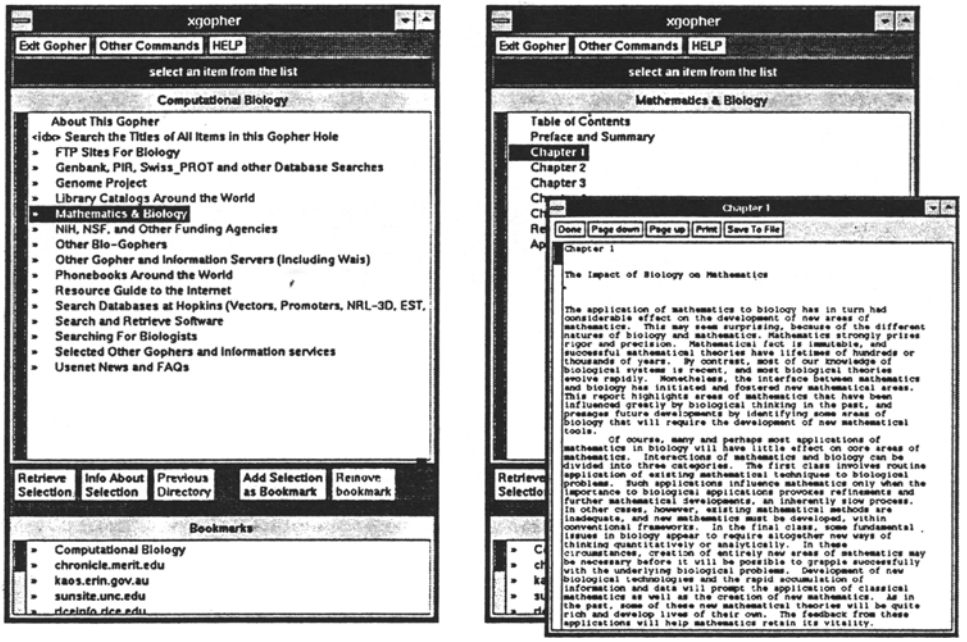

FIGURE 9. Gopher provides a menu-driven interface in which the user executes menu choices to retrieve either additional menus or files of interest. The image on the left is the first screen from the Gopher server at Johns Hopkins Applied Research Laboratory. Next is the menu that results when the "Mathematics & Biology" choice is made. Finally, the last panel shows what is returned when the user selects "Chapter 1" from the Mathematics & Biology menu.

database, or (4) initiates a search for more menu items. The power of Gopher lies in the invisibility of its infrastructure to users, who feel they are just making choices from options presented by a single system when in fact they can be jumping from computer to computer, around the world, without even noticing.

Gopher supports many different styles of interaction with the system. In general, Gopher can be used to (1) access a single server to obtain its resources, (2) interrogate a particular database, (3) roam through "GopherSpace" (i.e., all known Gopher servers), browsing for anything of potential interest, (4) retrieve specific data, text, graphics, or software, or (5) search GopherSpace to locate particular information resources. GopherSpace is large: as of April 1993, there were more than 1,250 known Gopher servers, containing more than 1.5 million unique items for retrieval.

At Johns Hopkins, we established a Gopher server[5] about a year ago with one initial purpose—to make available electronic versions of the chapters from the book *Mathematics and Biology*. Because the response was so positive, we have added many additional services so that now ours is one of the larger biologically oriented Gopher servers in the world and the "Mathematics & Biology" choice can almost be missed among the many other choices on our first menu (Figure 9).

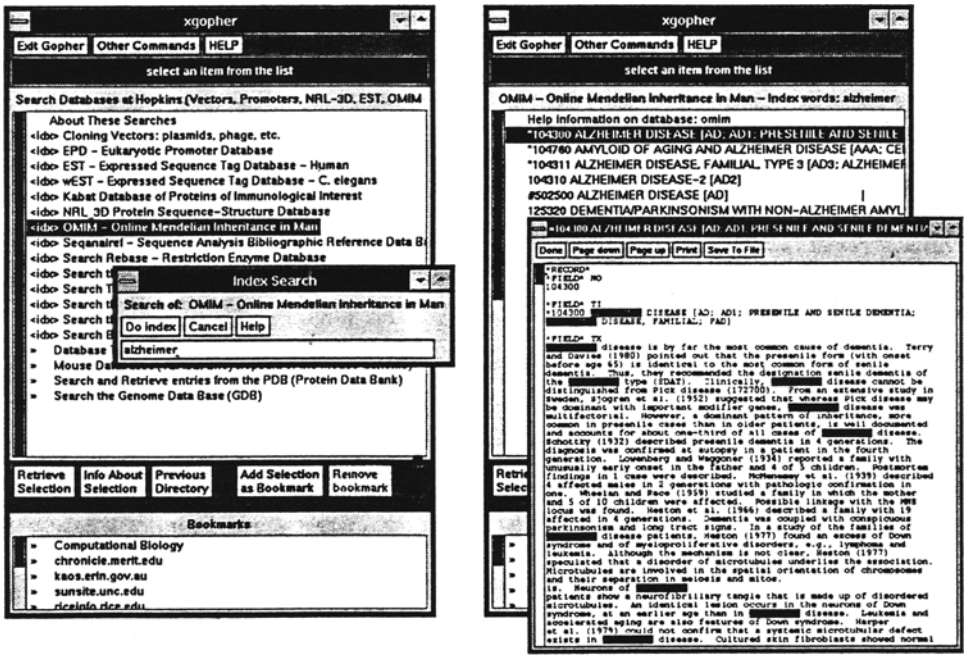

FIGURE 10. Individual databases are easily interrogated on a Gopher server. Executing a menu choice for the database brings up a dialog box, in which the terms of the search are entered. After the search is executed, another menu is generated and presented, with one entry for each item found that satisfies the query. Choosing an item results in that item being retrieved and displayed.

In the last six months, our services have expanded so that the server now contains more than 1,600 menu choices distributed over 85 menus and our databases contain tens of thousands of entries. Currently, our server handles about 4,000 transactions per day, with many coming from outside the United States. To date, requests have been logged from more than 40 countries, representing every continent except Antarctica.

Using Gopher to retrieve information from one server is easy. For example, choosing Mathematics & Biology from our top menu brings up another menu listing the book contents. Chapters chosen from this menu are returned in a separate window.[6] Any retrieved file can now be read on-line, printed, saved as a local permanent file, or discarded.

Searching a database is equally simple. One database available on our Gopher is OMIM, a collection of essays on human genetics. Retrieving a particular essay involves a few mouse clicks and typing a simple query. First, OMIM is selected from the "Search Databases at Hopkins" submenu. This results in the presentation of a dialog box in which the query, say, "Alzheimer," is entered (Figure 10). The search returns a custom menu, listing every OMIM entry that uses the word "Alzheimer" at least once. Clicking on the first choice returns the essay as shown.

Roaming GopherSpace is no more difficult. If the "Other Bio-Gophers" choice

is made on our top menu, the user receives a long menu of options that allows jumps to menus of other Gophers with a single click of the mouse. For example, selecting the Genethon Gopher retrieves a menu from Paris and displays it on screen in just a few seconds. From the user's perspective, all of GopherSpace is covered by one large, nested menu system, against which choices are made and results obtained.

One potential problem with Gopher is the size and dynamism of the resource. How do you find something if you do not know where to look? Soon after Gopher was developed, this problem was recognized and a new service was added: veronica—a system that lets you search all of the Gopher menus in the world with a single query and then returns a custom menu (sometimes huge) containing all of the menu choices that matched your request.[7] Initiating a veronica request is just like initiating a search of a dedicated database. A menu choice is made, the search string is entered into a dialog box, and the search is launched. The power of this is demonstrated with a test that I performed recently.

Nearly twenty years ago, I heard Richard Lewontin give a seminar in which he produced a particularly apt Shakespearean quote that began with the boast, *"I can call spirits from the vasty deep,"* to which the rejoinder was something like, "Big deal. I bet they don't show up." Doubting the accuracy of my recollection, especially the "big deal" part, I have long wished to obtain the correct quote. But how? A reading of Shakespeare's complete works would of course have sufficed, but the heft of the volume has always proved too intimidating.

Electronic publishing offered a solution. Assuming that Shakespeare's works would surely be on-line somewhere, I gave Gopher a try. A veronica query for "Shakespeare" produced more than 100 choices, including one with a searchable complete works on line. Sending a query to that system consisting of "vasty deep" produced the answer. Lewontin's quote was from *Henry IV*, Part 1, Act 3, Scene 1, Lines 51–53:

GLENDOWER	I can call spirits from the vasty deep.	51
HOTSPUR	Why, so can I, or so can any man;	52
	But will they come when you do call for them?	53

After twenty years, I had my answer in less than five minutes. With Gopher, anyone can call information from the vasty deep. And it does come when you do call.

I now have my quote. But is it accurate? Using Gopher is so transparent, I do not know, literally, where on earth the answer to my question originated (there is a way to find out, by asking Gopher of course, but I didn't bother). Is the source reliable? Can it be trusted? Are the contents of its system well edited and maintained? Until I know the answers to questions like these for every Gopher source that I use, I will remain slightly distrustful. Would I use Gopher material in a scholarly publication without checking further to verify it? Perhaps not. Or, perhaps so, provided I was confident in the editorial policies of the particular Gopher from which I obtained the material.

Discussion

Electronic data publishing is here, and Nobel laureate Walter Gilbert has argued that access to such resources is changing the way biology is done (Gilbert, 1991). However, many questions remain. Can the information in EDP systems be trusted? Will EDP lead to conceptual consensus, as does the traditional literature? What, conceptually, should be kept in EDP systems? How could one deal with the need for continuous editing? Will EDP authorship ever be seen as equivalent to paper authorship? Will EDP infrastructure become as efficient as that for print literature? And finally, how will EDP become an edited communication system, with established editorial policies and procedures?

Certification as a Publishing Function

Scientists must have reliable scientific literature with which to communicate, and this is possible only through the establishment and use of professional editing standards. A primary service provided by a traditional scientific journal is the *certification* of the results presented in its papers. Scientists are always more willing to accept, or at least to take seriously, findings presented in journals with well-established editorial policies and a reputation for stringent review. Findings presented in papers with little or no editing and review are not treated with the same respect.

The same holds true for EDP. With major, unitary projects such as GDB or GSDB or PIR, the issue is more clear. Scientists can become familiar with the editorial policies of a given electronic resource and can make judgments accordingly. With something like Gopher, the situation is more difficult. There is no central authority for all of GopherSpace, and stable management exists for very few individual Gopher sites. Resources come and go, making tracking difficult. Such ephemeral EDP systems have much in common with vanity presses, making user caution especially important.

However, Gopher sites are now being developed that are official outlets for professional societies or other professional organizations. Such Gopher servers will emerge as sources of quality information, with the managing society providing the same certification role for electronic information as it does for printed communication.

For example, the American Physiological Society is leading the way, having established its own Gopher server. The official announcement for that Gopher (obtainable from the Gopher) reads:

> **APS INFORMATION SERVER**
>
> The American Physiological Society (APS) Information Server provides for the electronic distribution of APS information, documents and publications via the National Research and Education Network (NREN)/Internet.

> This server permits the APS to systematically begin integrating its services and publications into the new informational infrastructure being spawned by the High-Performance Computing Program of the United States. . . .
>
> By implementing the APS Information Server, the American Physiological Society recognizes that a critical change is taking place in the way scholarly information will be gathered, archived and shared in the future. This recognition and associated actions by the Society will insure that APS is an active participant in the development of the new National Informational Infrastructure. Active participation by APS in this process will insure that its membership services are appropriately adapted to this new environment.
>
> The APS Information Server is operated for the American Physiological Society by the
>
> Office of Academic Computing
> University of Texas at Houston
> Health Science Center
> P.O. Box 20708
> Houston, Texas 77225
>
> E-mail: aps_server@oac.hsc.uth.tmc.edu

As more societies follow the lead of the American Physiological Society, users will begin to distinguish among publications on different Gopher servers the same way as they now distinguish between articles from the *Proceedings of the National Academy of Sciences* and the *National Enquirer*. The key will be the establishment of well-defined editorial and review policies for individual Gopher servers.

Scientific Publishing as Consensus Building

Quality editing allows scientific knowledge to grow, while preserving the skepticism that is central to scientific progress. Editing, however, is not censorship. While not prohibiting the publication of iconoclastic views, it does subject them to careful examination. John Ziman (1978) has described the situation, and the crucial role of editors, well: "[I]t is proper that [extraordinary] claims be examined seriously by competent experts, in case there is something in them. Since there is no official accrediting agency for 'scientific' knowledge this responsibility falls on the editors and referees of reputable scientific journals."

Ziman argues that the primary goal of scientific literature should be the achievement of consensus. This requires that established standards be maintained, both in editorial policy and in decorum. Words chosen for their rhetorical flourish or their stylistic grace may charm the already convinced, but they

are not as effective in generating consensus as words chosen for clarity and precision. Argumentation and bombast may be entertaining, but they are counterproductive in consensus building.

In the past, it was common to find scientists publicly maligning and ridiculing each other's work, as in Ernst Haeckel's response (1897) to a critic: "I find [no] reason to answer Semper's polemic on 'Haeckelism in Zoology' . . . ; for, apart from his defective education and his insufficient acquaintance with the whole subject of Zoology, this 'gifted' zoologist is so much at variance with logic, as also with truth, that refutation seems superfluous." Although such *ad hominem* outbursts are almost never encountered today in formal literature, they occur with sufficient frequency in some forms of computerized communications that they have been given a name—"flames." These electronic communications would benefit from a steadying editorial hand to help elevate them to the status of respectable literature.

Database Contents

Should a biological database be a compilation of scientific *truths*, or should it be a collection of scientific observations? The notion of a compilation of facts is appealing, so that one might consult the database to determine *the* amino-acid sequence of human beta-hemoglobin, or *the* map location of the beta-hemoglobin gene. But scientific "facts" have a way of changing with more scientific observations, and the growing burden of constant editing to achieve accuracy and internal consistency would be difficult. Ziman (1978) has made a relevant observation, although not in the context of database publishing:

> Science continually evolves. Scientific knowledge is under constant revision in the light of new evidence. From a practical point of view, it is not the ultimate truth of the scientific world picture that matters, but the [current] scientific answers to particular questions. . . . There is no *Encyclopedia* where *all* well-established science, and only well-established science, may be consulted. If such an institution existed, it would be in constant agitation, as new information was being added, and old facts and assertions struck out.

Building a database of scientific truths would be equivalent to creating an electronic version of Ziman's Encyclopedia of all well-established science. Maintaining perfect consistency in such a database would require that every existing entry in the database be checked for continuing validity every time any new entry is made. Even with a linear flow of new data, this seems impossible. Also, assertions about the real world may be initially believed, then rejected, then accepted again, albeit in a modified form. Catastrophism in geology is an example. Thus, maintaining a database of scientific truth would be an editorial nightmare, if not an outright impossibility.

Building a database of scientific observations is equivalent to creating an electronic version of the primary literature. Individual entries (i.e., published pieces) are stand-alone contributions, and there is no guarantee of consistency among pieces. Each published piece would have a recognized author and have been subjected to some form of editorial review, which would guarantee not its global truth but rather its adherence to current scientific practices. But even the primary literature is not without its problems. Ziman again:

> Amongst professional scientists, the corpus of what is called the *literature* of a subject consists of *papers* published in reputable *journals*, catalogued regularly in, say, an *abstract journal*. But the layman who attempts to consult all the papers relevant to a particular scientific question is soon wearied and appalled by the confusion and diversity of fact and opinion that he will find. At the research frontier, scientific knowledge is untested, unselected, contradictory and outwardly chaotic; only the expert can read, interpret and weigh such material.

The presence of contradictory information can make primary scientific literature difficult for the novice. Experts filter the literature in part through their personal knowledge of the field and its practitioners and in part through their familiarity with the editorial policies of different journals.

Without the existence of journals of differing editorial policies, some important but iconoclastic findings might never be published. Currently, building databases is very expensive, so that having an unlimited number of databases treating the same subject seems impractical, if not impossible. This constraint will call for some cleverness in database design, if they are truly to play the role of primary literature.

Could a single database ever possess a single editorial policy that was broad enough to support widely divergent scientific views but that was also clear enough to allow knowledgeable filtering? If scientific databases remain expensive, could a single database develop some method for simultaneously supporting multiple, well-defined editorial views? One possibility would be to have a unified database of entries, but with individual entries being coded as having passed the editorial review of different editorial bodies. This might be equivalent to a complete bibliographic database that would contain all articles published in all English-language periodicals, but which would allow users to request that articles from certain sources, say *National Enquirer*, not be returned in answer to queries.

In short, more attention needs to be directed toward defining, at a fairly abstract conceptual level, the contents of electronic data publishing systems.

Continuous Editing

Even databases of scientific observations benefit from constant editing. For example, it is difficult to justify filling a database with errata notices correcting

simple errors, when the actual entries can be updated. If this is done, however, then data items previously retrieved may not be locatable again. To assist with filtering data, perhaps the epistemological status of entries should be flagged periodically by reviewers to assist in data filtering by users.

For these and many other reasons, there will always be a need for some continuous review and editing of the entries in a database. This poses interesting conceptual challenges for a scientific publication, especially since the notion of continuous editing of print media is usually associated with the workings of a police state, in which efforts are made to keep all publications in line with current dogma. The Orwellian strangeness of this continuous editing is captured perfectly in the opening lines to Milan Kundera's novel *The Book of Laughter and Forgetting* (Kundera, 1981):

> In February 1948, Communist leader Klement Gottwald stepped out on the balcony of a Baroque palace in Prague to address the hundreds of thousands of his fellow citizens packed into Old Town Square.... Gottwald was flanked by his comrades, with Clementis standing next to him. There were snow flurries, it was cold, and Gottwald was bareheaded. The solicitous Clementis took off his own fur cap and set it on Gottwald's head.... The Party propaganda section put out hundreds of thousands of copies of a photograph of that balcony with Gottwald, a fur cap on his head and comrades at his side, speaking to the nation.... Four years later Clementis was charged with treason and hanged. The propaganda section immediately airbrushed him out of history and, obviously, out of all the photographs as well. Ever since, Gottwald has stood on that balcony alone. Where Clementis once stood, there is only bare palace wall. All that remains of Clementis is the cap on Gottwald's head.

Supporting continuous editing, while avoiding the scientific equivalent of Clementis' cap, is a standing challenge to electronic data publishing. That the answer to the question, "Was Clementis in the picture with Gottwald," should depend upon the political context in which the question is asked does not sit well with a scientific mind. Either he was there or he wasn't. Either the picture is doctored or it is not.

Let us consider a biological example. Suppose that Jones reports a DNA sequence to be AATCGA, but the database staff mistakenly enter the sequence as ATACGA. When the mistake is discovered, should the original entry be updated, or should a separate erratum entry be made? Later Jones discovers a laboratory transcription error and resubmits the sequence as AATGCA. What kind of change is appropriate here: an entry update or an erratum entry? Suppose that later yet, Smith discovers that with the equipment used by Jones real sequences of AAAT are almost always reported as AAT. Now what?

On the other hand, one might argue that, once released, electronic database entries, like the pages of a printed journal, must stand for all time in their original condition, with errors and corrections noted only by the additional

publication of errata and commentaries. However, this might quickly lead to a situation in which commentary outweighs original entries several fold. On the other hand, occasional efforts to "improve" individual entries could create a slippery slope leading toward Clementis' cap. Many potential solutions exist, but this is not the place to discuss them. The important point here is, if a "publication" is never finished, if it is always a work in progress, the notions of authorship, editing, and review begin to differ from those associated with print publications.

Authorship in EDP Systems

At the present time, the concept of authorship is not well established in most electronic data publishing systems. Although a few, such as OMIM, do consist of individually authored essays, most are seen as compilations with authorship being associated only with the literature citations connected to the data objects in the system. However, there are some trends away from this. GSDB has been moving more in the direction of making the on-line editing of entries available to the submitters of those entries. GDB considers genetic maps to be individual objects of intellectual creation with known authors. Entries in Flora North America have named authors and larger sections have named editors. Other databases are also naming curators with responsibility over certain subsections of the database. Some Gopher servers are now mounting collections of essays or other authored pieces.

One of the major impediments to establishing an authored electronic literature is author recognition. In a publish-or-perish world, the urge to publish is associated with the urge to receive credit and intellectual standing. Presently, almost no one would think to put a database submission on his CV as evidence of intellectual output, and many young scientists have been told explicitly by their mentors that involvement with electronic publishing can be a career stopper.

A major source of the perceived illegitimacy of electronic publication is its similarity to vanity press publishing. So long as EDP seems to have little or no editorial oversight, contributions to EDP will have little value. As edited systems become more prevalent, this will change.

Implementing Systems

The effort required to build an EDP system can vary significantly. Major projects like GenBank and GDB use custom software that required many person-years of effort before they became operational. Both require much equipment, large staffs, and budgets to match—millions per year to operate. Such information-infrastructure efforts can be justified only if they support an appropriately large scientific superstructure. Although such projects play and will continue to play crucial roles in some biological fields, their expense and size means that there will never be many of these established.

Gopher systems, comparatively speaking, can be built on a shoestring. Both server and client software are generic and freely available. A dedicated person, working only part time and using inexpensive computer equipment, can develop a respectable Gopher server. A few full-time workers with access to better computer equipment can produce a stunning information resource. There has already been a remarkable proliferation of Gopher servers, and the pattern continues. The minimal expense and relative ease with which these may be established means that Gophers, or some newer Gopher-like system, will likely play a major role in transforming electronic data publishing from a relatively obscure process into a new, vibrant form of scientific literature. Even major projects like GDB are now using Gopher to help distribute their data.

GDB and the sequence databases need to provide infrastructure to support both the publishing and the distribution infrastructure aspects of EDP, with support for publishing (database building and editing) representing a large part of their expenses. Gopher systems, on the other hand, emphasize only the distribution side of electronic data publishing. If the data have been assembled elsewhere, distributing them via Gopher can be fairly straightforward.

Increases in the efficiency of some aspects of database publishing are still needed. When they arrive and are coupled with good methods for electronic data distribution, EDP will flourish.

Editorial Involvement in EDP

Gopher is but one example of the growing technical infrastructure that is making some electronic data publication almost embarrassingly simple to implement. Other systems like WAIS (Wide-Area Information Server) or WWW (World Wide Web) are equally important, but space does not permit their discussion here. Relatively inexpensive working systems for electronic data distribution are now available. A few more technical developments will, however, be helpful. For example, one proposed modification to Gopher would allow each "published" document to include an embedded, hidden check value that reflects the bit-for-bit contents of the file. Changes, even to a single byte, would cause the check-value test to fail, and the Gopher client could alert the user that the document was corrupt and could not be trusted.

The technical problems associated with mounting cost-effective electronic data publishing are either solved, or solutions seem in reach. What is needed now, to take us all the way into electronic data publishing as a new, formal literature, is the development of more high-quality, professionally operated EDP sites. These are coming. One Gopher at Los Alamos provides papers of standard-literature quality for the physics community. Soon, more will appear. Other professional societies will follow the lead of the American Physiological Society and establish electronic information dissemination systems. The key to making all of these become components in a new scientific literature is the establishment of appropriate editorial and review policies for electronic data

publishing sites. Editors have the opportunity and the responsibility to work in the vanguard of a revolution in scientific publishing.

Notes

This paper is an expansion of the keynote address given to the 37th Annual Meeting of the Council of Biology Editors, San Diego, California, 8–11 May 1993.

1. The amount of human sequence collected to date would amount to less than one-third of that 60-mile tag. Obviously, the information-managing challenges of the genome project are yet to come.
2. In fact, the sequence databases such as GenBank and PIR began as traditional print reviews, with computer systems used only to help the author store and manipulate the data in the preparation of print publications.
3. Although editing is normally associated with the conceptual and intellectual content of publications, technical editing is also required. Someone must verify that figures are numbered properly, that data in tables add up, that statistical tests have been selected and applied properly, etc. Similarly, there are many technical tests that can be done to analyze the validity of sequence data. The fact that critical analyses of sequences are best done by computer in no way eliminates the central role of human reviewer. Instead, the thoughts and actions of expert reviewers are captured in computer programs that are run against all submissions. The results of these human-designed, computer-performed analyses are returned to the author of the sequence. These analyses play the same role for sequences as traditional reviews do for manuscripts. They either confirm and validate the author's interpretation or they call the author's (and the editor's) attention to potential problems in the submission.
4. With GDB, there is no charge for the resulting reports. However, modifying such a system so that charges could be made would be straightforward.
5. Our server can be accessed directly by invoking its name when the gopher program is started locally: e.g., "gopher gopher.gdb.org". We are also cross listed on many other Gopher systems. Our Gopher server was built by Dan Jacobson (danj@gdb.org). His skill, knowledge, and enthusiasm have been invaluable in creating this public resource.
6. These figures illustrate an X-windows interface into Gopher. Without X-windows, all transactions occur through a single screen, but the end results are the same.
7. The names of these programs are fanciful. Gopher was so named because it was developed at the University of Minnesota. "Veronica" stands for "very easy rodent-oriented net-wide index to computerized archives"—an incredibly forced acronym that was chosen only to complement the name "archie" which is a software package that provides a similar lookup service for public ftp (file transfer protocol) sites.

Bibliography

Cantor, C. R. 1990. Orchestrating the human genome project. *Science* 248:49–51.

Cinkosky, M. J., J. W. Fickett, P. Gilna, and C. Burks. 1991. Electronic data publishing and GenBank. *Science* 252:1273–1277.

Culliton, B. J. 1990. Mapping terra incognita (humani corporis). *Science* 250:210–212.

Cuticchia, A. J., K. H. Fasman, D. T. Kingsbury, R. J. Robbins, and P. L. Pearson. 1993. The GDB Human Genome Data Base Anno 1993. *Nucleic Acids Research.* 21:3003–3006.

DeLisi, C. 1988. The human genome project. *American Scientist* 76:488–493.

Fickett, J. W. 1989. The database as a communication medium, in R. R. Colwell [Ed.], *Biomolecular Data: A Resource in Transition.* New York: Oxford University Press, 295–302.

Gilbert, W. 1991. Towards a paradigm shift in biology. *Nature* 349:99.

Haeckel, E. 1897. *The Evolution of Man, Volume I, Third Edition.* New York: D. Appleton and Company.

Jacobson, N. 1991. Mapping the human terrain. *Hopkins Medical News,* Spring: 16–19.

Kabat, E. A. 1989. The problem with GenBank, in R. R. Colwell [Ed.], *Biomolecular Data: A Resource in Transition.* New York: Oxford University Press, 127–128.

Kundera, M. 1981. *The Book of Laughter and Forgetting.* New York: Penguin Books.

Lewin, R. 1986. DNA databases are swamped. *Science* 232:1599.

Pearson, M. L., and D. Söll. 1991. The human genome project: A paradigm for information management in the life sciences. *The FASEB Journal* 5:35–39.

Pearson, P. L. 1991a. Genome mapping databases: data acquisition, storage, and access. *Current Opinion in GENETICS & DEVELOPMENT* 1:119–123.

Pearson, P. L. 1991b. The genome data base (GDB)—a human gene mapping repository. *Nucleic Acids Research* 19, Supplement:2237–2239.

Pearson, P. L., N. W. Matheson, D. C. Flescher, and R. J. Robbins. 1992. The GDB Human Genome Data Base Anno 1992. *Nucleic Acids Research* 20, Supplement:2201–2206.

United States Department of Energy. 1990. Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years.

United States Department of Health and Human Services, Public Health Service, National Institutes of Health, National Center for Human Genome Research. 1991. *Annual Report I—FY 1990.* Washington, D.C.: Government Printing Office.

United States National Academy of Sciences, National Research Council, Commission on Life Sciences, Board on Basic Biology, Committee on Mapping and Sequencing the Human Genome. 1988. *Mapping and Sequencing the Human Genome.* Washington, D.C.: National Academy Press.

Watson, J. D. 1990. The human genome project: Past, present, and future. *Science* 248:44–48.

Ziman, J. 1978. *Reliable Knowledge: An Exploration of the Grounds for Belief in Science.* Cambridge: Cambridge University Press.