Financing a future for public biological data

Lynda B. M. Ellis¹ and Doyle Kalumbi²

¹Department of Laboratory Medicine and Pathology and ²Department of Applied Economics, University of Minnesota, Minneapolis, MN 55455, USA

Received on February 15, 1999; revised on April 12, 1999; accepted on April 20, 1999

Abstract

Motivation: The public web-based biological database infrastructure is a source of both wonder and worry. Users delight in the ever increasing amounts of information available; database administrators and curators worry about long-term financial support. An earlier study of 153 biological databases (Ellis and Kalumbi, Nature Biotechnol., 16, 1323–1324, 1998) determined that near future (1–5 year) funding for over two-thirds of them was uncertain. More detailed data are required to determine the magnitude of the problem and offer possible solutions.

Methods: This study examines the finances and use statistics of a few of these organizations in more depth, and reviews several economic models that may help sustain them.

Results: Six organizations were studied. Their administrative overhead is fairly low; non-administrative personnel and computer-related costs account for 77% of expenses. One smaller, more specialized US database, in 1997, had 60% of total access from US domains; a majority (56%) of its US accesses came from commercial domains, although only 2% of the 153 databases originally studied received any industrial support. The most popular model used to gain industrial support is asymmetric pricing; preferentially charging the commercial users of a database. At least five biological databases have recently begun using this model. Advertising is another model which may be useful for the more general, more heavily used sites. Micro-commerce has promise, especially for databases that do not attract advertisers, but needs further testing. The least income reported for any of the databases studied was $50,000/year; applying this rate to 400 biological databases (a lower limit of the number of such databases, many of which require far larger resources) would mean annual support need of at least $20 million. To obtain this level of support is challenging, yet failure to accept the challenge could be catastrophic.

Contact: lynda@tc.umn.edu

Introduction

The World Wide Web delights us with a plethora of information sources. We enjoy access to its Internet resources for personal and professional gain; they exponentially increase in number and size every day. What sustains this bubble? And when will it burst?

The death of the Internet has been repeatedly forecast almost since its inception, yet the physical network appears stable for the foreseeable future. The same cannot be said for the myriad scientific resources which reside on it. Regarding increasing opportunities for linking biological databases, Teresa Atwood of University College London said: ‘All these efforts, however, are jeopardized by the uncertain funding for many smaller databases… Many funding bodies will support the creation of a new database but are less willing to fund development and continuing support for them. It’s crazy, and I hope the tide of things will change.’ (Williams, 1997).

What is the magnitude of the problem? There are hundreds of biological databases on the Internet; 410 were listed in early 1999 in DBcat, the INFOBIOGEN biological database catalog (Discala et al., 1999). They are significant information resources that represent the fundamental research infrastructure and its new ease of access have outpaced the scientific and business communities’ ability to establish economic viability for these resources over the long term.

Immediate efforts must, therefore, be undertaken to develop economic models to describe this new infrastructure, quantitatively measure its cost and value, and propose mechanisms to ensure its long-term economic viability. In May 1997, the European Biotechnology Information Strategic Forum (BTSF) sponsored a workshop on Financing Biotechnology Databases (BTSF, 1997). The workshop concluded that a framework was required to guide support database producers in their quest for financial viability. To further the development of such a framework and to bring the problem to a wider audience, in 1998 we surveyed 321 managers of biological databases throughout the world on the sources and stability of their funding (Ellis and Kalumbi, 1998). Forty-eight percent of them responded and over two-thirds of the respondents reported that near future funding (1–5 years) was uncertain.

In September 1998, a related BTSF workshop on Building and Owning Biological Databases had as one conclusion that the economics of databases are similar to those of scientific
journals; they are essential to the scientific community and should be established on a firm and sustainable footing (BTSF, 1998). We agree. Here we report in more depth on the finances of a few of these databases, on use statistics from one, and review several economic models which may help sustain the biological database infrastructure.

Methods

We carried out a review of common economic models that may support public biological databases and a benchmarking study of representative databases, including use statistics for one of the benchmarked databases.

A benchmarking instrument, five pages in length (available on request from the authors), was developed and used to collect detailed financial information from administrators of selected public biological databases. We benchmarked six organizations. Most (5) were selected from those who responded to the initial survey (Ellis and Kalumbi, 1998); all administer molecular biology databases. One non-profit, state-supported organization, labeled ‘B’ in Table 1 and 2, was included; it concentrated on supplying the information needs of small manufacturing companies in that state. The focus of our benchmarking was on the cost/revenue structure of each organization. The benchmark studies were carried out during the third quarter of 1998 and covered each organization’s 1997–1998 fiscal year.

We contacted administrators of the selected organizations and invited their participation in this second study. If they agreed, we sent them the benchmark questionnaire via e-mail or fax and, ~2 weeks later, visited them to conduct a structured interview. We visited five of the six organizations; the sixth returned the questionnaire using e-mail. Four of the five interviews were carried out by two people; one was carried out by a single interviewer.

The six organizations were founded between 1993 and 1997. Thus, like most public biological web-based databases, they are of relatively recent origin. They were chosen to represent both relatively large (1) and small (5) databases; those with both assured (2) and uncertain (4) future funding; those spanning the geographical range of surveyed databases: US/Canada (3), Western Europe (1), Australasia (1) and Other (1); and those with different sources of funding: public funding plus user fees (2) and completely public funding (4). All database administrators contacted agreed to participate, with one exception. An administrator of a biological database with industrial funding declined for proprietary reasons. Currency values were converted to US dollars if needed, using the OANDA Classic 164 Currency Converter (1999).

Developing economic models for biological databases requires current knowledge of the users of these resources. Thus, we reviewed the use statistics for one of the smaller, more specialized biological database resources which had been benchmarked, using Analog log analysis software (Turner, 1999).

Results

Benchmark studies

Table 1 shows the number of people responsible for the management and daily operation of these organizations. The personnel requirement for these organizations is as varied as the organizations themselves. Large organizations may require up to 10–20 or more full-time staff, while small organizations may get by with one or two.

As are most public biological databases, these organizations were completely or for the most part funded from public sources. Two of the organizations receive some revenue from fees for services. The percentage of total expenses that each spent on administration, non-administrative personnel, computer-related costs, and other expenses, was calculated. The average of these percentages is shown in Figure 1. Non-administrative personnel and computer-related activities are major expense items and account for 77% of the total cost. The aggregate income and expenses for each organization are shown in Table 2. There is a wide variation in income, from $50 000 to $1 040 000. For organization F (the smallest organization), income exceeds expenses. It may be the only organization studied to make a profit.

Table 1. Personnel

<table>
<thead>
<tr>
<th>Organization</th>
<th>Full-time staff</th>
<th>Part-time staff</th>
<th>Full-time equivalents</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11</td>
<td>9</td>
<td>15.5</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>5</td>
<td>3.5</td>
</tr>
<tr>
<td>F</td>
<td>3</td>
<td>1</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Table 2. Income and expenses

<table>
<thead>
<tr>
<th>Organization</th>
<th>Income (US$)</th>
<th>Expenses (US$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1 040 000</td>
<td>1 040 000</td>
</tr>
<tr>
<td>B</td>
<td>600 000</td>
<td>600 000</td>
</tr>
<tr>
<td>C</td>
<td>330 000</td>
<td>330 000</td>
</tr>
<tr>
<td>D</td>
<td>200 000</td>
<td>200 000</td>
</tr>
<tr>
<td>E</td>
<td>166 000</td>
<td>166 000</td>
</tr>
<tr>
<td>F</td>
<td>50 000</td>
<td>24 000</td>
</tr>
</tbody>
</table>
domains and 3% (13 285) unknown. Thus, from US educational domains, 13% (49 678) from other US domains, 21% (84 331) from non-US countries, 20% (79 710) of these accesses were from US commercial users with resolvable IP addresses. Approximately 42% successfully accessed a total of 393 056 times by non-local smaller, more specialized biological databases. In 1997, it was an educational institution and is representative of the many users (visitors) per month and 12 accesses per distinct user.

90 countries on six continents, for an average of 3356 new database served 40 267 distinct non-local users from were from commercial domains. During that year, the access was from the US. Of the 295 440 US accesses, 56% supported by parts of six small grants, and ‘assured’ meant that the curator felt personally able to maintain this level of grant funding over the next 5 years; this consumes large amounts of curator time.

The biological database whose use statistics are reported here, labeled ‘D’ in Table 1 and 2, is hosted by a US educational institution and is representative of the many smaller, more specialized biological databases. In 1997, it was successfully accessed a total of 393 056 times by non-local users with resolvable IP addresses. Approximately 42% (166 052) of these accesses were from US commercial domains, 21% (84 331) from non-US countries, 20% (79 710) from US educational domains, 13% (49 678) from other US domains and 3% (13 285) unknown. Thus, ~75% of this access was from the US. Of the 295 440 US accesses, 56% were from commercial domains. During that year, the database served 40 267 distinct non-local users from 90 countries on six continents, for an average of 3356 new users (visitors) per month and 12 accesses per distinct user.

Economic models
We focus on selected models of current interest, including public funding, asymmetric pricing, advertising, deal-making and microcommerce. We include, where possible, examples of their application to biological databases and end with one successful application of a hybrid model.

Public funding
As the name implies, the public funding model finances Internet resources from the public purse and provides these resources free of charge to all users. The BTSF (1997) concludes that ‘there is a clear case for the public model where the main users are also in the public domain…’. Over 90% of current support for public biological databases presently comes from public funds (Ellis and Kalumbi, 1998). Some public funds are governmental, but others include volunteer labor by database staff and support by the academic institutions which host these databases. In exchange, the databases can function as recruiting tools and advertisements for the staff and institution’s capabilities. It is difficult to put a price on these intangible benefits.

Asymmetric pricing
The concept of asymmetric pricing used for support of biological databases is simple. The database remains free to academic/public users. However, commercial users incur charge and gain added value. While asymmetric pricing can be applied to both subscription and per use fees, biological databases have so far exclusively used it for subscriptions. In early 1998, three biological databases initiated this pricing mechanism: Swiss-Prot (GeneBio, 1998), EcoCyc (Pangea Systems, 1998) and STACKS (Miller et al., 1998). Swiss-Prot, the largest database to use this model, being marketed by Geneva Bioinformatics, Inc. (GeneBio), charges up to $90 000 per year (GeneBio, 1998). More recently, TRANSFAC (Heinemeyer et al., 1999) and V BASE (Tomlinson, 1999) have also begun to use this model. Other databases, such as MEROPS (Rawlings and Barrett, 1999), require commercial users to register, but have not yet decided how or even whether to impose registration fees (A.J.Barrett, personal communication).

The registration fee paid by commercial users may give access to exactly the same information that other users receive for free. Alternatively, to motivate registration, value may be added, and the registered users become subscribers, who receive special benefits based on their registration (Senné, 1998). The benefits given to subscribers by some of the biological databases that are using this model include the right to load the database on local servers, improved search functionality, CD-ROM versions of the database, special technical support and newsletters.

The biological databases who are using this model are generally attached to an academic institution, as in the case of STACKS (University of the Western Cape, South Africa), or spun out into small companies centered around one or several databases; for example, TRANSFAC is now part of BIOBASE (BIOBASE, 1999). These organizations typically do not have either the capabilities, orientation or scale to market the database effectively. Cognia Bioinformatics, Inc. (Larvol, 1999), the distributor and service provider for the TRANSFAC database in North America, represents a new type of company which serves as a partner and an intermediary between these organizations and the pharmaceutical and biotechnology industry. Partnering with such a company...
relieves the organization from worrying about the commercial and service aspects of pursuing an asymmetric pricing strategy.

**Advertising**

On-line advertising is another Internet business model. Businesses make money from advertising by delivering an audience in return for a payment. Often, advertisers agree to pay a website a fixed amount each time a web page is viewed by an Internet user referred to it from that site (Welch, 1997). Alternatively, websites can receive a percentage of sales generated by the ad, called commission advertising. For example, Amazon.com, an on-line bookseller, will pay a commission to other websites that link to it. Payments range from 5 to 15% of the orders received through each link (Amazon.com, 1999). Many other companies also offer such commissions (Welch, 1999).

Companies advertise on a website because the site draws people that the company wants to reach. If the target audience is the biological research and educational communities, advertising in biological databases might appear attractive. However, the amount of traffic, especially at the smaller sites, is not likely to attract commercial clients. Welch (1998) states: ‘In general, sites with fewer than 10,000 visitors per month are unlikely to draw advertisers’. The biological database whose use statistics are reported in this paper averaged fewer than 4000 unique visitors per month.

Perhaps for this reason, currently there are no advertisements in public biological databases, except, of course, ads for the host organization. However, a related resource, the BIOSCI webserver (BIOSCI, 1999), which supports many biological newsgroups and e-mail lists, has solicited ads since their public funding ended in June 1996. They now have several banner ads on their main (most heavily used) web pages. Most interior pages, lacking sponsors, carry their own ‘Your Logo Here’ banner. Since the service is still in operation, more than 2 years after public funding ceased, one may infer that they are financially solvent. They state (BIOSCI, 1997) that they will solicit ads until they obtain an operating reserve of $100,000 and then would cease until funds were again needed. Be that as it may, they continue to send monthly requests for support to each of their newsgroups.

**Deal-making**

According to James Moore (1996), current successful business behavior is described as a business ecosystem. A successful business ecosystem is built on deals made (value received) by all stakeholders (producers, users, suppliers, etc.) in an industry. Biological databases could in theory make deals with content providers, linked databases and/or users. The commission ads mentioned earlier are one type of deal. Others include reciprocal linkage (‘ads’ which neither side pays for), data validation (staff of one database report errors in a database they link to) and payment in kind (give commercial data or subscriptions free in exchange for data validation or similar services).

**Microcommerce**

The microcommerce model involves the use of technology to handle efficiently transactions as small as—and even smaller than—1 cent, over the Internet. One example is the MilliCent System (Compaq Computer Corp., 1998a). A 9 month successful trial of MilliCent was recently concluded. During the trial, $110,000 was placed into circulation by 10,000 consumers. They bought on-line content from 45 vendors from around the world, with a median transaction price of 2 cents (Compaq Computer Corp., 1998b). Although none of these vendors sold access to scientific databases, Oxford University Press sold dictionary searches for 0.2 cents each and Vorarlberg Online, an Austrian German-language news service, sold searches of a database of newspaper articles and access to the articles for from 1 to 2 cents each (Compaq Computer Corp., 1998c). The organizers of the trial suggest that ‘microcommerce is of most interest to websites that are unable to develop viable advertising revenue streams’ (Compaq Computer Corp., 1998b).

**A successful hybrid**

The ChemFinder chemical database (Brecher, 1998) is a good example of a public database which has paid its own way from the start. ChemFinder started operation in 1995, owned and operated by CambridgeSoft Corp., which publishes scientific software and databases. The ChemFinder database website presently has ads for CambridgeSoft products, ads from other vendors, and links which allow on-line purchase of CambridgeSoft products, such as subscriptions to a value-added web version of ChemFinder with improved search and other capabilities, and CD-ROMs containing the database (academic discounts are available for this). It has ‘made deals’ with some of the other databases it links to, informing these others of errors in their information, acknowledging errors these others report, and correcting them quickly. This hybrid model works well for ChemFinder, a high-volume, general-purpose, scientific database.

**Discussion and conclusions**

**Benchmarking**

Benchmarking, the study of organizations with recognized excellence in the area of interest, is a widely used and important strategic business tool for uncovering and understanding possible paths to organizational improvement. Our benchmark study covered a representative sample of

---

L.B.M.Ellis and D.Kalumbi
public biological databases, i.e. members of the biological database peer group. However, Tucker et al. (1987) state that, in order to move in a new direction (such as toward economic sustainability of biological databases), it is important to collect benchmark information outside a group of industry peers, to do ‘wide benchmarking’ to find those solutions that have not penetrated the peer group seeking improvement. In this way, new to the group solutions can be translated from one industry group to another and the possibility for the total innovative solutions being developed is greatly enhanced. Furthermore, the study of non-peer operations is a valuable tool for eliminating the interfering bias present in benchmarking of a group from a single industry or organizational type.

Accordingly, we attempted to benchmark members of the non-peer group. However, three we contacted were not cooperative. One of them openly refused to take part in the study for proprietary reasons. The others did not respond to our repeated inquiries. As stated above, even the administrator of a peer biological database that received some funds from industry declined to participate in the benchmark study for proprietary reasons.

This is not surprising; Internet commerce is still in its infancy. Commercial firms are currently unwilling to give up any information which might make the different between success and failure in such a competitive and potentially lucrative field as electronic commerce.

**Economic models**

The public model will work best for those databases that are vital to all (or a large part) of the scientific community. It is simple and can reduce bureaucracy, creates equal access to all users, can reflect public policy and can encourage exploitation of information by others. However, it has some serious drawbacks. The model is unresponsive to the market, it can threaten commercial operations, it cannot meet all needs and it is hard to know whether the public is getting value for its money (Cameron, 1997).

The BTSF (1997) states, that, even if public funding is used to create data, ‘… there is nothing politically wrong with then allowing this basic data to be used by others in added value products which can indeed be charged for’. Asymmetric pricing is the phrase we use to describe such a model involving registration, subscriptions, added value and/or licenses. Despite the familiarity of registration or subscriptions among both buyers and sellers, the asymmetric pricing model has several drawbacks: subscriptions are expensive to administer and purchase, impulse buying from infrequent users is discouraged, subscription access codes can easily be inappropriately shared by multiple users, and subscribers can be frustrated by the need to remember their access codes (Compaq Computer Corp., 1998d).

Advertising may help the most general, heavily used, sites. The deal-making examples given here appear to be able to generate only modest income for a database. It remains to be seen if it is possible to make more innovative deals which might offer more support. Although the microcommerce model is interesting and its developers state that it might be useful to sites which lack the traffic needed for ads, microcommerce is in its infancy. It needs more testing in the commercial sector, to see if users are willing to make even micropayments for what they may have formerly received for free, and if micropayments can meet macroexpenses.

For the US database whose use statistics were reported here, non-local use from resolvable IP addresses is ~75% US and 25% elsewhere in the world. Looking more closely at US accesses, a majority (56%) is from commercial users, yet these users pay little, if any, in support of this or other biological databases. Only 2% of public biological databases around the world have funding from commercial sources (Ellis and Kalumbi, 1998). Database administrators are attempting to obtain more funding from commercial sources using an asymmetric pricing model. Does this model work? Will it be stable over the long run? Is it adaptable to smaller public biological databases? It was first implemented in 1998; only time will tell.

**Conclusions**

Over two-thirds of biological databases face uncertain funding in the very near future and the biological database community attaches great importance to this problem. The problem is worldwide and expected to increase with continuing pressure on public funds and increasing numbers of databases, all demanding support (Ellis and Kalumbi, 1998). Financial data from a few representative databases show that most expenses are related to computers and non-administrative personnel, budgets range from $50 000 to over $1 000 000, and even databases whose funding was self-reported as ‘assured’ may not be in the best of financial health.

Several economic models to address this problem have been discussed. Which model(s) offer greater scope in addressing the financial sustainability of public biological databases is an open question. Except for asymmetric pricing, most of the models discussed have not been applied to public biological database organizations, or may even not be applicable.

If a small public database requires $50 000/year to maintain, applying this rate to the 400 databases in DBcat (a lower limit to the true number of such databases, many of which require far more resources) would require annual support of at least $20 million and could easily need twice that much. To obtain this level of support is challenging, yet failure to accept the challenge could be catastrophic.
Acknowledgements

We thank Jeffrey Tate and Doug Hershberger for assistance with development of the benchmark instrument and with the benchmark interviews. We thank them and Teresa Attwood, Alan Barrett, Jonathan Brecher, Mark Cantley, Jack Franklin, Theresa Gaasterland, Winston Hide, Roger Hurd, Minoru Kanehisa, Bruno Larvol and Ron Tarro for helpful discussions. This work is supported in part by National Science Foundation Award DBI-9753282.

References

All URLs were correct as of 2 February 1999.