A Century of Science Publishing
E.H. Fredriksson (Ed.)
IOS Press, 2001

Chapter 19
Biological and Medical Publishing via the Internet

Matthew Cockerill
BioMed Central Ltd., London, UK

The early days

Biomedical researchers have been enthusiastic users of web technology since its early days. Even before the release of the first usable web browsers in 1995, scientists were downloading software from biological ftp (file transfer protocol) archives, accessing gopher servers (precursors to web servers) to search databases, and using email both to communicate with colleagues and to run sequence comparisons against biological databases such as *EMBL* in Europe, the *DDBJ* in Japan, and *GenBank* in the USA (http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html) [1,2].

The exponential growth which these nucleotide and protein sequence databanks have experienced over the last 20 years (Figure 1) probably in large part explains biologists early enthusiasm for the Internet. It quickly became impractical to physically distribute such databases, both because of the amount of data involved and the frequency with which new data was being added. Making use of the data remotely via a network was a far more efficient solution.

When the web began to take off in 1994–1995, biologists (especially bioinformaticists) took full advantage, and biological websites of various kinds sprang up, almost all of them non-commercial, and maintained by scientists in their spare time. At this time the web was still predominantly non-commercial nature, and one of the most trafficked scientific pages on the Internet was Pedro's Biomolecular Research Tools (http://www.public.iastate.edu/~pedro/research_tools.html), maintained by a graduate student, which kept track of many of the most useful biological web pages and online sequence analysis tools.

Bibliographic databases

Since the late 19th Century the US National Library of Medicine (NLM) has been compiling bibliographic details of medical research articles published each year into a printed publication, *Index Medicus*. In the 1940's, an alternative to *Index*

Medicus arrived in the form of *Excerpta Medica*, now owned by Elsevier.

Index Medicus has been distributed electronically since the late 1960's as the *MEDLINE* database, which now contains more than 10 million records. Similarly, *EMBASE* is the electronic version of *Excerpta Medica*. Neither *EMBASE* nor *MEDLINE* is fully comprehensive, and many scientists use both. See Ch. 16 for more on the history of *Index Medicus* and *Excerpta Medica*.

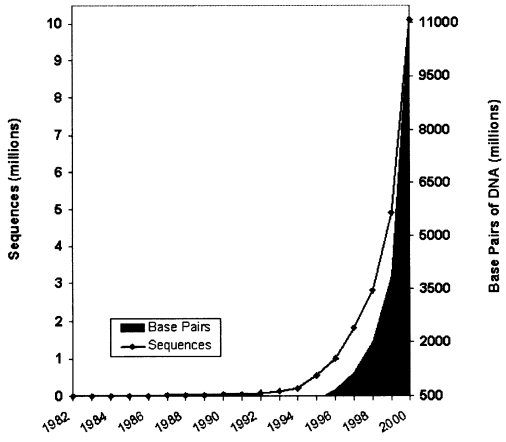

Figure 1: Growth of *GenBank*.

In the 1960's, Eugene Garfield, at the Institute for Scientific Information (ISI), created the *Science Citation Index*, which added an important twist to the idea of a bibliographic database, by including details of all the citations from the reference list of each indexed article. ISI and the *Science Citation Index* are described in more detail in Ch. 15.

By the 1980's, companies such as Ovid, SilverPlatter and DIALOG were offering paid-for access to *MEDLINE* and *EMBASE*, via their own proprietary software, typically on CD-ROM, or via a text-based online interface.

In 1988, as it became clear that molecular biology was producing an explosion of data that would require processing with advanced computational tools, the National Council for Biotechnology Information (NCBI) was founded as part of the NLM. One of the most important roles played by the NCBI is managing a collection of globally accessible databases of biological sequences and structures — of which the *GenBank* nucleic acid sequence databank is perhaps the most significant. In doing so the NCBI works closely with similar organisations in Europe and Japan who maintain their own databanks. The NCBI recently also became responsible for

the data collected by the US human genome project.

One of the most significant developments at the NCBI in recent years has been the Entrez system for retrieving sequences and related information [3,4]. Entrez allows sequences, structures and related bibliographic records from *MEDLINE* to be retrieved, either by keyword searching, on the basis of a similarity based clustering scheme, or by explicit links between the various databases (e.g., from a protein sequence to a corresponding structure).

Entrez was initially released as a quarterly CD-ROM in 1993, but the following year a networked version was released, which meant that updates could be far more frequent. As web browsers took off in 1995, the web became the dominant mode of access to Entrez. Initially, only a small molecular biology related subset of *MEDLINE* was included in the Entrez database, but the popularity of even this limited free web-based *MEDLINE* searching was such that in 1997, the US government decided to make the whole of *MEDLINE* searchable on the web without charge. This aspect of Entrez, known as *PubMed*, was an immediate success. By the end of 1999 *PubMed* was handling 700,000 searches per day. It was clear that the power of the web to provide open electronic access to research information would profoundly change the way scientists communicated.

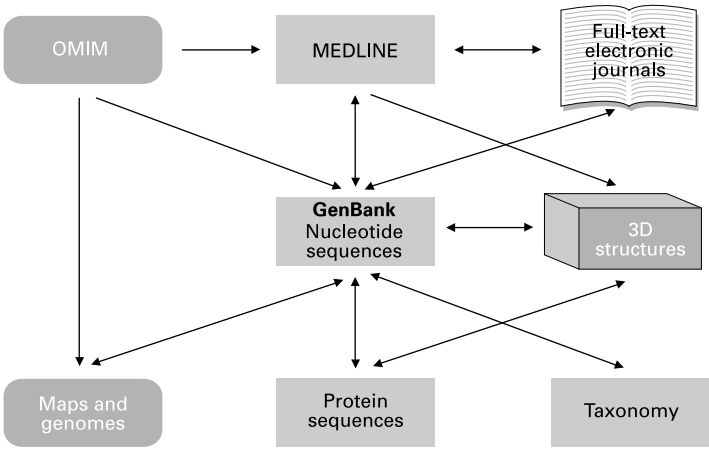

Figure 2: Links between Entrez databases.

The response of science publishers to the web

Traditional scientific publishers had been pondering the coming importance of electronic access to scientific research for some time. However, the speed of the web revolution took everyone by surprise. Publishers began to place their scientific jour-

nals on the web in large numbers, and disparate as these efforts were, there usefulness quickly outstripped what was being achieved by expensive proprietary digital library experiments such as Red Sage (http://www.ckm.ucsf.edu/projects/RedSage) [5].

By the mid 1990's, enlightened publishers had already begun to digitize the content that they published in a structured form such as SGML, and were able to take full advantage of the evolving capabilities of the web. The establishment of Adobe's Portable Document Format (PDF) as a standard also played an important role in encouraging online journal access, since PDF's are easy for the publisher to produce, quick to download, and when printed via a laser printer, produce results far superior to a traditional photocopy or fax.

The larger scientific publishers devoted significant resources towards building their own comprehensive electronic journal solutions. Examples of these services include Academic Press's "IDEAL" (http://www.idealibrary.com), Springer's "Link" (http://www.link.springer.de) and Elsevier's "Science Direct" (http://www.sciencedirect.com). These services were typically targeted at existing print institutional subscribers, who by paying a small supplement on top of their existing subscriptions could get online access to their holdings. In some cases, as a hedge against online access causing immediate drop-off of print subscriptions, publishers encouraged libraries to enter into online access agreements which committed the library to retain all its existing print subscriptions for a 3–5 year period, in return for broad online access. Thus, online access became an important tool which these publishers could use to address their ongoing problem of attrition (losing old subscribers faster than new ones can be acquired).

Some smaller publishers (e.g. *Nature*, http://www.nature.com) also built their own sites, either in-house or via outsourcing. But many publishers did not have the resources to build a full-featured website from scratch. HighWire Press (http://www.highwire.org), a non-profit offshoot of Stanford University Libraries, filled this gap by developing systems to host online journals in a standard way. The first journal offered online through HighWire, in early 1995, was the *Journal of Biological Chemistry* (*JBC*). HighWire set a high standard with its online journals, and many society journals and others followed *JBC*'s lead. HighWire currently hosts 225 different journal sites. Most HighWire sites restrict access to subscribers only, but with HighWire's encouragement, an increasing number of these journal sites make their content freely available to non-subscribers after an 'embargo period', typically ranging from six months to two years, has elapsed. HighWire Press now (December 2000) boasts that its sites offer a total of nearly 200,000 full-text articles for download without charge. (For Highwire's list of open-access research archives, see: http://www.highwire.org/lists/largest.dtl)

Bibliographic linking

From a scientist's point of view, one of the major problems with the explosion of different journal websites has been the lack of reliable citation linking. In the early days, publishers would link to articles on their own websites, but would not link to other publishers articles, either for technical or political reasons. Some publishers even went so far as to block other sites from linking to their articles. Eventually however, the message got through to publishers that readers wanted to be able to follow any citation they came across and find the full-text article concerned, and the CrossRef initiative was born (http://www.crossref.org).

CrossRef, which exploits the Digital Object Identifier (DOI) standard, is designed to be a generic system for resolving citation links. Most major scientific publishers are participating in the CrossRef initiative, but has yet to be widely implemented. In the meantime, the increasing numbers of full-text links from databases such as the *Science Citation Index* and *PubMed* go some way to filling the gap.

Online communities

Not all scientific publishers used the web simply to make their existing published content available online. In the early 1990's, Current Science Group, then publisher of the Current Opinion series of review journals, developed *BioMedNet* (http://www.bmn.com), an internet-based community service for biologists and medical researchers. Initially, access to *BioMedNet* required the use of dedicated 'client' software, but as browsers such as Netscape became available, the service was quickly switched over to the web. *BioMedNet* not only offers the full text of the Current Opinion journals, but also brings together facilities such as a job exchange, discussion forums, news, a bookshop, databases and a scientific webzine, *HMS Beagle* (http://www.hmsbeagle.com), to which many scientists contribute. One of *BioMedNet's* most popular innovations is its enhanced *MEDLINE* service, which uses evaluations from Current Opinion reviewers to highlight the most interesting articles in *MEDLINE*. Access to most of these services is free, but requires registration, although access to review articles requires a subscription. By mid-1999 *BioMedNet* had more than half a million registrants.

The success of *BioMedNet* was repeated by *ChemWeb*, a joint venture between Current Science Group and MDL Information Systems Inc. (http://www.chemweb.com). *ChemWeb* offered registrants access to chemical journals and databases, along with community facilities similar to *BioMedNet's*. *ChemWeb's* unique feature, when it launched, was the use of MDL technology to offer structure-based searching of many of its databases. This allowed chemists to draw a specific chemical structure (using a browser plug-in), and then search for references to struc-

turally similar molecules in any of *ChemWeb's* databases.

In subsequent years, many more scientific community sites (sometimes known as vertical portals or vortals) have followed in the footsteps of *BioMedNet* and *ChemWeb*. In biomedicine these include *Medscape* (http://www.medscape.com) and the *Community of Science* (http://www.cos.com), while in chemistry, the American Chemical Society launched *ChemCenter* (http://www.chemcenter.org) and the Royal Society of Chemistry, *ChemSoc* (http://www.chemsoc.org).

Many community sites have been started by existing scientific publishers, but in other cases they have been started by new companies. For example, VerticalNet (http://www.verticalnet.com), founded in 1995, operates a variety of industry-specific sites, including *Bioresearch Online*. VerticalNet's sites provide various kinds of community information and services, but their prime function is to act as a front-end for e-commerce marketplaces. Internet-based scientific e-commerce has proven to be a difficult area however, as witnessed by the closure in late 2000 of the Chemdex online life science marketplace, which less than 12 months previously had had a market capitalisation of more than $10 billion.

Databases

Publishers of commercial scientific literature databases were also quick to adopt the web. For example, the Institute for Scientific Information developed a web based front end for its citation databases, *Web of Science* (http://www.isinet.com/isi/products/citation/wos), and set up linking agreements with several journal websites.

Other bibliographic database providers followed suit, but free access to *PubMed* has changed the competitive landscape significantly.

Aside from bibliographic databases, the web has also allowed scientists to easily and conveniently self-publish databases which collate biological information of various kinds in specific niche areas. *Nucleic Acids Research* (http://www.nar.oupjournals.org) publishes an annual database issue [6], which catalogs some of these databases. A problem that frequently occurs, however, is that the curation of the databases becomes an unmanageable long-term burden on the lab or individual that set them up.

SWISS-PROT (http://www.expasy.ch/sprot/sprot-top.html) offers one model of a solution to this problem [7]. *SWISS-PROT* is a curated, non-redundant protein sequence database containing annotations that describe evidence of protein function (both experimentally and theoretically determined).

Begun in 1986, initially maintained by the laboratory of Amos Bairoch at the University of Geneva, and later in collaboration with the European Bioinformatics

Institute, *SWISS-PROT* grew to be a widely used resource, but by 1996, it was in funding crisis. The solution reached was to form a separate non-profit body, the Swiss Institute of Bioinformatics (SIB) to be responsible for maintaining *SWISS-PROT*. The SIB (http://www.isb-sib.ch) receives some funds from the Swiss Government, but supplements these with income obtained by licensing *SWISS-PROT* to the commercial sector. SWISS-PROT remains freely accessible to academics. It is likely that this model will be emulated by other high value but high maintenance databases in the future [8]. Alternatively, many existing databases may disappear or cease to be maintained. Many commercial alternatives are already appearing, from the growing number of companies such as Incyte, Celera, DoubleTwist.com and Rosetta Inpharmatics which specialize in such bioinformatics databases and tools.

Markup languages and file formats

The explosion in the use of internet and software tools to analyse biological information has led to an urgent need for standard file formats for the exchange of this data.

Many different ad hoc file formats, mostly text-based, have become widely used in molecular biology. Often these file formats are named after the software or database which make use of them (e.g. FASTA format, PDB format, SWISS-PROT format).

As discussed in Ch. 17, at the same time as these biological data formats were coming into use, important work was also going on in the development of standard markup languages, to allow data to be structured in a flexible way, while facilitating its exchange and its conversion to other formats.

In an attempt to bring some standardization to database formats, NCBI initially experimented with the use of an ISO markup standard known as Abstract Syntax Notation 1 (ASN.1). Recently, though, ASN.1 has been overshadowed by the emergence of XML as the predominant standard markup language. Many biological databases including *GenBank* now allow data to be downloaded in an XML format of some kind. XML is not really a file format, however. It is a meta-file format — a standard way of describing file formats. The full benefits of XML cannot be realised until domain-specific XML formats (known as Document Type Definitions, or more recently, Schemas) are agreed and used throughout the scientific community. Two of the most well-developed scientifically relevant XML formats include Chemical Markup Language (CML) (http://www.xml-cml.org), [9] and Mathematics Markup Language (MathML) (http://www.w3.org/TR/REC-MathML). After several years of experimentation, CML and MathML are finally on the verge of mainstream use. XML markup standards for biological data are at a much earlier

phase in their development. Initiatives such as the Gene Expression Markup Language (GEML) (http://www.geml.org) [10], for describing microarray data, are an important starting point, but it may be some time before any such standard gains widespread acceptance.

Pre-prints and distributed archives

Biologists got their first taste of broad open access to research information through the web with the launch of *PubMed* in 1997. But *PubMed* includes only abstracts, not full text articles. Many physicists, on the other hand, had been accessing a large collection of full text research articles through the web at no charge for several years. The Los Alamos Physics Preprint Archive (now known as arXiv.org; http://www.arxiv.org) began in 1991 first as an email service, and subsequently as a widely-mirrored web archive, which allows researchers to exchange 'preprints' — articles that have not yet been accepted into a peer-reviewed journal. Initially the archive covered only high-energy physics, but its scope has expanded until it now covers all areas of physics, and also some areas of mathematics and computer science.

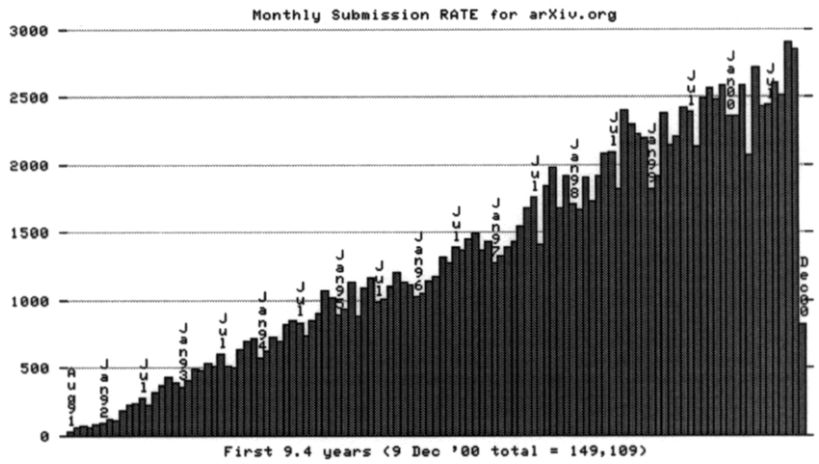

Figure 3: Growth of arxiv.org pre-print repository.

Within the physics community, there was already a long tradition of preprint circulation, in paper form, and as a natural electronic extension of this system, arXiv.org has been widely accepted by both physicists and physics publishers. Many of the articles made available through the arXiv.org servers do go on to be published in peer-reviewed journals, but in particular sub-fields of physics, arXiv.org is

now the primary mode of access to the research literature.

In the chemical and biomedical sciences, no such established tradition of broadly circulated preprints existed, and although non-physicists have looked enviously at the arXiv.org example, it is not clear to what extent the same model can succeed in other sciences.

Many worry that in medicine especially, relying on an archive of research which has not been subject to peer-review could have dangerous consequences. Also, many scientists are nervous about submitting their research to pre-print servers, worrying that their work will subsequently not be accepted for publication in traditional journals, many of whose rules prevent authors from submitting work that has previously been made available elsewhere.

Nonetheless, several initiatives have started which aim to allow researchers in areas other than physics. These include the *British Medical Journal's* Netprints (http://www.clinmed.netprints.org/home.dtl) and *ChemWeb's* Chemistry Preprint Server (http://www.preprint.chemweb.com), and CogPrints (http://www.cogprints.soton.ac.uk), a preprint archive for Cognitive Science operated at the University of Southampton.

One recent development arising from the interest in pre-print servers is the Open Archives Initiative, an emerging set of XML standards for the interchange of metadata (such as titles, abstracts, and subjects/keywords) between research archives in different physical locations (http://www.openarchives.org). Originally envisioned as a way of connecting pre-print archives in the biomedical sciences, the initiative has expanded into a generic framework for exchange between distributed archives of scholarly literature of any kind. For example, one participant in Open Archives is the Networked Digital Library of Theses and Dissertations (http://www.theses.org/), which aims to bring together archives of digital theses and dissertations from universities around the world.

Another area in which the collection of metadata from many sources is becoming important is clinical trials. Publication of the results of clinical trials in conventional journals is problematic, since clinical trials producing inconclusive or negative results are less likely to be published. This can significantly skew the balance of results that appear in the published literature.

ClinicalTrials.gov is the US National Institutes of Health's response to this problem — a comprehensive archive of all in-progress NIH-sponsored clinical trials. Taking the same idea further, the recently released metaRegister of Controlled Trials (http://www.controlled-trials.com), published by Current Science Group in collaboration with the UK Cochrane Center, Glaxo-Wellcome and others, brings together information from many registers into a single web-searchable database.

When combined with the online publication of results from all clinical trials, however inconclusive, this approach promises to eliminate the problem of 'publication bias'.

Open access to research — PubMed Central and BioMed Central

The huge success of NIH's decision to make *MEDLINE* freely available, via *PubMed*, led to the recognition that open access to biomedical research was highly desirable from a scientific point of view.

Since it started, *PubMed* has continued to increase the number of links from *PubMed* records to fulltext articles (as of January 2001, *PubMed* includes links to full text articles from more than 1600 journals). But for many scientists these full text links lead to frustration, as the articles concerned are not accessible without a personal or institutional subscription. Journal subscription prices have greatly outpaced inflation for many years, and so even relatively well-off institutions cannot afford to subscribe to all the publications they would like. Currently, a major funding organisation such as the National Institutes of Health spends tens of billions of dollars each year on biomedical research, but then has to pay once again to get access to the resulting research articles for its scientists.

Against this background, in August 1999, after a period of consultation with the research community, Harold Varmus, then head of the NIH, announced the PubMed Central initiative (http://www.pubmedcentral.nih.gov). PubMed Central's mission [11] was defined as the creation of a permanent archive of peer-reviewed biomedical research which would be available to all, without subscription charges or other barriers to access. PubMed Central is not itself a publisher, and does not control the peer-review process, although it does set minimum standards for what constitutes peer review and therefore what can and cannot be included in PubMed Central.

Publishers were encouraged to allow existing journals to be archived in PubMed Central, but recognizing that many would be reluctant to do so because of the impact it might have on their subscription revenue, Varmus encouraged the scientific community to set up new open-access journals specifically intended to be archived in PubMed Central. Several aspects of the PubMed Central proposal were designed to speed the acceptance of new online-only journals. Firstly, all research archived in PubMed Central is listed in *PubMed*, and is highly visible to scientists since *PubMed* is the single most widely used biomedical bibliographic database in the world. Secondly, by providing an independent NIH-backed electronic archive, PubMed Central provides a credible guarantee of permanent accessibility for those electronically published articles, which a new small publisher could not provide

alone. Finally, by openly supporting the development of new online only journals, the NIH provided a reasonable indication that electronic only publication, in new journals, would be treated fairly when making funding and career decisions on the basis of a publication track record — i.e., it reassured scientists that publishing in an established journal was not the only option if they wanted to obtain kudos and career advancement.

Several existing journals already participate in PubMed Central, including *Proceedings of the National Academy of Sciences* and *Molecular Biology of the Cell*, but these journals operate an embargo period, and so articles appear on PubMed Central only after a several month delay, during which time they are available to subscribers only. The *British Medical Journal* (*BMJ*), funded largely from membership dues paid to the British Medical Association and therefore not wholly reliant on subscription revenue, makes its content available through PubMed Central without delay.

Another publisher which has embraced PubMed Central wholeheartedly is Current Science Group. With the announcement of PubMed Central, Current Science Group saw the opportunity to create an alternative to traditional research journals and in late 1999 set up BioMed Central, a website which allows scientists and clinicians to publish research articles in any area of biology or medicine (http://www.biomedcentral.com). In total, BioMed Central offers a choice of 60 subject-specific online journals, each of which has a panel of expert subject advisers.

BioMed Central also works with groups of scientists to create electronic 'niche journals'. The editorial process for these journals will be controlled by the group of scientists concerned, who will make use of the online manuscript submission and peer-review tools that have been developed for the main BioMed Central journals.

BioMed Central is a commercial initiative — it plans to reduce the cost of publishing original research to a minimum through the use of the web and technologies such as XML which facilitate the automation of the publication process. It then plans to recoup the remaining cost through advertising, e-commerce linkups, and by offering value-added services which scientists are prepared to pay for, such as high quality databases and commissioned review articles. In doing so, BioMed Central aims to develop a new model for commercial scientific publishing, which incorporates open access to original research as a basic tenet.

The slowness of commercial publishers to allow open-access to newly published research has become a significant frustration for many scientists, who believe that the potential of the web to facilitate scientific communication is being squandered. Several thousand scientists have gone as far as to sign an open letter, pledg-

ing to boycott any journal that fails to provide open electronic access to the research it publishes within 6 months of publication. [12,13].

Building a permanent digital archive

As scientists increasingly rely on electronic means to view journal articles, librarians are in many cases considering cancelling their print subscriptions, and are being encouraged to do so by publishers, for whom printing and distribution is now an unnecessary expense.

However, this has prompted concerns amongst some librarians as to the longevity of the digital record. Past experience suggests that paper journals, stored carefully, will remain accessible on a timescale of centuries or even millennia. But keeping a similarly permanent digital record is not straightforward. Typical digital media such as magnetic and optical disks have a physical lifespan of just years, or at most decades. Furthermore, the pace of technological advance means that even if the digital medium remains intact, the equipment for reading it may have long ago become obsolete and unavailable (5 $^1/_4$" floppy disk-drives are already something of a rarity).

Concerns such as these have prompted a variety of proposals, ranging from Stanford Universities LOCKSS (Lots of Copies Keep Stuff Safe) project (http://www.lockss.stanford.edu) which allows libraries to maintain their own copies of important web content, all the way through to more outlandish suggestions such as periodically micro-engraving important data in analog form onto nickel disks, as proposed by the Long Now Foundation (http://www.longnow.org).

As reliance on online journals increases inexorably, this issue will certainly have to be addressed in years to come.

References

[1] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J. & Ostell, J. (2000) Genbank. *Nucleic Acids Research*, **28**, 15–18.

[2] Stoesser, G., Baker, W., van den Broek, A.E., Camon, E., Hingamp, P., Sterk, P. & Tuli, M.A. (2000) The EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, **28**, 19–23.

[3] Woodsmall, R.M. & Benson, D.A. (1993) Entrez: Sequences and Entrez: References; NCBI's Genbank on CD-ROM. *Biotech Knowledge Sources*, **6**, 3–4.

[4] Baxevanis, A.D. & Landsman, D. (1995) The Internet biologist: Network Entrez. *FASEB J*, **9**, 994.

[5] An experimental digital journal library for the health sciences. *D-Lib Magazine* (1995, August) http://www.dlib.org/dlib/august95/lucier/08lucier.html.

[6] Baxevanis, A.D. (2000) The Molecular Biology Database Collection: an online compilation of relevant database resources. *Nucleic Acids Research*, **28**, 1–7.

[7] Bairoch, A. & Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, **28**, 45–48.

[8] Bairoch, A. Should the model proposed by SWISS-PROT be considered by other databases? http://www.expasy.ch/announce/abpp98_1.html.

[9] Murray-Rust, H.S. & Rzepa, P. (1999) Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. *J Chem Inf Comput Sci*, **39**, 928–942.

[10] A useful list of resources relating to biological applications of XML. http://www.maggie.cbr.nrc.ca/~gordonp/xml/.

[11] Varmus, H. PubMed Central: An NIH-operated site for electronic distribution of life sciences research reports. http://www.nih.gov/about/director/pubmedcentral/pubmedcentral.htm.

[12] Brown, P. (2000) What must scientists do to exploit the new environment. *Freedom of Information Conference -The impact of open access on biomedical research July 6th-7th, 2000, New York, Academy of Medicine.* http://www.biomedcentral.com/info/brown-tr.asp.

[13] Public Library of Science. http://www.publiclibraryofscience.org.