Donald A.B. Lindberg

National Library of Medicine,
Bethesda MD, USA

Preface

Seeking the Digital Library

Congratulations are once more due to the editors and contributors of this Yearbook. After ten years of outstanding achievements, the baton has been passed from Jan van Bemmel and Alexa McCray to Reinhold Haux and Casimir Kulikowski. These have been extremely useful volumes, drawing from the best work in the world to bring together informative and insightful assemblies on the selected special topics. This year's topic, "Digital Library and Medical Informatics" is indeed a fascinating one.

In the U.S., the Digital Library was a program initiated by the White House High Performance and Communications Program.[1] Initially, six university libraries were supported jointly by the National Science Foundation, the Defense Advanced Research Projects Agency, and the National Aeronautics and Space Administration. Subsequently, the National Library of Medicine joined the consortium, and then biomedical applications were also supported. [2]

While many readers will be familiar with the term "digital library", some will associate the idea with a traditional physical library that has enthusiastically deployed computers to assist its staff and its users, and others will imagine a conceptual store of knowledge in an important special field that has been designed for remote computer access. Of course, both ideas are valid, but both fall quite short of the rather more global and more experimental nature of the challenge conveyed in the Yearbook for 2001. There are many good libraries, my own institution included, in which virtually every step from literature selection to information retrieval is intimately bound up with computers, networks, and information technology. Yet it is only in certain of the newer library functions that virtually all of the processes from data creation to interpretation of the experimental results are truly digital, and the whole process is without any preceding non–digital form. The storage and retrieval of gene expression data is an example of such a process. [3]

In this exemplary application, "gene chips" of one sort or another are used to determine which particular genes in experimental or even human surgical specimens are actually coding for protein production, and which are not. Tens of thousands of individual genes can be looked for simultaneously by this technique. In the case of the NLM, the storage of such results has been anticipated by David Lipman and colleagues in the National Center for Biotechnology Information by provision of GEO, an information storage and

retrieval system for this purpose. [4,5] GEO is intended to store the basic results along with data fields that permit the contributor or user to identify the experimental "run," the result, and emerging standards data such as the chip name and the physical sequence of genes in the array. Outside this library system, it will be the responsibility of the contributor to store (doubtless also in computer form) the information that surrounds the rationale of the experiment, the assumptions and hypotheses, in short the conjectural background that gives scientific meaning to the raw experimental results stored in the digital library.

This system will provide for sharing of these new data between investigators. In this respect the concept resembles GenBank (and Entrez, its information access system). GenBank data arise from large–scale sequencing centers funded by NIH (and other American, European, Asian, and national research funding sources, and in many cases from laboratories funded by industrial corporations), from individual investigators funded by NIH and other sources, and from an amazing variety of individual scientists from virtually all parts of the world.[6] All make their data publicly available. Indeed an investigator with a potentially "new" sequence (whether nucleotide or protein) cannot determine if it really is "new", nor interpret its genetic function without reference to the discoveries that have preceded. These are contained in the great corpus of knowledge stored in GenBank (and its European and Asian counterparts). Thus, sharing of scientific sequence data in a public data base library is sensible and productive–whether mandatory or not.

In the case of the new gene expression data, however, one does not know if sharing will occur, nor whether sufficient standards and metadata descriptors will be created to permit sharing of gene expression results, even if scientists wish to do so. Thus, part of the mission of the digital library must be to experiment by attempting to develop entirely new information methods and services. On the other hand, we must be prepared for some to fail, or fail to be used, because of social not technical considerations that may be outside the library's control.

For the National Library of Medicine, the emerging digital library is at once an intriguing daily challenge and also an evanescent image receding into the future. This is because all of the dimensions of our work are changing at the same time. These are, first, the basic information we acquire, organize, store, and disseminate. Second, we see the emergence of the Internet (and its next generation relatives and evolving network software) as the new means to share this information. Third, we welcome wholly new audiences for our library services: namely, patients, families, and the public. Since NLM made MEDLINE searching a service free to all [7,8], the total volume of searching has increased twenty times, and the use of this bibliographic file of biomedical literature by the public has reached 34% of the total. Subsequently we have created additional information services designed primarily for consumers. MEDLINEplus[9] and ClinicalTrials.gov[10] are two such systems. Both are available to and used by both the public and by health professionals. Consequently both exemplify our efforts to present interfaces and linguistic exchanges with users who span a great range of education and familiarity with the knowledge conserved by the digital library.

Modern biomedical science must, as a field, retain its commitment to preserve as well as to collect the voluminous new digital data. This problem is a good bit more demanding and complex than it seems at first. Students of the problem have taken a number of approaches, including writing to hopefully "permanent" media, imagining and designing systems to rewrite periodically the important data to archival but not perfectly permanent media, and finally, an approach currently being investigated by NLM. This is an experiment in which we explore if "permanent access" to digital information (regardless of whatever technological changes are needed to provide the regular access), is the best way in the end to assure "permanence" of the digital data itself. To test this empirical approach, we have challenged ourselves to declare explicitly with respect to our own NLM public files those which will be guaranteed permanently available. Immediately one sees that different levels of availability are required and that some problems exist merely in creating such taxonomies. For example, MEDLINE files will surely be available permanently but not in a completely unaltered state. In fact, errors are corrected daily. At the moment, we think that four categories of permanent access will suffice to describe this policy: first, Permanent, Unchanging Content (example: a scanned image of correspondence in the History of Medicine collection); second, Permanent, Stable Content (example: a MEDLINE record); third, Permanent: Dynamic Content (example: NLM's Home Page); and fourth, Permanence Not Guaranteed (example: a training schedule, or opening hours). Naturally, more testing is required to determine if these are indeed sufficient for NLM–and whether a similar approach might be workable for other institutions.[11] Our goal is to gain experience from carefully defined experiments so that we can contribute to the development of workable national standards and strategies.[12]

Many important public policy questions accompany the interesting technical and scientific problems.[13] Generally these are matters that must be dealt with by laws or regulations that are specific to a nation. We can all benefit, however, by study of those such policies that prove workable elsewhere. Clearly in the technical domain of standards, none are of much matter nowadays if they are not acknowledged internationally.

The evolving field of digital library studies may well go beyond the few examples I have mentioned. Indeed it may exceed even the richer sample of research pictured in this volume. Some have suggested an ultimate fusion of the digital library work and the computer–based patient record, yielding a system that finally provides ready knowledge on which to base medical decision–making at the time and place the patient and the caregiver need it. If so, the contributors to the present Year Book will have earned the right to be doubly proud.

References

1. National Science and Technology Council (US), Committee on Information and Communications. High performance computing and communications: foundation for America's information future: a report. Supplement to the President's FY 1996 budget. Washington, DC: The Council; 1995.
2. McCray AT. Digital library research and application. Stud Health Technol Inform 2000;76:51–62.
3. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nat Genet 1999 Jan;21(1 Suppl):33–7.
4. Marshall E. Do–it–yourself gene watching. Science 1999 Oct 15;286(5439):444–7.
5. Gene Expression Omnibus: GEO [Database on the Internet]. Bethesda (MD): National Library of Medicine, National Center for Biotechnology Information; 1999 Oct [Updated 2000 Aug 1; cited 2000 Oct 19]. Available from: http://www.ncbi.nlm.nih.gov/geo
6. Benson DA, Boguski MS, Lipman DJ, et al. GenBank. Nucleic Acids Res 1999 Jan 1;27(1):12–7.
7. Tenopir C. MEDLINE on the Web: databases for free. Library J 1997 Oct 1;122(16):37–8.
8. Landro L. Health web sites get better at explaining complex medical data. Wall Street Journal 2000 Jul 14; Sect. B:1 (col. 1).
9. Miller N, Lacroix EM, Backus JE. MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service. Bull Med Libr Assoc 2000 Jan;88(1):11–7.
10. McCray AT, Ide NC. Design and implementation of a national clinical trials registry. J Am Med Inform Assoc 2000 May–Jun;7(3):313–23.
11. Byrnes MM. Defining NLM's commitment to the permanence of electronic data. ARL 2000 Oct;212:8–9.
12. Humphreys BL. Electronic health record meets digital library: A new environment for achieving an old goal. J Am Med Inform Assoc 2000 Sep–Oct;7(5):444–52.
13. Lindberg DA, Humphreys BL. Medicine and health on the Internet: the good, the bad, and the ugly. JAMA 1998 Oct 21;280(15):1303–4.

Address of the author:
Donald A.B. Lindberg, M.D., Director
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA