

DEPARTMENT OF HEALTH AND HUMAN SERVICES

NATIONAL INSTITUTES OF HEALTH

PUBMED CENTRAL NATIONAL ADVISORY COMMITTEE

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION

NATIONAL LIBRARY OF MEDICINE

December 2, 2003

**NLM Board Room
National Center for Biotechnology Information
National Library of Medicine
8600 Rockville Pike
Bethesda, Maryland 20894**

**DEPARTMENT OF HEALTH AND HUMAN SERVICES
NATIONAL INSTITUTES OF HEALTH
NATIONAL LIBRARY OF MEDICINE
NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION
PUBMED CENTRAL NATIONAL ADVISORY COMMITTEE**

Function of the PubMed Central National Advisory Committee

Since the mission of NIH is to conduct and support medical research and to disseminate the results of that research widely to the public and the scientific community, it will make use of electronic publishing technology to fulfill this role by establishing and maintaining PubMed Central. This new service is a Web-based repository, housed at the NCBI that will archive, organize, and distribute peer-reviewed reports from journals in the life sciences, as well as reports that have been screened but not formally peer reviewed. The Committee shall advise the Director, NIH, the Director, NLM, and the Director, NCBI, concerning the content and operation of the PubMed Central repository. Specifically, it is charged to establish criteria to certify groups submitting materials to the system, monitoring the operation of the system, and ensuring that PubMed Central evolves and remains responsive to the needs of researchers, publishers, librarians and the general public.

SUMMARY MINUTES OF MEETING – DECEMBER 2, 2003

The meeting of the PubMed Central National Advisory Committee was convened on December 2, 2003 in the Board Room of the National Library of Medicine (NLM), Bethesda, Maryland. The meeting was open to the public from 9:30 a.m. to 2:00 p.m. Dr. Joshua Lederberg presided as Chair.

Members Present

Joshua Lederberg, Ph.D., The Rockefeller University, PubMed Central National Advisory Committee Chairman
Michael Eisen, Ph.D., University of California at Berkeley
Anthony Delamothe, M.D., BMJ Publishing Group
Richard Johnson, The Scholarly Publishing & Academic Resources Coalition (SPARC)
Samuel Kaplan, Ph.D., University of Texas Medical School at Houston
Paula Kaufman, M.B.A., University of Illinois at Urbana-Champaign
Chaitan Khosla, Ph.D., Stanford University
Ajit Varki, M.D., University of California, San Diego
Linda A. Watson, M.L.S., University of Virginia
James Williams, M.S., University of Colorado at Boulder
David J. Lipman, M.D., Director, National Center for Biotechnology Information, NLM, NIH, PubMed Central National Advisory Committee Executive Secretary

Ad-hoc Members Present

Mariam Balaban, Desalination
Debra Lappin, J.D., Princeton Partners Ltd.

NLM Senior Staff Present

Donald A.B. Lindberg, NLM Director

Donald King, M.D., Deputy Director for Research and Education, NLM

Kent Smith, Deputy Director, NLM

Betsy Humphreys, Associate Director for Library Operations, NLM

Visitors Present

Bridget Coughlin, Proceedings of the National Academy of Science of the United States of America

Dr. Michael Rogawski, NINDS, NIH

I. Call to Order and Opening Remarks

The meeting was called to order at 9:45 a.m. Dr. Lipman welcomed members of the PubMed Central National Advisory Committee. The Committee officially adopted the minutes from the previous meeting. The date of May 18, 2004 was confirmed for the next meeting. Proposed dates for the following meeting will be sent to members for consideration. Committee members and guests were then introduced.

Dr. Lindberg, NLM Director addressed the group and began by thanking members for their time and commitment to the committee. He reviewed that the committee's function is to set policy for PMC and provide guidance on how PMC should operate. He also welcomed the launch of PLoS and its participation in PMC. Dr. Lindberg then raised two issues for consideration by the committee: 1) possible location for repository for large-scale data results of studies reported in literature and 2) education outreach and training of bioinformatics specialists.

Along the same lines, Dr. Lipman briefed the group on a meeting held by NHGRI based on the need for a new human-based integrated database. Two of three breakout groups at this meeting concluded that before a new database is implemented, consumers need to be made more aware of the resources currently available. Therefore, they suggested better training before a new database is built. The group suggested that NIH leadership needs to be sensitized to this issue and make an investment in training the medical and user community.

II. PubMed Central Update

Journal Status

Dr. Lipman began the PMC update with listing new journals. *PLoS Biology* started publishing in August and is available in PMC. Other new journals include the *Journal of Athletic Training*, *Canadian Veterinary Medicine Association Journals (2)*, and *Amphibian and Reptile Conservation*. *Annals of Internal Medicine* agreed to participate in PMC, but then reversed their decision due to copyright concerns. The *New England Journal of Medicine* was considering participation but at this point is not expected to participate.

A discussion followed about the reasons journals state for deciding against participation. There are often concerns about ownership and copyright issues. Some journals have requested the option to stop making back issues freely available at some point, which is not something that PMC allows as it would not adhere to the definition of a permanent archive. Some journals also do not want to lose the potential of making additional revenue in the future from past

publications. Dr. Lipman noted that journals with an advocate for PMC and open access within their leadership are more likely to follow through with participation.

Open Access

Dr. Lipman informed the group of a new development in which *Science* has deposited an article in PMC authored by Harold Varmus and others. This marks the beginning of a hybrid policy for depositing individual articles in PMC where negotiations occur with authors, even though the journal in which it appears does not participate. Guidelines at this time for the open access policy include:

- The journal of publication must meet PMC qualification standards.
- The full text of the article will be accepted in NLM Journal Publishing DTD or the Keton DTD only, on or before the date of publication.
- No journal-specific pages will appear for these articles, a general PMC Open Access banner will be used at the top of the page. No links will be provided from the article page in PMC to the publishing journal site.
- PMC Open Access Initiative (OAI) service provides XML for full text of the OA article. All other files (XML, PDF, images) will be available via FTP.

Data Conversion Labs have agreed to a piece-work service to mark up OA article content in the NLM DTD for approximately \$5 per page, with a minimum charge of \$50 per article. BioMed Central is also offering a similar service and other companies are being researched.

In response to a member question, it was stated that older articles will be accepted as long as the policy criteria are met and PMC has publisher assurance that open access is allowable for the article. Dr. Lipman added that PMC does not want to have publisher restrictions on these articles. It is believed that the OA service gives journals and authors flexibility in their desire to offer open access. Dr. Lipman also added that Oxford University Press would like to have a hybrid model for author choice of open access for a fee.

A question was raised regarding the possibility of PMC accepting a third party funded publication that is not copyrighted, such as a government publication. Dr. Lipman deferred to the committee for a consensus and suggested that it may be an issue for NLM legal counsel.

Dr. Rogawski, an NIH scientist in attendance, shared his personal experience with this topic and stated that the issue revolves around the embodiment of the text, i.e. how it appears in the journal. If only the text, in Word format for example, is provided via open access, there should be no problem on the part of the publishing journal and authors. He personally favored the idea of all of his work being as widely distributed as possible, especially review articles published in books.

Dr. Lindberg suggested that there be further investigation into releasing information authored by government employees. The committee agreed that NLM will investigate this issue with the NIH and provide feedback on its position to the committee.

Another question was raised regarding open access negotiation methods and the presence of a mechanism that assists in negotiation of archiving information. Rick Johnson and Debbie Lappin mentioned that SPARC has a new working group looking into formalized support for open access. At the next SPARC meeting, the subcommittee will be discussing guidelines for working

with publishers and academic centers on OA issues. Topics for discussion will include a model agreement, principles for an author agreement, and an addendum for authors to retain certain rights. This will assist authors with guidelines for working within the new open access initiative. Rick will update the PMC committee on the SPARC group's progress at the next PMC meeting. It was clarified that a draft agreement will be circulated to key players in the field to be sure that it addresses recognized needs. Endorsement will then be sought from societies, journals, and other institutions.

Sam Kaplan mentioned that OA cost for authors and copyright agreements will be discussed at the next ASM meeting as well. Ed Sequeira stated that he crafted an agreement with *Science* for the first OA article containing some Bethesda principles on open access. Open Access information is available in the 'About PMC' section of the PMC web site.

The cost of publishing for authors was also discussed in terms of traditional versus open access publishing. A submission charge for authors has been proposed for some publications but it has been difficult to come to a fixed charge due to various circumstances. Some publishers have suggested a menu of charges for different options. Dr. Lederberg suggested a discussion regarding scientific association attitudes toward open access and their rights.

Back Issue Scanning

Dr. Lipman provided the committee with an update on the NLM back issue scanning project. As of November 17, 1.1 million pages have been scanned with 1.6 million more pages in the queue. The scan cost per page is currently averaging \$.60 and current funding should accommodate approximately 4 million pages.

The *Bulletin of the Medical Library Association*, dating back to 1911, has been released. Full text searching is available across the collection and viewing options include PDF format, individual scanned pages, enlarged images and OCR text. A demonstration of searching and images were provided to the group. Correctional links are also available for the scanned information.

The Wellcome Trust is expected to arrange for and fund back issue scanning of selected journals for deposit into the PMC archive. They will help with recruitment for the project and back issues will be from journals participating in PMC and will abide by the PMC regulations. The archives will add to the current back scanning project, amounting to one pipeline with two sources of funding. In order for European groups to maintain an archive, a portable PMC system will be needed for their use, which developers are currently designing.

In answering questions by the committee, Betsy Humphreys estimated that the total back issue print archive of all journals located and maintained at NLM is estimated at less than 20 million more articles.

NLM Archiving and Publishing DTDs

At this time, several journals are using the NLM Publishing DTD. HighWire Press is basing their next version of its system on the NLM DTD. JSTOR and CSIRO are also committed. A technical advisory group has been set up to provide feedback from the publishing and archiving communities on further development of the DTD. Jeff Beck provided a synopsis of the group's first meeting in August where discussions focused on tools to support the DTD, and work on a

schema version for modern XML tools. MicroSoft has agreed to collaborate on an MS word-based content editor that will output XML conforming to the NLM and JP DTDs.

Committee members posed questions regarding MicroSoft's project, asking if their development could make the "tool specific" problem worse. It is believed by the PMC group that MS involvement will help the community in the long run with open access publishing. A question was asked regarding the general availability of the NLM DTD via MS editors. Dr. Lipman responded that the goal is for users to be able to download additional software needed, i.e. the schema and add-ons for MS word, so that documents can be produced in the NLM DTD for submission. Also, archiving sites will eventually be able to download documents in the NLM DTD and be rendered for onsite archiving. Another question focused on the possibility of embedding XML data into the document to include information such as user rights, and provide external links to copyright information. Jeff responded that the idea of embedding XML into documents to create external links is being discussed within the working group.

PMC Usage

Dr. Lipman next provided the group with a graph of PMC use by month in 2003. There are currently 750,000 unique users per month. This number has grown in part due to the new global query feature in Entrez which searches a subset of databases. Google has indexed PMC articles so additional users are finding the site via this search engine. These figures are important for PMC participants as their articles are being viewed by an increasing number of users.

Discussion

The group then entered into a discussion about the new open access initiative. Mike Eisen told the group that PLoS had a lawyer draft a sample license for authors which he will send to David Lipman for distribution. A member asked if all authors own copyright equally on a paper. Technically the answer is no, and was expressed that open access is impeded when all own equal shares of copyright. A license agreement of rights for document usage is something that PLoS would like to see happen more frequently as it could make distribution easier.

The committee felt that it is important that NCBI promote the Open Access Initiative due to its significance. Dr. Lipman agreed, and information will be added to the NCBI homepage and the PMC homepage. Also, information will be given to committee members for their own personal use and distribution.

Dr. Lipman informed the group that he attended a JISC meeting in England recently where the subject was self-archiving. Some proponents of open access felt that the solution is for authors to put their articles on their own sites. In the physics community 70% of journals are freely available on the Los Alamos site but journals are still heavily used although the subscription cost of many has decreased. Dr. Lipman mentioned that LinkOut could provide some of the self-archiving functionality by allowing institutions to link their holdings with PubMed. Committee members noted that LinkOuts cannot be searched directly. Instead, articles must be viewed individually to get all LinkOut links highlighting a difference between LinkOut holdings and an archive.

It was expressed that other archives can potentially appear with which PMC could exchange data but Dr. Lipman said that institutional repositories are not emerging on a large scale for various

reasons. A member asked if there is a repository for NIH scientist publications and whether PMC could offer some type of guidance for NIH authors interested on open access. At this time there is no such repository and Dr. Lindberg stated that he does not know of any formal request being made on this topic. It was mentioned that an awareness-raising activity for NIH scientists would be beneficial. Constituencies exist on college campuses to inform faculty of new developments and available options and NIH could benefit from such a group.

Dr. Lipman then reviewed action items regarding open access. NLM will look into the issues in terms of a policy on accepting articles that do not have explicit publisher restrictions on the article. Richard Johnson will report back on progress of the SPARC subcommittee for author agreements with publishers. Mike Eisen will send an agreement that PLoS uses as an example for review. Information will be sent to members and feedback accepted on how to handle issues and distribute information to colleagues about open access.

A committee member asked about plans for new modules at NCBI related to chemical structures. Dr. Lipman informed the group about PubChem, a chemical structures project in the beginning stages, as part of the NIH Molecular Libraries Initiative. A freely available high throughput chemical screening repository will be created at NCBI that will include chemical structures with links to other NCBI databases. The database will also include a small molecule repository consisting of information from multiple sources.

Another question about non-participation in PMC was raised. A member asked specifically about failed negotiations with the *Annals of Internal Medicine*. Dr. Lipman stated that part of the problem could be fear of the unknown on the part of societies as far as copyright and usage. Members expressed that many technical options for journals now make it easier to participate. Dr. Lipman expressed that there is an open market for companies to provide societies with technical assistance that will allow them flexibility in their publishing. If private entities assist societies with electronic publishing, they may better understand the technical issues and provide more open access to their information.

Lunch 12:35 to 1:15

III. Dr. Lederberg Remarks

Dr. Lindberg thanked Dr. Lederberg for chairing the PMC committee for the past few years. Dr. Lederberg addressed the group as this is his final meeting as chair. He expressed that there has been enormous progress in promoting open dissemination of information during his involvement with the PMC committee. He stated that when the opportunity exists for openly accessible information, the science community should embrace it and provide as much as possible. Also, the gap between the technology potential for free information and what is actually available is great and he hopes that it will lessen in time. He thanked the group for the opportunity to act as chair.

IV. Genomics, Proteomics, and Expression Data

The NLM Board of Regents and the NCBI Board of Scientific Counselors (BoSC) made recommendations that NCBI devise a plan for dealing with the growing areas of genomics, proteomics, and expression data. Therefore, NCBI established a planning group comprised of past and present BoSC members to tackle the issues. Dr. Lipman briefed the group on two recent planning meetings held in July and October of 2003.

The topics of the planning group meetings centered around the core activity of NCBI defined as sequence resources and the role of NCBI in dealing with new types of scientific data, specifically functional genomics data. A committee consensus was the need to increase and improve usability of all NCBI resources. Dr. Lipman was requested to draft a proposal focusing on a solution to this issue which he submitted to the NLM administration.

Dr. Lipman stated that Entrez Global Query has been a step in increasing usability because of its capability of searching multiple databases at one time and NCBI has seen increased usage of many resources such as PMC, Books, and the NCBI web site search. Planning group participants unanimously agreed on the value of the upcoming PubChem resource. Other subjects broached had proponents and opponents because of possible resource intensity and current instability and infancy of the data, such as expression data and proteomics. The group consistently approved of projects with stable data and experimental methods that will not change drastically or require changes in database infrastructure over time.

The Gene Expression Omnibus (GEO), NCBI's gene expression database does not have significant resources allocated to it at this time and NCBI is trying to resolve long range plans for data collected. The planning group and BoSC advised that no additional resources be allocated for at least one year, when it will be reassessed. In the area of proteomics, results-oriented and post translational modification data is desired but is difficult to obtain at this time. A member asked about mass spec data. Dr. Lipman replied that NCBI is working on mass spec search tools at this time but they are not publicly available. Protein-protein interaction data will be forthcoming from a government funded group in Canada. It is likely that a planning meeting will be needed each year in order to analyze the development of growing fields and determine NCBI's role.

A committee member inquired about the full time employee (FTE) allocation difference between development of databases versus tool building. The three main areas that Dr. Lipman considered in his report to the planning committee are sequence core (primary data), curation, and comparison/classification. Within these areas, the resources allocated to database building are much higher than those committed to retrieval tools. Another member asked if there is a role for PMC in the areas between data and literature, such as a raw data archive. Dr. Lipman responded that data specification is needed by submitters and a resource would need to be developed by PMC. However, it is difficult to understand data without interpretation and participation of the scientific community is needed. Dr. Lederberg mentioned a new free courseware database by MIT and speculated on its impact.

Another member remarked on a new type of user he is noticing from students in late high school and early college and inquired about building PMC tools focused on less sophisticated users. Dr. Lipman answered that there are such opportunities and NCBI would be able to collaborate with others, either within or outside NIH, in providing a service that people could use in education. Sam Kaplan mentioned the idea of having a lay summary for all articles, which may require additional resources on the part of the journal. An NIH scientist is working with PMC on a project where top researchers in the field of epilepsy write critiques of research articles in order to help fellow scientists in the field as well as to identify key articles in the field. This resource is available in the form of an online electronic and paper journal and will hopefully be available soon in PMC.

V. Adjournment

The PubMed Central National Advisory Committee adjourned the public meeting at 2:00 p.m.

CERTIFICATION

I hereby certify that the foregoing minutes are accurate and complete.

(date)

Joshua Lederberg, Ph.D., Chair
PubMed Central National Advisory Committee

(date)

David J. Lipman, M.D., Director,
National Center for Biotechnology
Information, NLM