DEPARTMENT OF HEALTH AND HUMAN SERVICES

NATIONAL INSTITUTES OF HEALTH

PUBMED CENTRAL NATIONAL ADVISORY COMMITTEE

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION

NATIONAL LIBRARY OF MEDICINE

June 25, 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 – JUNE 25, 2003

The meeting of the PubMed Central National Advisory Committee was convened on June 25, 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 3:00 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
Paul Ginsparg, Ph.D., Cornell University
Richard Johnson, The Scholarly Publishing & Academic Resources Coalition (SPARC)
Heather D. Joseph, M.A., BioOne
Samuel Kaplan, Ph.D., University of Texas Medical School at Houston
Sarah Thomas, Ph.D., Cornell University
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

NLM Senior Staff Present
Kent Smith, Deputy Director, NLM

Donald King, M.D., Deputy Director for Research and Education, NLM
Betsy Humphreys, Associate Director for Library Operations, NLM

Visitors Present
Melissa Junior, American Society for Plant Biology
Diane Sullenberger, Proceedings of the National Academy of Science of the United States of America
Nancy Winchester, American Society for Plant Biology

I. Call to Order and Opening Remarks

Dr. Lipman welcomed members of the PubMed Central National Advisory Committee. The Committee officially adopted the minutes from the previous meeting. Tentative dates of November 17, 18, or 20, 2003 and May 10 or 17, 2004 were discussed for the upcoming two meetings. Members will be contacted to firm a date. Committee members and guests were then introduced. Dr. Lipman announced new PMC committee members who were not present at the meeting: Ajit Varki of the University of California, San Diego, Chaitan Khosla of Stanford University, Marc Kirschner of Harvard Medical School, and Gerald Rubin, Vice President, Howard Hughes Medical Institute.

Dr. Lipman notified the committee that a working group will meet this summer to discuss the direction of NCBI growth and project planning. NCBI would like a plan on managing projects without growing in personnel. NCBI would also like engage in new projects, but there is a need to ensure that existing goals and commitments are met. The working group will be comprised of NCBI senior staff, NIH institute directors, NLM senior staff, and past and current members of the NCBI Board of Scientific Counselors. A report will be drawn up and presented to Dr. Lindberg, the NLM Board of Regents, and other NIH institutes with an interest in working with NCBI. The main areas of concentration will be sequence databases, sequence classification, and text information.

II. PubMed Central Update

Journal Status

Dr. Lipman began the PMC update with journal information. The *Journal of Clinical Investigation* is available at this time and is being released on the issue date, with no delays. Cold Spring Harbor has deposited minimal content from 2002, and the PMC group is working with them to improve delivery. The *American Journal of Human Genetics* has started submitting production back files. The *EMBO Journal* is switching to the Nature Publishing Group later this year and *EMBO Reports* has already moved, which raises some questions regarding participation but EMBO has informed PMC of its intent to continue participation in PMC. *The Annals of Internal Medicine* agreed to participate and will be sending files in a few weeks. The PMC database currently contains just under 100,000 articles and other items.

A question was raised regarding Highwire Press participation and quality of information. Dr. Lipman responded that there is a production flow in place with most journals. The overall trend

is positive and all data suppliers are taking PMC's data quality standards more seriously. Committee members also posed questions regarding the PMC DTD and new participants. PMC will encourage new participants to use the PMC DTD if they do not already have one in use. Otherwise, NCBI will take theirs and convert it to the PMC DTD. The committee inquired about the overhead cost for conversion and quality assurance. Dr. Lipman replied that QA is an essential part of PMC and conversion is part of PMC's design philosophy. In fact, once converters are in place for a certain DTD, there is little additional work to do if another new journal uses it as well.

PMC is in discussion with the Institute for Scientific and Technical Information (Institut de L'Information Scientifque et Technique – INIST) of France and the National Institute of Genetics of Japan (NIG) to experiment with maintaining copies of the PMC archive. At this time, the ability of these groups to handle such a project is being assessed. If their ability looks promising, PMC will seek agreement from current participants. If PMC initiates the exchange for the information to go live, the databases will abide by the same content agreement as PMC.

Dr. Lipman provided background on the institutes who will be participating in this project. INIST provides access to a range of information, hosted at their central location to a mostly scientific community. They are currently able to make SGML and XML data available on the web and will do the same for PMC. NIG is working with the National Institute for Informatics, which is the site of the DNA DataBank of Japan (DDBJ), with whom NCBI has an established working relationship for exchange of sequence data.

These sites will not be mirror sites, but will contain the same information as PMC with a different interface and their own management of data. The exchange of data will be in XML, and each group's software will index the information and provide search and retrieval access. It was mentioned that diversity of access for the community can have a positive effect in data quality control. This has been demonstrated with the tripartite cooperation of the sequence databases, GenBank, DDBJ and EMBL.

Questions were raised by the committee relating to the possibility of defining a general set of requirements for contracts rather than customizing agreements for each publisher. Some publishers, however, may not want a general agreement since they have specific conditions for participating in PMC. A question was also raised regarding metadata for files in the PMC archive. It was pointed out that the XML contains article identification information such as the journal title, citation information, and copyright statement. The committee suggested a notification of inclusion for publishers regarding the "mirror" project in order to ensure a complete mirror of the PMC archive.

Break 10:50 to 11:05

Dr. Lipman explained to the group some advantages to open access in general. One advantage is integration with factual databases and literature. An example is the recently automated text analysis that identifies references to GenBank accession numbers that have been published (by the author) in the literature. This has resulted in the release of approximately 250 sequences in

the past three months that would have been otherwise held until GenBank was contacted about their publication in literature.

The committee inquired about the amount of text analysis research that NLM is supporting. A group within NCBI is conducting research. In addition, NLM is involved in intramural research within the Lister Hill Center in various textual analysis projects including the Unified Medical Language System (UMLS). Extramurally, the NLM funds several medical informatics centers. NHGRI, NIGMS, and NCI are also funding projects extramurally. Dr. Lipman is willing to send citation information to those interested in recent progress and notable findings in text analysis research in the life sciences. A question was also raised regarding outside groups using the PMC files for their own text analysis research. In the future, there may be information available via FTP on a publisher-by-publisher basis.

Upward Trend in PMC Use

Dr. Lipman next addressed PMC usage statistics, reporting an overall upward trend in PMC use. Since January 2003 there have been significant increases both in unique users of PMC and total retrieval of full text articles, due to improved access to PMC through Entrez, links from other areas of the NCBI web site, and Google referrals. Increased usage may also be due to over 1,000 libraries participating in LinkOut. Graphs illustrating corresponding increases for the ASM journals, BMJ, and the ASPB journals were also presented.

A discussion ensued regarding referrals from Google. Google indexes PMC articles and we are working with them to provide additional updated information. Special files are created for Google that contain the URLs of PMC records, so they can retrieve and index content that is normally dynamic and unavailable to web crawlers. Members were also interested in the categories of users for PMC.

Release of Archiving and Publishing DTDs

The Archiving and Publishing DTDs have been released and Dr. Lipman informed the group that TechBooks is using the journal publishing DTD to send content for the *Journal of Athletic Training*. Other small journals are using the DTD and HighWire is studying the DTD at this time. The release of the DTD prompted coverage in various publications, for which committee members requested citations. An XML Advisory Board is being established to advise NLM on continuing development of the DTDs. Representatives from major companies, publishers, and information sources are participants in this group.

Back Issue Scanning Project

An update of the back issue scanning project was presented. This project is moving according to plan and at this time approximately 90,000 pages have been delivered, with delivery of all content expected by April 2004. While details are still being worked out, output is progressively increasing. Some new groups have expressed interest in the project, and the possibility of their inclusion is being assessed. Mr. Sequeira showed the committee examples of articles delivered thus far. Images can be enlarged by clicking on a thumbnail and loaded separately, or viewed via

a PDF file. Scanned articles are searchable via OCR maps to the text image where the search word will be highlighted. The first complete journal, the *MLA Bulletin*, will be available online in September. More journals will be added incrementally as they are completed.

Committee members raised questions regarding cost efficiency. If PubMed does not have an abstract for an article the information is keyed, otherwise it is mapped to PubMed. Experiments on automatic citation matching are being done at this time, using bibliographic references taken directly from the unedited OCR, which is of high quality. Dr. Lipman noted that there will be no journals archived that are not PMC participants. Another question was raised regarding the political base for support of this project and for PMC as a whole. Dr. Lipman feels that the basis for the support of this project is due to the service it provides to both participants and the public, making it a stable project. Mr. Smith read a favorable comment from the Congressional Appropriations Committee regarding PMC.

III. Proposed PMC Policy Changes

The first policy issue discussed was the criterion that states participating journals must have three scientists on the masthead who are currently grantees from a major funding agency. The *Journal for Independent Medical Research* has petitioned PMC to relax this policy. Members expressed the opinion that this criterion is not difficult to meet. It was unanimously decided to maintain current policy, and that there is no reason to make this rule either more or less restrictive.

Lunch 12:15-1:00

PubLink Policy

Dr. Lipman reviewed the current policy for participants using the PubLink option in PMC, which allows articles to be viewed on the publisher site via links from PMC. While content must be made available within one year, it is not necessarily available on the PMC site. This policy was enacted two years ago in order to enable publishers with specific concerns to participate in PMC. Some publishers cited concerns about quality of information and the possibility of a reduced presence on the web as their reasons for nonparticipation. The PubLink option was instrumental in the participation of ASM. However, it did not change the decision of other high profile journals to join as they had indicated. It has been the experience of the PMC team that the linkout option does not impact the decision of most publisher participation. It is also the opinion of the PMC team that this option clouds the focus of PMC and the current, open-ended agreement limits the quality of an archive because the content is not actually viewable in PMC. Since the data is not in PMC it does not benefit from the data validation provided by users confirming that the material is readable and usable.

Dr. Lipman proposed to eliminate the PubLink option for new participating publishers, returning to the original PMC model. Current PubLink journals will be grandfathered but will be asked to consider allowing all content to be viewable within PMC. The committee was unanimously in favor of this motion.

Advisors suggested that guidelines be drawn up as follows: content deposited within one month after publication, primary research articles available on PMC within one year, all other content available on PMC within three years.

An update was provided about a meeting sponsored by scientists from the Howard Hughes Medical Institute which is supporting open access publishing by the investigators it funds. The June 9 meeting included members of the academic, publishing, and funding communities. The goal of the meeting was to draw up a document with a formal definition of 'Open Access Publication' along with a set of recommendations for implementation. Along with the Open Access definition, another document was drawn up, titled the "Bethesda Manifesto" in support of the open access principles, which is to be signed by the various participants. PMC was mentioned as a site for long term biomedical sciences content archiving. The group is drawing up a license that would be available for publishers, with a definition of open access information.

IV. Conclusion

Drs. Lederberg and Lipman thanked both the Committee members and invited guests for their valuable time and input.

V. Adjournment

The PubMed Central National Advisory Committee adjourned the public meeting at 3:00 p.m.

CERTIFICATION

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

_____	_____
(date)	(date)
Joshua Lederberg, Ph.D., Chair PubMed Central National Advisory Committee	David J. Lipman, M.D., Director, National Center for Biotechnology Information, NLM