Testimony before the
Subcommittee on Information Policy, Census and National Archives
Committee on Oversight and Government Reform
United States House of Representatives

Public Access to Federally-Funded Research

Statement of
David J. Lipman, M.D.

Director
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
U.S. Department of Health and Human Services

For Release on Delivery
Expected at 2:00 p.m.
July 29, 2010

Mr. Chairman, members of the Subcommittee, it is my pleasure to testify before you today. My name is David J. Lipman. I am the Director of the National Center for Biotechnology Information (NCBI), which is part of the National Library of Medicine (NLM) at the National Institutes of Health (NIH), an agency of the Department of Health and Human Services. NCBI was chartered by Congress in 1988 to employ computer systems to collect and disseminate the results of biotechnology research, and we have been doing so ever since. NCBI is the home of more than 40 free and Internet-accessible databases, including GenBank, the database of all publicly available DNA sequences. It is also the home of dbGaP, a research database of studies that investigate the links between genetic variations and diseases. And, closer to the theme of today's hearing, NCBI is the home of PubMed Central – the publicly accessible, online archive of peer-reviewed biomedical sciences literature and the repository for NIH-funded papers submitted in compliance with the NIH Public Access Policy.

NIH has operated PubMed Central for more than a decade and has had a Public Access Policy in place for the last five years. During this time, NIH has gained considerable experience that I would like to share with you today as the subcommittee considers legislation to expand public access policies to other Federal science agencies and examines the systems and processes that might be put in place to do so. Our experience has demonstrated that policies such as the NIH Public Access Policy and repositories such as PubMed Central are important elements of efforts to develop an information infrastructure that will advance basic science, accelerate its application to solving today's problems, and satisfy a growing public desire for transparency and access to scientific information.

In launching PubMed Central in 2000, NIH aimed to follow the successful example of the Human Genome Project and promote scientific discovery by taking advantage of opportunities created by information technology and the Internet. Development of a digital archive of biomedical journal articles was seen as a way to improve access to cutting-edge research and to provide a long-term, stable repository of the scientific literature that researchers could continue to draw on in their work, recognizing the cumulative nature of science. From the beginning, we were fortunate to have the collaboration of a number of publishers who offered to deposit their journals in PubMed Central to make them widely accessible. As PubMed Central grew, we gained considerable experience in building and operating a digital repository for journal articles. Among the highlights of these efforts was establishment of a structured digital format for representing journal articles – the NLM DTD. The format has been adopted by some major publishers and libraries, including the Library of Congress, the British Library, and HighWire Press, and is in the process of becoming a standard recognized by the National Information Standards Organization.

Early experience with PubMed Central illustrated the benefits that a centralized repository of the biomedical literature could have, not only for scientists, but for medical practitioners, companies involved in the development of medical products and services, and the public. Without a resource like PubMed Central, the general public does not have ready access to much of the biomedical literature, and even large academic institutions and drug and device companies can lack access to the broad set of journals that might be relevant to their efforts. It also became apparent that PubMed Central could serve as an institutional archive for articles describing the research that results from NIH funding – articles for which no other systematic archive had been assembled. In 2005, NIH announced its first public access policy. The policy

was viewed as a way to keep a central archive of NIH-funded research publications and preserve vital medical research results and information for years to come; to advance science by creating an information resource that would make it easier for scientists to mine medical research publications; to help NIH better manage its research investment; and to provide ready access to NIH-funded published research for patients, families, health professionals, scientists, teachers, and students. This initial policy was voluntary. Specifically, the policy *requested* recipients of NIH funding to deposit a copy of their peer-reviewed manuscripts in PubMed Central upon acceptance for publication. They were permitted to delay public availability of the article in PubMed Central for as long as 12 months after the official date of publication.

Only some 5% of the articles that were subject to the initial policy were voluntarily submitted by their authors. Other NIH-funded articles were received directly from journals that participated in PubMed Central, but still only 19% of the articles subject to the NIH Public Access Policy between May 2005 and December 2007 were included in PubMed Central. To improve compliance, Congress, as part of the Consolidated Appropriations Act for FY 2008, instructed NIH to make the public access policy mandatory, which it did starting in April 2008. As of that date NIH-funded researchers have been *required* to submit copies of their peer reviewed journal articles to PubMed Central upon acceptance for publication. As with the voluntary policy, up to a 12-month delay for public access to the articles can be requested.

The transition to a mandatory policy has had a dramatic effect on the deposit of papers into PubMed Central. Of the estimated 88,000 NIH-funded articles published in 2009, approximately 70% have been submitted to PubMed Central. That figure continues to climb as NIH works with the research community to promote awareness of the policy, improves its ability

to track papers resulting from NIH research awards, and develops new systems to assist sponsored research offices at universities and medical research centers in tracking their compliance with the policy

NIH has also taken steps to simplify the submission process for authors. For articles that are published in a journal that participates in PubMed Central, the authors need to do nothing once their article has been accepted for publication. The publisher directly deposits the author's final article into PubMed Central. For articles that are not published in participating journals, authors submit the articles themselves using the NIH Manuscript Submission System, a process that takes only about 10 minutes. At present, more than 900 journals have formal agreements with PubMed Central to deposit the published version of all NIH-funded articles in PubMed Central, a number that has doubled in the 2 years since the policy became mandatory. As a result of these arrangements, approximately 40% of the articles submitted to PubMed Central in 2009 were deposited directly by the publisher, with no additional intervention by the author. That percentage is expected to continue to climb as more journals make arrangements for submitting articles on behalf of their authors.

As a result of these efforts, PubMed Central has continued to grow. Between April 2008 (when the policy became mandatory) and June 2010, approximately 700,000 articles were added to PubMed Central, bringing the total content of the archive to more than 2 million full-text articles. Of those 700,000 added articles, approximately 130,000 report on NIH-funded research. With increased content has come increased usage. In the two years between March 2008 and March 2010, the monthly number of articles retrieved from PubMed Central doubled from 10 million to 20 million. On a typical weekday in March 2010, some 420,000 different users

retrieved 740,000 articles from PubMed Central. Those visitors included more than 2,800 users from Missouri, 21,000 users from California, and 4,800 users from North Carolina. And they access a significant portion of the available content. Last year, 99% of the articles in PubMed Central were downloaded at least once, and 28% were downloaded more than 100 times.

Although we can collect only aggregated information about users of PubMed Central, we can infer they represent a mix of people from the education and business sectors, as well as private citizens. Based on the type of Internet domain from which they access PubMed Central (e.g., .com, .edu, .net, .gov), we estimate that approximately 25% of our users are from universities, 40% are private citizens or those using personal Internet accounts, and 17% are from companies (the remainder consists of government users or others). These kinds of numbers support the notion that PubMed Central has become a broad-based repository for researchers, students, clinicians, entrepreneurs, patients and their families.

The success of the NIH model has stimulated similar efforts in other countries. Major biomedical research funding organizations in the United Kingdom, including the Wellcome Trust, Medical Research Council, and National Institute for Health Research, have access policies similar to NIH's that require funded authors to ensure that articles are publicly accessible. The Canadian Institutes of Health Research also requires funded researchers to ensure that research papers are publicly accessible. In both the U.K. and Canada, funding agencies are using a portable version of the PubMed Central software (developed by NLM) to build their repositories. NIH's collaboration with these organizations has demonstrated the capability to establish interoperable archives at other sites. It has also expanded the access that

users in the United States have to research results resulting from the growing amounts of biomedical research that are conducted in other countries.

But to look at PubMed Central as just a repository for scientific articles is to miss the bigger picture. PubMed Central has become an integral part of a larger information infrastructure that is accelerating scientific discovery in the biomedical sciences. Articles contained in PubMed Central are another entry point into the larger body of biomedical information that is maintained by NCBI and NLM. As noted above, NCBI produces more than 40 databases, including GenBank and dbGaP. NCBI and NLM also maintain information about small, biologically significant molecules that are assayed through the NIH Molecular Libraries program, information about the results of clinical trials – the result of recent legislation – and 3-dimensional structures of proteins. Every day, users download over 13 trillion bytes of data from NCBI – equivalent to all the books in the Library of Congress. Interpreting and understanding this data requires access to the knowledge that is embodied in scientific articles. By having journal articles in PubMed Central in a machine readable format, we are able to create linkages among these resources that can aid and advance scientific discovery. For example, during the recent H1N1 flu pandemic, NCBI was the major site for collecting all of the known flu sequences. Within months, NCBI had over 20,000 sequences from around the world. Taking advantage of the deep integration among NCBI systems, a researcher reading a paper on the spread of drug-resistant variants of the flu sequences could, with the click of a mouse, compare the new isolates to all other flu variants and gain insight into the epidemiological consequences. With equal ease, the researcher could map the variant viral proteins to known 3D protein structures to see how the mutations affect binding of the antiviral drug.

Already, a significant fraction of the users who access data from an NCBI database on any given day also retrieve articles from PubMed Central and vice-versa. More than 17% of dbGaP users (for studying the genetic basis of disease), for example, also use PubMed Central. This type of iteration between the literature and data increasingly reflects the way that research in biomedical sciences is done, as biomedical science becomes an ever more data-intensive science. This interoperability is difficult to achieve if the literature – the knowledge – is widely dispersed and unconnected to other databases of biomedical information.

Furthermore, because the searching and navigating among databases takes place within our integrated database structure, we are able to continually refine our information systems to make them more helpful to our users. We can examine on a regular basis how the system is used and how users navigate from one database to another, and we can improve the systems to help users find the information they are looking for. For example, NCBI recently began adding what we call "discovery ads" to pages in PubMed Central. These ads, placed adjacent to an appropriate passage in the text, provide references to other related articles that are indexed in NLM's PubMed database of more than 16 million journal abstracts. Since adding this capability, we have almost doubled – in a 1-year period – the rate at which users move from PubMed Central to PubMed as they review the scientific literature. Links from PubMed Central to other NCBI databases connect users to related data.

Equally important, we are able to do these activities cost-effectively. Startup costs for developing the system that handles articles submitted under the NIH Public Access Policy were about $500,000. Annual operating costs for the system, including ingest of articles, refinement of the submission system and search tools, staffing of a help desk and a central coordinating

office for NIH, are approximately $3.5-$4.0 million per year. This represents a small fraction of NIH's budget authority of more than $30 billion per year. We keep our costs low because of the incredibly skilled staff we have assembled at NCBI and because we can leverage NLM's existing infrastructure and services, as well as many other resources available at NIH.

In summary, our experience with PubMed Central and the NIH Public Access Policy show that such approaches can be a cost-effective means to enhance access to the results of scientific research – in particular federally funded research – to preserve and increase the use of research results, and to enhance scientific discovery. The NIH Public Access Policy is a critical element of the agency's efforts to enhance opportunities for scientific discovery. It ensures that the scientific knowledge that is generated by the Government's investment in biomedical research and that is documented in peer reviewed articles is integrated into the information infrastructure that has become fundamental to continued progress in biomedical science. Having a comprehensive resource that integrates knowledge and data speeds the discovery process that is critical for improving human health.

Thank you for the opportunity to present our experiences to you. I would be happy to answer any questions you might have.