Statement of Stuart M. Shieber before the
Committee on Science, Space and Technology
Subcommittee on Investigations and Oversight
U.S. House of Representatives

March 29, 2012

Chairman Broun and Members of the Subcommittee:

My name is Stuart Shieber. I am the James O. Welch, Jr. and Virginia B. Welch Professor of Computer Science at Harvard University. My primary field of research is computational linguistics, the study of human language from a computer science perspective, often with application to the engineering of useful computer systems that manipulate language. As a faculty member, I led the development and enactment of Harvard's open-access policies. Since October of 2008, I have served in the additional role as the faculty director of Harvard's Office for Scholarly Communication. Thank you for the opportunity to speak with you today about some of the actions that we have taken at Harvard to provide the broadest possible access to the results of our research.

THE POTENTIAL FOR OPEN ACCESS

The mission of the university is to create, preserve, and disseminate knowledge to the benefit of all. In Harvard's Faculty of Arts and Sciences (FAS), where I hold my faculty post, we codify this in the FAS Grey Book, which states that research policy "should encourage the notion that ideas or creative works produced at the University should be used for the greatest possible public benefit. This would normally mean the widest possible dissemination and use of such ideas or materials."

At one time, the widest possible dissemination was achieved by distributing the scholarly articles describing the fruits of research in the form of printed issues of peer-reviewed journals, sent to the research libraries of the world for reading by their patrons, and paid for by subscription fees. These fees covered the various services provided to the authors of the articles — management of the peer review process, copy-editing, typesetting, and other production processes — as well as the printing, binding, and shipping of the physical objects.

Thanks to the forward thinking of federal science funding agencies, including NSF, DARPA, NASA, and DOE, we now have available computing and networking technologies that hold the promise of transforming the mechanisms for disseminating and using knowledge in ways not imaginable even a few decades ago. The internet allows nearly instantaneous distribution of content for essentially zero marginal cost to a large and rapidly increasing proportion of humanity. Ideally, this would ramify in a universality of access to research results, thereby truly achieving the widest possible dissemination.

The benefits of such so-called *open access* are manifold. The signatories of the 2002 Budapest Open Access Initiative state that

> The public good [open access] make[s] possible is the world-wide electronic distribution of the peer-reviewed journal literature and completely free and unrestricted access to it by all scientists, scholars, teachers, students, and other curious minds. Removing access barriers to this literature will accelerate research, enrich education, share the learning of the rich with the poor and the poor with the rich, make this literature as useful as it can be, and lay the foundation for uniting humanity in a common intellectual conversation and quest for knowledge.

From a more pragmatic point of view, a large body of research has shown that public research has a large positive impact on economic growth, and that access to the scholarly literature is central to that impact. Martin and Tang's recent review of the literature concludes that "there have been numerous attempts to measure the economic impact of publicly funded research and development (R&D), all of which show a large positive contribution to economic growth."[1] It is therefore not surprising that Houghton's modeling of the effect of broader public access to federally funded research shows that the benefits to the US economy come to the billions of dollars and are eight times the costs.[2]

Opening access to the literature makes it available not only to human readers, but to computer processing as well. There are some million and a half scholarly articles published each year.[3] No human can read them all or even the tiny fraction in a particular subfield, but computers can, and computer analysis of the text, known as *text mining*, has the potential not only to extract high-quality structured data from article databases but even to generate new research hypotheses. My own field of research, computational linguistics, includes text mining. I have collaborated with colleagues in the East Asian Languages and Civilization department on text mining of tens of thousands of classical Chinese biographies and with colleagues in the History department on computational analysis of pre-modern Latin texts. Performing similar analyses on the current research literature, however, is encumbered by proscriptions of copyright and contract because the dominant publishing mechanisms are not open.

---

[1] Ben R. Martin and Puay Tang, The benefits from publicly funded research, SEWPS Paper No. 161, SPRU—Science and Technology Policy Research, University of Sussex, Brighton (2007). `http://www.sussex.ac.uk/spru/documents/sewp161`

[2] John Houghton, *Economic and Social Returns on Investment in Open Archiving Publicly Funded Research Outputs* (July 2010). `http://www.arl.org/sparc/bm~doc/vufrpaa`

[3] Scholarly Publishing Roundtable, *Report and Recommendations from the Scholarly Publishing Roundtable* (January 2010). `http://www.aau.edu/WorkArea/DownloadAsset.aspx?id=10044`

In Harvard's response to the Office of Science and Technology Policy's request for information on public access,[4] Provost Alan Garber highlighted the economic potential for the kinds of reuse enabled by open access.

> Public access not only facilitates innovation in research-driven industries such as medicine and manufacturing. It stimulates the growth of a new industry adding value to the newly accessible research itself. This new industry includes search, current awareness, impact measurement, data integration, citation linking, text and data mining, translation, indexing, organizing, recommending, and summarizing. These new services not only create new jobs and pay taxes, but they make the underlying research itself more useful. Research funding agencies needn't take on the job of provide all these services themselves. As long as they ensure that the funded research is digital, online, free of charge, and free for reuse, they can rely on an after-market of motivated developers and entrepreneurs to bring it to users in the forms in which it will be most useful. Indeed, scholarly publishers are themselves in a good position to provide many of these value-added services, which could provide an additional revenue source for the industry.

Finally, free and open access to the scholarly literature is an intrinsic good. It is in the interest of the researchers generating the research and those who might build upon it, the public who take interest in the research, the press who help interpret the results, and the government who funds these efforts. All things being equal, open access to the research literature ought to be the standard.

SYSTEMIC PROBLEMS IN THE JOURNAL PUBLISHING SYSTEM

Unfortunately, over the last several years, it has become increasingly clear to many that this goal of the "widest possible dissemination" was in jeopardy because of systemic problems in the current mechanisms of scholarly communication, which are not able to take full advantage of the new technologies to maximize the access to research and therefore its potential for social good.

By way of background, I should review the standard process for disseminating research results. Scholars and researchers — often with government funding — perform research and write up their results in the form of articles, which are submitted to journals that are under the editorial control of the editor-in-chief and editorial boards made up of other scholars. These editors find appropriate reviewers, also scholars, to read and provide detailed reviews of the articles, which authors use to improve the quality of the articles. Reviewers also provide advice to the editors on whether the articles are appropriate for publication in the journal, the final decisions being

[4] Alan Garber, Harvard response to the White House RFI on public access to research (January 2012). `http://osc.hul.harvard.edu/stp-rfi-response-january-2012`

made by the editors. Participants in these aspects of the publishing process are overwhelmingly volunteers, scholars who provide their time freely as a necessary part of their engagement in the research enterprise. The management of this process, handling the logistics, is typically performed by the journal's publisher, who receives the copyright in the article from the author for its services. The publisher also handles any further production process such as copy-editing and typesetting of accepted articles and their distribution to subscribers through print issue or more commonly these days through online access. This access is provided to researchers by their institutional libraries, which pay for annual subscriptions to the journals.

Libraries have observed with alarm a long-term dramatic rise in subscription costs of journals. The Association of Research Libraries, whose members represent the leading research libraries of the United States and Canada, have tracked serials expenditures for over three decades. From 1986 through 2010 (the most recent year with available data), expenditures in ARL libraries have increased by a factor of almost 5. Even discounting for inflation, the increase is almost 2.5 times. These increases correspond to an annualized rate of almost 7% per year, during a period in which inflation has averaged less than 3%.[5]

Another diagnostic of the market dysfunction in the journal publishing system is the huge disparity in subscription costs between different journals. Bergstrom and Bergstrom showed that even within a single field of research, commercial journals are *on average* five times more expensive per page than non-profit journals.[6] When compared by cost per citation, which controls better for journal quality, the disparity becomes even greater, a factor of 10 times. Odylzko notes that "The great disparity in costs among journals is a sign of an industry that has not had to worry about efficiency."[7] Finally, the extraordinary profit margins, increasing even over the last few years while research libraries' budgets were under tremendous pressure, provide yet another signal of the absence of a functioning competitive market.

The Harvard library system is the largest academic library in the world, and the fifth largest library of any sort. In attempting to provide access to research results to our faculty and students, the university subscribes to tens of thousands of serials at a cost of about 9 million dollars per year. Nonetheless, we too have been buffeted by the tremendous growth in journal costs over the last decades, with Harvard's serials expenditures growing by a factor of 3 between 1986 and

[5]Association of Research Libraries, Monograph and Serial Costs in ARL Libraries, 1986-2010 (2010). `http://www.arl.org/bm~doc/t2_monser10.xls`

[6]Carl T. Bergstrom and Theodore C. Bergstrom, The costs and benefits of library site licenses to academic journals, *Proceedings of the National Academy of Sciences*, volume 101, number 3 (20 January 2004). `http://dx.doi.org/10.1073/pnas.0305628101`

[7]Andrew Odlyzko, The Economics of Electronic Journals, *First Monday*, volume 2, number 8 (4 August 1997). `http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/542/463`

2004.[8] Such geometric increases in expenditures could not be sustained indefinitely. Over the years since 2004 our journal expenditure increases have been curtailed through an aggressive effort at deduplication, elimination of print subscriptions, and a painful series of journal cancellations. As a researcher, I know that Harvard does not subscribe to all of the journals that I would like access to for my own research, and if Harvard, with its scale, cannot provide optimal subscription access, other universities without our resources are in an even more restricted position.

Correspondingly, the articles that we ourselves generate as authors are not able to be accessed as broadly as we would like. We write articles not for direct financial gain — we are not paid for the articles and receive no royalties — but rather so that others can read them and make use of the discoveries they describe. To the extent that access is limited, those goals are thwarted.

The economic causes of these observed phenomena are quite understandable. Journal access is a monopolistic good. Libraries can buy access to a journal's articles only from the publisher of that journal, by virtue of the monopoly character of copyright. In addition, the high prices of journals are hidden from the "consumers" of the journals, the researchers reading the articles, because an intermediary, the library, pays the subscriptions on their behalf. The market therefore embeds a moral hazard. Under such conditions, market failure is not surprising; one would expect inelasticity of demand, hyperinflation, and inefficiency in the market, and that is what we observe. Prices inflate, leading to some libraries canceling journals, leading to further price increases to recoup revenue — a spiral that ends in higher and higher prices paid by fewer and fewer libraries. The market is structured to provide institutions a Hobson's choice between unsustainable expenditures or reduced access.

The unfortunate side effect of this market dysfunction has been that as fewer libraries can afford the journals, access to the research results they contain is diminished. In 2005, then Provost of Harvard Steven Hyman appointed an ad hoc committee, which I chaired, to examine these issues and make recommendations as to what measures Harvard might pursue to mitigate this problem of access to our writings. Since then, we have been pursuing a variety of approaches to maximize access to the writings of Harvard researchers.

ADDRESSING INSUFFICIENT ACCESS THROUGH AN OPEN-ACCESS POLICY

One of these approaches involves the self-imposition by faculty of an open-access policy according to which faculty grant a license to the university to distribute our scholarly articles and commit to providing copies of our manuscript articles for such distribution. By virtue of this kind of policy, the problem of access limitation is mitigated by providing a supplemental venue for access to the

[8] Association of Research Libraries, Monograph and Serial Costs in ARL Libraries, 1986-2010 (2010). `http://www.arl.org/bm~doc/t2_monser10.xls`

articles. Four years ago, in February of 2008, the members of the Faculty of Arts and Sciences at Harvard became the first school to enact such a policy,[9] by unanimous vote as it turned out.

In order to guarantee the freedom of faculty authors to choose the rights situation for their articles, the license is waivable at the sole discretion of the author, so faculty retain control over whether the university is granted this license. But the policy has the effect that by default, the university holds a license to our articles, which can therefore be distributed from a repository that we have set up for that purpose. Since the FAS vote, six other schools at Harvard — Harvard Law School, Harvard Kennedy School of Government, Harvard Graduate School of Education, Harvard Business School, Harvard Divinity School, and Harvard Graduate School of Design — have passed this same kind of policy, and similar policies have been voted by faculty bodies at many other universities as well, including Massachusetts Institute of Technology, Stanford, Princeton, Columbia, and Duke. Notably, the policies have seen broad faculty support, with faculty imposing these policies on themselves typically by unanimous or near unanimous votes.

Because of these policies in the seven Harvard schools, Harvard's article repository, called DASH (for Digital Access to Scholarship at Harvard),[10] now provides access to over 7,000 articles representing 4,000 Harvard-affiliated authors. Articles in DASH have been downloaded almost three-quarters of a million times.[11] The number of waivers of the license has been very small; we estimate the waiver rate at about 5%. Because of the policy, as faculty authors we are retaining rights to openly distribute the vast majority of the articles that we write.

The process of consultation in preparation for the faculty vote was a long one. I started speaking with faculty committees, departments, and individuals about two years before the actual vote. During that time and since, I have not met a single faculty member or researcher who objected to the principle underlying the open-access policies at Harvard, to obtain the widest possible dissemination for our scholarly results, and have been struck by the broad support for the kind of open dissemination of articles that the policy and the repository allow.

This approach to the access limitation problem, the provision of supplemental access venues, is also seen in the extraordinarily successful public access policy of the National Institutes of Health (NIH), which Congress mandated effective April, 2008. By virtue of that policy, researchers funded by NIH provide copies of their articles for distribution from NIH's PubMed Central (PMC) repository. Today, PMC provides free online access to 2.4 million articles downloaded a million times per day by half a million users.[12] NIH's own analysis has shown that a quarter of the users

[9]Text of the FAS policy and the other Harvard open-access policies is available at `http://osc.hul.harvard.edu/policies`.

[10]`http://dash.harvard.edu/`

[11]`http://dash.harvard.edu/mydash`

[12]National Institutes of Health, *NIH Public Access Policy Implications* (2012). `http://publicaccess.nih.gov/public_access_policy_implications_2012.pdf`

are researchers. The hundreds of thousands of articles they are accessing per day demonstrates the large latent demand for articles not being satisfied by the journals' subscription base. Companies account for another 17%, showing that the policy benefits small businesses and corporations, who need access to scientific advances to spur innovation. Finally, the general public accounts for 40% of the users, some quarter of a million people per day, demonstrating that these articles are of tremendous interest to the taxpayers who fund the research in the first place and who deserve access to the results that they have underwritten.

THE STANDARD OBJECTION TO OPEN-ACCESS POLICIES

The standard objection to these open-access policies is that supplemental access to scholarly articles, such as that provided by institutional repositories like Harvard's DASH or subject-based repositories like NIH's PubMed Central, could supplant subscription access to such an extent that subscriptions would come under substantial price pressure. Sufficient price pressure, in this scenario, could harm the publishing industry, the viability of journals, and the peer review and journal production processes.

There is no question that the services provided by journals are valuable to the research enterprise, so such concerns must be taken seriously. By now, however, these arguments have been aired and addressed in great detail. I recommend the report "The Future of Taxpayer-Funded Research: Who Will Control Access to the Results?" by my co-panelist Elliott Maxwell,[13] which provides detailed support for the report's conclusion that "There is no persuasive evidence that increased access threatens the sustainability of traditional subscription-supported journals, or their ability to fund rigorous peer review." The reasons are manifold, including the fact that supplemental access covers only a fraction of the articles in any given journal, is often delayed relative to publication, and typically provides a manuscript version of the article rather than the version of record. Consistent with this reasoning, the empirical evidence shows no such discernible effect. After four years of the NIH policy, for instance, subscription prices have continued to increase, as have publisher margins. The NIH states that "while the U.S. economy has suffered a downturn during the time period 2007 to 2011, scientific publishing has grown: The number of journals dedicated to publishing biological sciences/agriculture articles and medicine/health articles increased 15% and 19%, respectively. The average subscription prices of biology journals and health sciences journals increased 26% and 23%, respectively. Publishers forecast increases to the rate of growth of the medical journal market, from 4.5% in 2011 to 6.3% in 2014."[14]

[13]Committee for Economic Development. *The Future of Taxpayer-Funded Research: Who Will Control Access to the Results?* (2012). `http://www.ced.org/component/blog/entry/1/765`

[14]National Institutes of Health, *NIH Public Access Policy Implications* (2012). `http://publicaccess.nih.gov/public_access_policy_implications_2012.pdf`

Nonetheless, it does not violate the laws of economics that increased supplemental access (even if delayed) to a sufficiently high proportion of articles (even if to a deprecated version) could put price pressure on subscription journals, perhaps even so much so that journals would not be able to recoup their costs. In this hypothetical case, would that be the end of journals? No, because even if publishers (again, merely by hypothesis and counterfactually) add no value for the readers (beyond what the readers are already getting in the [again hypothetical] universal open access), the author and the author's institution gain much value: vetting, copyediting, typesetting, and most importantly, imprimatur of the journal. This is value that authors and their institutions should be, would be, and are willing to pay for. The upshot is that journals will merely switch to a different business model, in which the journal charges a one-time *publication fee* to cover the costs of publishing the article.

I state this as though this publication-fee revenue model is itself hypothetical, but it is not. Open-access journals already exist in the thousands. They operate in exactly the same way as traditional subscription journals — providing management of peer review, production services, and distribution — with the sole exception that they do not charge for online access, so that access is free and open to anyone. The publication-fee revenue model for open-access journals is a proven mechanism. The prestigious non-profit open-access publisher Public Library of Science is generating surplus revenue and is on track to publish some 3% of the world biomedical literature through its journal *PLoS ONE* alone. The BioMed Central division of the commercial publisher Springer is generating profits for its parent company using the same revenue model. Indeed, the growth of open-access journals over the past few years has been meteoric. There are now over 7,000 open-access journals,[15] many using the publication-fee model, and many of the largest, most established commercial journal publishers — Elsevier, Springer, Wiley-Blackwell, SAGE — now operate open-access journals using the publication-fee revenue model. Were supplemental access to cause sufficient price pressure to put the subscription model in danger, the result would merely be further uptake of this already burgeoning alternative revenue model.

In this scenario, the cost of journal publishing would be borne not by the libraries on behalf of their readers, but by funding agencies and research institutions on behalf of their authors. Already, funding agencies such as Wellcome Trust and Howard Hughes Medical Institute underwrite open access author charges, and in fact mandate open access. Federal granting agencies such as NSF and NIH allow grant funds to be used for open-access publication fees as well (though grantees must

[15]According to the Directory of Open Access Journals, `http://www.doaj.org/`.

prebudget for these unpredictable charges). Not all fields have the sort of grant funding opportunities that could underwrite these fees. For those fields, the researcher's employing institution, as de facto funder of the research, should underwrite charges for publication in open-access journals. Here again, Harvard has taken an early stand as one of the initial signatories — along with Cornell, Dartmouth, MIT, and University of California, Berkeley — of the Compact for Open-Access Publishing Equity,[16] which commits these universities and the dozen or so additional signatories to establishing mechanisms for underwriting reasonable open-access publication fees. The Compact acknowledges the fact that the services that journal publishers provide are important, cost money, and deserve to be funded, and commits the universities to doing so, albeit with a revenue model that avoids the market dysfunction of the subscription journal system.

ADVANTAGES OF THE OPEN-ACCESS PUBLISHING SYSTEM

The primary advantage of the open-access journal publishing system is the open access that it provides. Since revenue does not depend on limiting access to those willing to pay, journals have no incentive to limit access, and in fact have incentive to provide as broad access as possible to increase the value of their brand. In fact, open-access journals can provide access not only in the traditional sense, allowing anyone to access the articles for the purpose of reading them, but can provide the articles unencumbered by any use restrictions, thereby allowing the articles to be used, re-used, analyzed, and data-mined in ways we are not even able to predict.

A perhaps less obvious advantage of the publication-fee revenue model for open-access journals is that the factors leading to the subscription market failure do not inhere in the publication-fee model. Bergstrom and Bergstrom[17] explain why:

> Journal articles differ [from conventional goods such as cars] in that they are not substitutes for each other in the same way as cars are. Rather, they are complements. Scientists are not satisfied with seeing only the top articles in their field. They want access to articles of the second and third rank as well. Thus for a library, a second copy of a top academic journal is not a good substitute for a journal of the second rank. Because of this lack of substitutability, commercial publishers of established second-rank journals have substantial monopoly power and are able to sell their product at prices that are much higher than their average costs and several times higher than the price of higher quality, non-profit journals.

[16] `http://www.oacompact.org/`. See also Stuart M. Shieber, Equity for open-access journal publishing, *PLoS Biology*, volume 7, number 8 (2012). `http://dx.doi.org/10.1371/journal.pbio.1000165`

[17] Theodore C. Bergstrom and Carl T. Bergstrom, Can 'author pays' journals compete with 'reader pays'?, *Nature Web Focus* (2004). `http://www.nature.com/nature/focus/accessdebate/22.html`

> By contrast, the market for authors' inputs appears to be much more competitive. If journals supported themselves by author fees, it is not likely that one Open Access journal could charge author fees several times higher than those charged by another of similar quality. An author, deciding where to publish, is likely to consider different journals of similar quality as close substitutes. Unlike a reader, who would much prefer access to two journals rather than to two copies of one, an author with two papers has no strong reason to prefer publishing once in each journal rather than twice in the cheaper one.
>
> If the entire market were to switch from Reader Pays to Author Pays, competing journals would be closer substitutes in the view of authors than they are in the view of subscribers. As publishers shift from selling complements to selling substitutes, the greater competition would be likely to force commercial publishers to reduce their profit margins dramatically.

Again, the empirical evidence supports this view. Even the most expensive open-access publication fees, such as those of the prestigious Public Library of Science journals, are less than $3,000 per article, with a more typical value in the $1,000–1,500 range. By contrast, the average revenue per article for subscription journal articles is about $5,000. Thus, the open-access model better leverages free market principles: Despite providing unencumbered access to the literature, it costs no more overall per article, and may end up costing much less, than the current system. The savings to universities and funding agencies could be substantial.

CONCLUSION

I began my comments by quoting the mission of academics such as myself to provide the widest possible dissemination — open access — to the ideas and knowledge resulting from our research. Government, too, has an underlying goal of promoting the dissemination of knowledge, expressed in Thomas Jefferson's view that "by far the most important bill in our whole code is that for the diffusion of knowledge among the people."[18] The federal agencies and science policies that this committee oversees have led to knowledge breakthroughs of the most fundamental sort — in our understanding of the physical universe, in our ability to comprehend fundamental biological processes, and, in my own field, in the revolutionary abilities to transform and transmit information.

Open access policies build on these information technology breakthroughs to maximize the return on the taxpayers' enormous investment in that research, and magnify the usefulness of that research. They bring economic benefits that far exceed the costs. The NIH has shown one

[18]Thomas Jefferson, Letter to George Wythe (13 August, 1786). `http://hdl.loc.gov/loc.mss/mtj.mtjbib002184`

successful model, which could be replicated at other funding agencies, as envisioned in the recently re-introduced bipartisan Federal Research Public Access Act (FRPAA).

Providing open access to the publicly-funded research literature — amplifying the "diffusion of knowledge" — will benefit researchers, taxpayers, and every person who gains from new medicines, new technologies, new jobs, and new solutions to longstanding problems of every kind.