The Digital Public Library of America: The Idea and Its Implementation

Robert Darnton
Harvard University, rdarnton@hulmail.harvard.edu

Robert Darnton, "The Digital Public Library of America: The Idea and Its Implementation" (2011).
Proceedings of the Charleston Library Conference.
http://dx.doi.org/10.5703/1288284314873

The Digital Public Library of America: The Idea and Its Implementation

Robert Darnton, Professor and Director, H. Pforzheimer University, Harvard University Library

In a famous letter of 1813, Thomas Jefferson compared the spread of ideas to the way people light one candle from another: "He who receives an idea from me, receives instruction himself without lessening mine; as he who lights his taper [candle] at mine receives light without darkening me."

The eighteenth-century ideal of spreading light may seem archaic today, but it can acquire a twenty-first century luster if one associates it with the Internet, which multiplies messages at virtually no cost. And if Internet enthusiasm sounds suspiciously idealistic, one can extend the chain of associations to a key concept of modern economics—that of a public good. Public goods such as clean air, efficient roads, hygienic sewage disposal, and adequate schooling benefit the entire citizenry, and one citizen's benefit does not diminish that of another. Public goods are not assets in a zero-sum game, but they do carry costs—up-front costs, usually paid through taxation, at the production end of the services and facilities that the public enjoys as users. The Jeffersonian ideal of access to knowledge as a public good does not mean that knowledge is costless. We enjoy freedom of information, but information is not free. Someone had to pay for Jefferson's taper.

I stress that point, because I want to offer a work-in-progress report on the Digital Public Library of America (DPLA) and to argue that it is a feasible, affordable project as well as an opportunity to realize the Enlightenment ideals on which our country was founded.

Although fantasies about a mega-meta-macro library go back to the ancients, the possibility of actually constructing one is recent. It dates from the creation of the Internet (1974) and the web (1991). Google demonstrated that the new technology could be harnessed to create a new kind of library, one that, in principle, could contain all the books in existence. But Google Book Search is a story of a good idea gone bad. As first conceived, it promised to do what Google did best: searching for pertinent information. Google would digitize millions of books provided for free from research libraries, and users would be able to locate material in them by entering key words and examining short snippets called up from the database. Google would not produce the texts of the books, and it might even indicate where they could be found in the nearest library. But because most of the books were covered by copyright, the Authors Guild and the Association of American Publishers (AAP) brought suit for alleged infringement of their intellectual property in *Authors Guild v. Google*. Google could have defended itself by invoking the doctrine of fair use—tricky business, to be sure, because it hangs on arguments based on sections 107 and 108 of the 1976 copyright act, whose obscurities and ambiguities have occupied lawyers for decades. But Google could have hired the best lawyers in the country to make a convincing case. If it won, it would have scored a double victory for the public good: It would have promoted the accessibility of literature and established a broad and firm legal basis for the fair use of that literature.

Instead, Google chose the path of commercialization. After three years of secret negotiations with the plaintiffs, it reached a settlement with them which transformed the original search operation into a speculation based on the data base of books. Access to the texts of the books would be sold back to libraries, including the libraries that had originally provided them free of charge, for an annual subscription fee, which would be set by representatives of the authors and publishers along with Google. Free of pressure from competition and from oversight by any public body, the cost of the subscription could escalate as disastrously as the price of academic journals has risen in the last two decades. The settlement therefore came down to an agreement about how to divide a pie: 37% of the profits would go to Google and 63% would go to the Authors Guild and the AAP.

The settlement had to be accepted by a federal court, because it involved a class action suit, and a

DOI: http://dx/doi.org/10.5703/1288284314873

judge had to verify that the Authors Guild and the AAP represented authors and publishers in general. The Guild has only 8,000 members, but several hundred thousand Americans have published at least one book, and 6,800 authors had taken advantage of an opt-out clause in the settlement by notifying Google that they did not want to participate in its enterprise. Conflicting interests made it difficult to believe that the plaintiffs spoke for any coherent class. Judge Denny Chin of the Southern Federal District Court of New York therefore rejected the settlement in a decision announced on March 23, 2011. He also emphasized other, equally strong objections to it, including the fact that it threatened to constitute a monopoly and that it would give Google exclusive control over orphan works—that is, books whose copyright owners have not been identified. So far, Google and the plaintiffs have failed to rework the settlement in a way that would make it acceptable to the court. At a hearing on September 15, Judge Chin set a trial schedule for the resumption of the original suit, which would extend proceedings until next July. The publishers have indicated that they might reach a separate settlement with Google, but the Authors Guild appears to be less ready to compromise. And on August 17, a parallel class-action suit over copyright, which involved a group of freelance writers, also failed to get clearance from another court in New York. The legal obstacles therefore seem formidable. It may be too early to declare Google Book Search dead, but I do not see how it can be revived.

Whatever the fate of Google's attempt to commercialize access to digitized books, the time has come to relight Jefferson's taper. We now have it in our power to create a digital library that will make our cultural heritage available, free of charge, to all Americans—and to the entire world.

On October 1, 2010, a group of librarians, foundation heads, and computer scientists met at Harvard to discuss the possibility of constructing a Digital Public Library of America. The basic idea was simple: form a coalition of foundations to provide the funding; form a coalition of libraries to supply the books. But the task is enormously complex. After taking its measure, the group formed a steering committee to provide general guidance and to recruit support from diverse constituencies scattered around the country. A secretariat was appointed and set to work with the help of a grant from the Sloan Foundation to organize study of the most difficult questions. Six working groups produced reports, which cleared the way for a master plan. A preliminary version of the plan was presented to the public on October 21st at a meeting in Washington hosted by the National Archives with the support of the Library of Congress, the National Endowment for the Humanities, and the Institute of Museum and Library Services. By now, therefore, it is possible to have a clear view, or at least a preview, of the DPLA's most important features. Here are some thoughts—my own, not those of the steering committee—about five of them.

1. Scope and content. The DPLA will not draw on one gigantic data base. It will be a distributed system, which will aggregate collections from many research libraries, museums, and other institutions. It will provide one-click access to documents in many formats, including images, recordings, and videos. At first, however, it will consist primarily of books, books in the public domain. Google digitized about two million of them, and copies of its digital files have been deposited at HathiTrust, a digital repository set up in Michigan to preserve the output of Google's digitizing. The Internet Archive—a not-for-profit, open-access digitizing operation founded by Brewster Kahle—also can make available millions of files. Research libraries everywhere have digitized great swaths of their special collections independently of Google. For example, Harvard has digitized and made freely accessible 2.3 million pages of public-domain material for its Open Collections Program, and it is cooperating with China in a program to digitize 51,500 rare Chinese works from its Yenching Library. Government sources are particularly rich. All fifty states have digitized most of their newspaper archives, and their holdings have been aggregated by the Library of Congress, which has offered to make this great trove of information available to the DPLA. By combining these and other sources, the DPLA can lay a foundation of incomparable depth and breadth.

Unfortunately, copyright laws prevent the public domain from extending beyond 1923. Most twentieth-century literature will therefore remain out of bounds for the DPLA, unless some legal way can be

found to include it. And even assuming that copyright could be adjusted, where should the boundary be drawn? Some argue that the DPLA's holdings should stretch right up to the present, provided that an agreement can be reached to compensate rights holders. Were that possible, the DPLA would become a truly "public" library for the entire country. But it also might alienate the public libraries that already exist, because local authorities could cut the funding for their libraries on the erroneous pretext that the DPLA will provide their basic material. For my part, I think the DPLA's mission should be defined in a manner that would make its services clearly distinct from those of existing public libraries. It should leave them to supply their users with current material—whether best-selling novels or magazines or dvds—and supplement that function by providing free access to the general corpus of books that constitute the world's literary heritage. Where then would its collections stop? Most books go out of print with astonishing rapidity. In fact, if they make it into book stores (most don't), their shelf life is a matter of days [info here?]; and few of them continue to sell, even as e-books, after a year. I suggest that the DPLA exclude everything published within the last five or ten years, and that a moving wall, which would advance a year at a time, keep it from interfering in the current market.

2. Costs. When the DPLA opens as expected in 2013, it probably will contain only a basic stock of public domain works and special collections furnished at a minimal cost by research libraries. From that point forward, it will grow as fast as funding permits, but its initial expenses will be devoted for the most part to the creation of its technical architecture and administration. It will be designed in a way that will make it interoperable with major digital libraries in other countries. In fact, it has already reached an agreement to cooperate with Europeana, the pan-European digital library that aggregates collections from 27 countries. Europeana now runs on a budget of 5 million euros a year, but it does not become directly involved in digitization, collection development, or preservation. Therefore, the example of Europeana suggests the bare minimum of what it will cost to get the DPLA up and running.

What would it cost if the DPLA led a major effort to digitize books that are covered by copyright but out of print, assuming there were no legal impediments? Brewster Kahle, who has digitized more than a million works for his Internet Archive, says he can digitize a book for ten cents a page or $30 for an ordinary work of about 300 pages, and he estimates that he could digitize the entire contents of a great library—one with 10 million volumes, somewhat larger than that of Princeton and smaller than Yale's—for $300 million. Other experts find those costs too low. They consider a dollar a page as a conservative estimate; and they note that, aside from the scanning, a great deal of work must be done to perfect the metadata and to assure preservation, not to mention other possible services such as curation and the development of apps. But the costs of digitization and preservation are decreasing, and the technology is improving. The DPLA will begin with a base of several million volumes, and it will grow incrementally by digitizing at a rate that conforms to its budget. What will that budget be? No one knows until a final model is perfected sometime before April 2013. By combining ball-park and back-of-the-envelope estimates, one could imagine digitizing a million books a year on an annual budget of $75-100 million. (The budget of the Library of Congress in fiscal 2010 came to $684.3 million.) If a grand coalition of foundations contributed $100 million a year, a great library would exist within a decade. Double that rate, and the library soon would be the greatest that ever existed. But we needn't rush. We must do the job right, because the DPLA should last for centuries, and it could grow gradually on a budget of $5-10 million a year.

3. Legal issues. The DPLA must respect copyright. How far it can go in making accessible books that are out of print but covered by copyright depends on the interpretation of copyright laws by the courts and the possibility of modifying them by Congressional action. The history of copyright in the United States goes back to article one, section eight, clause eight of the Constitution, which sets two objectives: "to promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries." The first copyright law, passed in 1790, struck a balance between those objectives by giving authors an exclusive right to the sale of their work for fourteen years, renewable once. At that time, Jefferson's taper was burn-

ing bright, and American statesmen took heed of British precedent. Parliament had adopted the 14/28 year limit in the original copyright law of 1710. Claims for perpetual copyright had been debated in a series of court cases until they were definitively rejected in the great decision of *Donaldson v. Becket* in 1774. During the debate over the Sonny Bono Copyright Term Extension Act of 1998, Jack Valenti, the head of the Hollywood lobby, was asked how long he thought copyrights should last if they could not be perpetual. "Forever minus one day," he replied. Since then, the flame of Jefferson's taper has nearly died out.

The current limit of copyright—the life of the author plus seventy years—tips the balance decisively in favor of private interests at the expense of public good. The public domain extends only to 1923. Every book published after 1963 is now covered by copyright, whether or not its copyright has been renewed, according to Congressional acts of 1976, 1992, and 1998. The status of many books published between 1923 and 1964 remains ambiguous, because at that time copyrights had to be renewed, and the record of renewals does not leave a clear trail leading to the copyright holders today, if any have survived. Hence, the problem of orphans.

Further legislation could solve the problem. But lobbyists had such a heavy hand in attempts to pass orphan book legislation in 2006 and 2008 that some consider it impossible to redress the balance of copyright law in a way that would "promote the progress of science and useful arts." The only recourse may be to sections 107 and 108 of the copyright law of 1976, which, as mentioned, opens the way for the "fair use" of copyrighted materials. Unfortunately, that way passes over some very uncertain terrain (a Section 108 Study Group composed of librarians and lawyers worked through the problems for two years and came up with some proposals but nothing that has had any effect). Fair use normally applies to non-commercial activities such as criticism, scholarship, and teaching. Google's original, search-and-snippets enterprise involved advertisements intended to bring in revenue for a profit-minded business. By contrast, the DPLA will be a not-for-profit association dedicated to the public good, and therefore it might stand a better chance with a fair-use defense, in case it should be sued by owners of rights to books that it had digitized in the mistaken belief that they were orphans. But should the DPLA run such a risk? Probably not. Orphan book legislation might provide immunity from litigation and set up an escrow fund to compensate rights holders of books that had been treated as orphans. And if Congressional action really is hopeless, the DPLA could try to reach an agreement with authors and publishers whose copyrighted books have gone out of print. Google had attempted to do so in the settlement, which included an opt-out default: all authors were deemed to have accepted the terms of the settlement unless they notified Google to the contrary. This aspect of the case especially troubled Judge Chin, because it seemed to give Google monopolistic control over the entire body of orphan books—and there are likely more than a million of them. Could an opt-out provision pass muster if it were applied for the benefit of the public by a not-for-profit organization?

Again, the answer is probably no. But a solution might be found in legal arrangements known as extended collective licenses (ECL), which have been successfully developed in Scandinavian countries. In Norway, a broad-based association of authors allied with publishers has developed an ECL that represents the interests of all copyright owners in a program to digitize and make accessible, free of charge, all Norwegian books to readers located in Norway. The rights holders will be compensated from a fund according to a fixed fee per page of use by readers, who can consult the texts on their screens but not download them, and authors can opt out of the system. In some respects—the creation of a "class" that represents all authors and the opt-out default—the Norwegian program resembles Google Book Search, except that it was authorized by legislation and is subject to government oversight.

Of course, the United States has little experience with collective management of rights—although the Copyright Clearance Center and JSTOR might provide models—and America's culture is much less homogeneous than Norway's. The Authors Guild may refuse to yield an inch in defending the interests of professional authors. But most authors probably would prefer to have digitized versions of their out-of-print books made available for a small fee or even for free, rather than leaving them to

languish unread on the shelves of a few libraries. Above all, authors want readers, and the minority of authors who live from their pens could opt out of this arrangement. Some of the best legal minds are now developing plans for an American ECL regime, which would make it possible for our digital library to include everything that was published in the twentieth century.

4. Technical architecture. Last June the steering committee of the DPLA opened an international "Beta Sprint" competition for the best pilot projects, tools, and tentative blueprints of the infrastructure that will hold the system together and make it operate seamlessly for users. More than 60 potential applicants expressed interest. Nearly 40 of them submitted projects by the deadline of September 1. A panel of experts from around the country selected the six most promising projects, and the six were presented to the public at the general meeting in Washington on October 20th and 21st. The technical subcommittee of the DPLA will oversee the effort to cull and combine the best ideas of the winners and to come up with a draft prototype by April 2012. The prototype will be perfected during the next six months, and it should be ready to go into operation when the DPLA is launched in April 2013.

The race to this deadline may seem breathtaking, but it is fueled by enthusiasm and energy. Leading figures in computer science, information technology, and library science have assured us that the task is do-able, and we will get it done.

5. Governance. I have arrived at the last of my five topics, and here I must be brief, because the governance committee of the DPLA has only begun to study the possibilities for administering it after it is launched a year and a half from now. Where should it be located? Who should lead it? To whom should it be responsible? How will it formulate policy and administer its services? The present secretariat under the able leadership of John Palfrey of the Harvard Law School Library will continue to direct affairs during the final eighteen months of the embryonic DPLA's existence. By April 2013, the newly born DPLA will have set up headquarters—probably at a considerable distance from Harvard. The Harvard phase of its existence had to do with its original conception by a group of self-appointed enthusiasts. The mature DPLA will belong to the entire country and will serve a broad constituency, including ordinary readers, independent researchers, the multi-faceted public of public libraries, K-12 school children, students in community colleges, university students and faculty, and book lovers of every stripe. In order to fulfill its broad mission, the DPLA will be responsible to a board of trustees representing a wide variety of interests. It will need a staff of professionals and, no doubt, a director with plenty of expertise and energy. It might become absorbed in an NGO that has a strong record of excellence in library affairs, or it could operate as an independent corporation by taking advantage of section 501(c)(3) of the Internal Revenue Code, which favors nonprofit organizations. At present, most people think it should not be part of the federal government so that it will be free from political pressures. It might resemble the National Academy of Sciences or the BBC.

In fact, however, it won't resemble anything, because nothing like it has ever existed. A library without walls that will extend everywhere and contain nearly everything available in the walled-in repositories of human culture... E pluribus unim! Jefferson would have loved it.