

How Free Culture Will Save Digital Libraries

Aaron Krowne (Emory University)

July 23, 2005

Abstract

Today, we are watching as web search engines—especially Google—push libraries further down the service-provision hierarchy, towards roles as anonymous information-silos, and therefore diminished importance. The production of free culture in a digital library context will reverse this trend, making digital libraries themselves compelling, useful, intentionally-visited virtual places, and ensuring the continuing importance of libraries in the future. We will see that this outcome largely follows from digital libraries doing what we already expect libraries to do—but translated into a digital context, where forces of commons-based peer production (CBPP) operate in a milieu of free culture.

Note: This is a preprint of an article to appear in The Symposium on The Free Culture and the Digital Library to be held at Emory University, October 14, 2005. <http://www.metascholar.org/events/2005/freeculture/>.

1 Introduction

A few years ago, as a graduate student mining my advisor’s lengthy CV for interesting publications, I came across one called “How Digital Libraries Will Save Civilization” (Fox et al., 1998). While I had access to nothing more than the citation string of this article until recently, the title struck me as embodying a wonderful sort of sunny optimism, which—being someone working in the field of digital libraries—I found very motivating. And still do.

Nothing in the Fox article attempts to rigorously substantiate the claim implicit in the title. However, the ideas in the paper, when projected to their greatest-possible impact, could realistically attain this end. The possibilities have tantalized technologists and dreamers as far back as Vannevar Bush (Bush, 1945).

The current paper is written in the spirit of that article—and as sort of a prelude to it. While I won't attempt to seriously substantiate the implicit claim of the title here, I hope that my optimism for free culture and commons-based peer production (CBPP) will similarly inspire others. And perhaps the outcome will be that the digital library will persevere as a socially-significant institution, and maybe—through bringing people together to build our common culture of knowledge—actually save civilization.

1.1 A Note on “Digital Libraries”

It will be useful for this paper to establish what, exactly, I mean when I say “digital libraries” (DLs). In the present context, I am not speaking about systems which have basically all the functions and functionality of what are formally defined to be digital libraries (e.g., with the 5S framework, introduced in Gonçalves et al. (2004)). Indeed, under such a definition, the current incarnation of Google even qualifies as a digital library—as well perhaps it should. Instead, however, I mean systems that self-consciously call themselves “digital libraries,” and which are likely to be supported in a top-down fashion, from large, official, often national funding organizations, or as outgrowths of physical libraries.

These systems, while inarguably important, tend to be governed by different forces than commercial or grassroots public projects. And it is largely to them I direct this lecture, out of a concern for seeing them maximize their potential, and deliver the most value to the general public—from which their support and lifeblood ultimately derives.

I should also point out here that this article will to a great extent conflate classical libraries and digital libraries. This is largely unavoidable for two reasons: the first being that many of the trends and principles discussed here affect both kinds of libraries, and the second being that classical libraries are increasingly *becoming* digital libraries. The latter is evidenced by a number of trends, including web-based OPACs that basically behave like digital library search interfaces, and an increasing number of all-electronic holdings.

The reader should exercise common sense as to which kind of library I mean in particular passages, whether or not cited unambiguously.

2 The “Threat” to Digital Libraries

In recent years, the chagrin and frustration of librarians has been raised by web search engine Google (e.g., Gorman (2004)). The problem is that library patrons are using web search engines, especially Google, as the primary means of meeting their research needs (Lippincott and Kyrrillidou, 2004). The reality is that almost every research undertaking starts at Google—whether or not it ends up at a digital library record or a physical library. Librarians want patrons to use their electronic library catalog search systems, their specialized domain database search interfaces, and their digital library search and browse services. But patrons are staying away in droves.

To add insult to injury, if a search starts on the internet, there’s a good chance the research task will end there, completely bypassing the library or digital library. But even when the research task continues on to some electronic library interface, it is usually because it is *mostly done*. The user knows what “record” or records they want; the role of the library is simply to “serve it up.” This is a significantly diminished role for libraries, compared to their historical purpose and present aspirations.

A major part of the explanation for this is that web search engines have essentially become universal metasearch engines. These search engines are not limited by mere “kinds” of records; if its on the web and the typical web surfer can see it, the search engine can get them to it. Google has provided the template for how this is to be done “right:” a single search text box, in which the user can enter keywords (even natural language is safe to use), with a minimalistic interface design, free for all to use, supported by low-key text advertisements, and providing instantaneous response. By contrast, library search engines are still all-too-often fragmented over disparate databases, utilize legacy, Boolean-based OPAC query syntax, have cluttered and unclear interfaces, and are slow to boot.

3 Responses from the Digital Library World

3.1 The Technical Approach

Libraries and DLs are beginning to respond with their own metasearch solutions. But these may be a day late and a dollar short. From the user's perspective, it is difficult to determine why they should go first to their local library's metasearch interface (possibly one of many local libraries) when they can simply "google it"—and probably find all the information they could possibly need.

Seeming to realize this reality, in the past few years, the digital library community has been working to find ways to make DLs important and useful to end users. Digital library researchers have even embraced Google as a way to get users in "through the back door," as services such as DP9¹ illustrate (Liu et al., 2002). But most effort has been in digital library architecture—in essence coming up with reasons for users to show up (or at least, return) through "the front door." This effort has been focused on two broad technical categories: (1) exploiting domain specificity, (2) and metadata-based services.²

Domain specificity refers to building digital libraries *for* X, where X is some subject. Metadata-based services refers to providing ways of retrieving and organizing information that web search engines can't, because the digital library has access to (and "understands") richer metadata. These two categories are not unrelated, e.g., often most of the advantage in building domain-specific digital libraries is in the fact that such a library can support metadata elements which are unique to the domain. But this need not be the case.

Much progress has been made along both of these technical fronts.³ Surely they are a part of the solution. However, it is unclear that these efforts alone are paying off in significant "front door" use of digital libraries.

¹DP9 makes Open Archives repository records automatically accessible as web pages. An immediate consequences of this is that web search engines can crawl and index them. See <http://dlib.cs.odu.edu/dp9/>.

²One needs only to peruse the DLI-1 and DLI-2 initiative projects or the proceedings of JCDL in the past few years to get a feel for this trend.

³For example, see our MetaCombine (<http://www.metacombine.org/>) and OCKHAM (<http://www.ockham.org/>) efforts. These projects seek to more meaningfully combine digital library resources and services, and to methodically propagate these services throughout the library world.

One problem is that digital library researchers and developers have an incredibly tough task: they must make these value-added technical services so compelling, so convenient, and so easy to use, that they will be as attractive and high-priority to users as Google. But there is an inherent conflict here: there are many digital libraries, yet as a universal meta-search system, Google precisely *is* a machine for avoiding the repetition of searches through many interfaces. The problem is not just convenience, but the “economics” of attention and efficiency.

A second problem is that the technical route to shoring up digital libraries implicitly expects to “out-innovate” Google in the areas of information retrieval and information integration in general. One could call this a technical “arms race” with Google—but Google has such a war-chest of resources and talent that it doesn’t actually seem to be facing any serious challenge (from the most established of tech companies, let alone digital libraries—e.g., Vogelstein (2005)).

Indeed, Google gets mentioned by name in this paper, by librarians, and even in vernacular verbiage, because it cannot be treated the same as “other” internet search services. Google says, ever ambitious, that its mission is “to organize the world’s information,” and it has proven it is very serious about this goal. The problem is that this is basically the mission of libraries too, so almost every step of progress Google makes toward this goal seems to shrink the universe of importance of libraries (*especially* DLs).

For example:

- The launch of Google Scholar⁴ in the past few months shows that Google is acting more like a universal metasearch service than any other entity. The function of this service is to provide searching of scholarly materials—from library and digital library catalogs—through Google’s interface. If the scholarly materials don’t happen to be available for free online, then the user, properly authenticated, can access it through a library interface. Hence, the library’s role here: filling in the blanks at the final step of the research process.
- The library world is also watching in slow motion as Google undertakes its audacious book digitization effort. This venture demonstrates the power of Google’s clout—not only can they fund such a massive effort, but they are completely unafraid of the inevitable copyright

⁴See <http://scholar.google.com/>.

headaches. Solutions will probably have to be found purely to accommodate Google.

There are other reasons to have pause about pursuing technical “competition” with Google. Google Local and some binding with the Open Directory web categorization hierarchy shows that the company can indeed provide domain-specific information services when it wants—even over a generalized web search. Binding with category hierarchies could be even tighter, with minor additional engineering. Link analysis could be used to automatically discover and guide searchers to subject domains (Flake et al., 2000, 2002; Gibson et al., 1998; Reddy and Kitsuregawa, 2001). And the nascent treatment of metadata, at which Google Scholar hints, could be extended.

Libraries and digital libraries can of course always rest assured that they will be required elements of the research process, inasmuch as patrons require access to copyrighted items which are distribution-restricted. However, this may have to be in the most pedestrian of roles: doing nothing but delivering records, once found. In fact, libraries may face an “identity-stripping,” whereby tools meant (innocently) to interface web search with digital libraries have the unintended result of almost completely abstracting away and black-boxing the provider of the end artifact (i.e., the library or digital library). This is an inevitable consequence of tools which make research easier for the user, which we can already see prototyped by the WAG localizer (Singer, 2005).⁵

3.2 The Closed-Access Approach

Even if the above is not considered an issue, relying on proprietary digital content as the *raison d’être* of digital libraries still seems to be a risky bet. As the Sabo “Public Access to Science Act,” H.R. 2613 indicates,⁶ the tide of public sentiment is turning against the present practice of letting remain generally-unavailable scholarly content which is produced using public monies. And research libraries rely heavily on this content—mostly expensive paper and electronic journals—for patronage. While H.R. 2613 seems stalled, what happens when a bill in the same spirit finally succeeds, and

⁵The WAG-The-Dog web localizer, from Georgia Tech, smoothly integrates library and digital library holdings with Google Scholar and other web sites. It is available at <http://rsinger.library.gatech.edu/localizer/localizer.html>.

⁶See <http://thomas.loc.gov/cgi-bin/query/z?c108:H.R.2613:>.

publicly-funded research is unshackled from restricted distribution? When one considers that there is very little private scholarly research left, and on top of this the fact that this kind of law would necessarily force blanket open access policies on most journals, it is difficult to see which scholarly content could legally remain closed.

There are other problems as well. As pointed out in (Regazzi, 2004),

The early 1970s was a time when, for the most part, research libraries could buy all new research material, thus keeping up with virtually all R&D developments. But for the 20-year period from 1975 to 1995, university library expenditures increased only at the rate of 2.2%, ... while research and development spending increased by 4.6%, nearly double that of the library. The result is a huge gap in the university library's ability to keep up with the production of research and development.

Exacerbating this, as also reported in (Regazzi, 2004), library expenditures as a measure of total university spending have decreased from 3.7% to 2.8% annually since 1982. Thus, there is more to buy, and less money being allocated to buy it. This has resulted in a situation where it is becoming less likely an individual will be able to access a given journal article, because their member institution probably doesn't have a subscription to it (or perhaps not the *right* subscription).⁷

Thus, it seems closed-access is unwanted, inefficient, and a poor way for libraries and DLs to make a meaningful identity for themselves. Table 1 speaks to these points: collaborative (CBPP) digital libraries are compared with open access but strictly-controlled digital libraries, as well as one completely closed one. According to this data, it appears that closed libraries such as ACM DL are being seriously challenged by open, collaborative ones, such as CiteSeer.

Perhaps more important than the above reasons to avoid putting all of the DL "eggs" in the proprietary-content "basket," I want to suggest that relying on these materials to give DLs a purpose runs completely counter to the philosophy of libraries. This philosophy is one of disseminating knowledge as widely as possible and furthering scholarly activities. I would like to put

⁷The "solution" of inter-library loan for this problem still incurs a significant convenience hit. It also seems a bit convoluted, given the already-digitized state of most of the inaccessible articles.

forth here that the way out of this narrow and possibly terminal future for digital libraries is to embrace free culture, and in doing so, embrace their true calling.

4 In Free Culture, a Better Solution

That a fuzzy, social concept like “culture” is part of the solution I am proposing signals a radical departure from the extant, technical attempts at solutions outlined in part above. The cue that culture is a part of the solution comes from perhaps the least likely of places one would expect to find inspiration for surviving in the digital age: classical, brick-and-mortar libraries.

The key fact about classical libraries is that they are not seen as, or used as, information retrieval *machines*. They are seen as social and cultural *places*. People go there not only to retrieve information in the form of books, but to study it, to conduct work derived from the knowledge these books contain, to discuss with others the ramifications of what they are reading and researching, or to interact with the library staff to help give direction to their research activities. In short, they go to act in a scholastic way, in a social context, with peers and experts.

This notion has been all but lost in digital libraries—or at least, systems that self-consciously call themselves digital libraries (as distinguished above). I will argue in this paper that replacing this “lost” social notion is the key to adding compelling value to digital libraries, and sustaining them as a useful, meaningful institution.

As someone who works in a library, I routinely observe that people will come to physical libraries to act studiously and scholastically, despite alternatives that let them stay at home—because of the special and social nature of the space. This continuing fact is reported in (Lippincott and Kyrillidou, 2004). In fact, for all intents and purposes, Starbucks and Borders benefit from the same phenomenon. The practice of providing a social atmosphere for intellectual activities seems to be alive and well, and if anything, growing.

The key, then, to “saving” digital libraries is to similarly re-establish a notion of a social place—within the context of the digital library. In essence, this allows the patron to undertake intellectual, cultural activities, resulting in the actual creation of culture.

4.1 Free Culture, CBPP, and Digital Libraries

For this article, I define free culture as the social milieu of information artifacts which may be disseminated and modified without permission (*libre* free), for which there is also zero structural monetary cost to do so (*gratis* free).⁸ Note that open access is necessary but not sufficient for free culture; it provides the *gratis*, but may lack the *libre* component.

The tie-in of free culture to digital libraries as social, culture-producing places is for two main reasons. The first is that, in the digital context, all interaction is potentially subject to copyright restrictions. That is, every communication is an artifact created, and the dissemination of such cultural communications by the digital library is generally reduced or ruled-out entirely under the current, permission-default copyright regime (Lessig, 2004). I will talk more about this in a later section.

The second reason, which is the focus of this paper, is that free culture both enables and is produced by commons-based peer production (CBPP). CBPP is the name given to the distinct mode of production of intellectual artifacts which has emerged in the past few years, enabled by the internet and the appropriate software layer on top of it. The GNU/Linux operating system and Wikipedia are prominent examples of this phenomenon. CBPP is considered a mode of production as distinct, and perhaps important, as markets or firms. For more on CBPP see (Benkler, 2002) and (Iannacci and Mitleton-kelly, 2005).⁹

Making intellectual artifacts free (in both the *libre* and *gratis* senses) enables CBPP, which then produces more free intellectual artifacts. The upshot of this feedback loop is a powerful economics of intellectual production, which we will explore in detail later. A simple illustration of this feedback

⁸By *structural* I mean costs which are built into *access*. Carrying costs of dissemination (such as paying for media and handling) do not count. The litmus test would be: if the distributor or seller of the work has a legitimate complaint in preventing do-it-yourself copying or dissemination of the work, then it is not truly free of structural dissemination costs, and is hence not *gratis* free.

⁹CBPP can actually be used in a closed fashion, e.g., for internal corporate collaboration solutions. Applied in such situations, it is still an innovative approach, due to its flattening of productive hierarchies. Thus, the reader should not get the impression that CBPP cannot be used for more narrow communities and with more limitations on permitted activities. For a more technical and generic treatment of CBPP frameworks, see (Corneli and Krowne, 2005) in this volume. For this article, however, I deal with the more widely-open sense of CBPP.

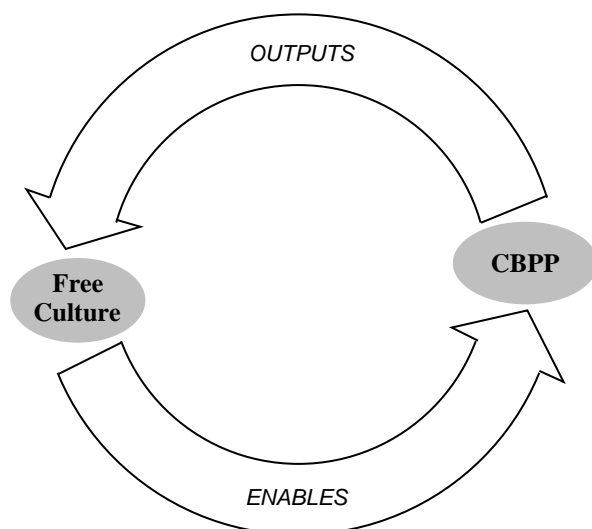


Figure 1: The relationship between free culture and commons-based peer production (CBPP).

loop is shown in Figure 1.

In Figure 2, the “pillars” which support free culture are shown. These pillars are not identical with CBPP—they are “made out of” not only the technical elements of by CBPP systems, but also social protocols within and above these systems, and copyright law and licensing which permit the necessary productive activities to take place. This figure also shows how production of free culture progresses “through” the pillars, then feeds back on itself (as in Figure 1).

I contend that free culture is so compelling that users will be drawn to virtual places that allow them to work with it, use it, manipulate it, and bring it into the context of their lives and their interests. By integrating CBPP services, digital libraries can become “engines” of free culture, crossing a threshold where their *services* and their *communities* become more important than their collections. Such a transformation undermines the deleterious effects (to DLs) of the recent Google-fueled trend toward commoditization of collections I described earlier.

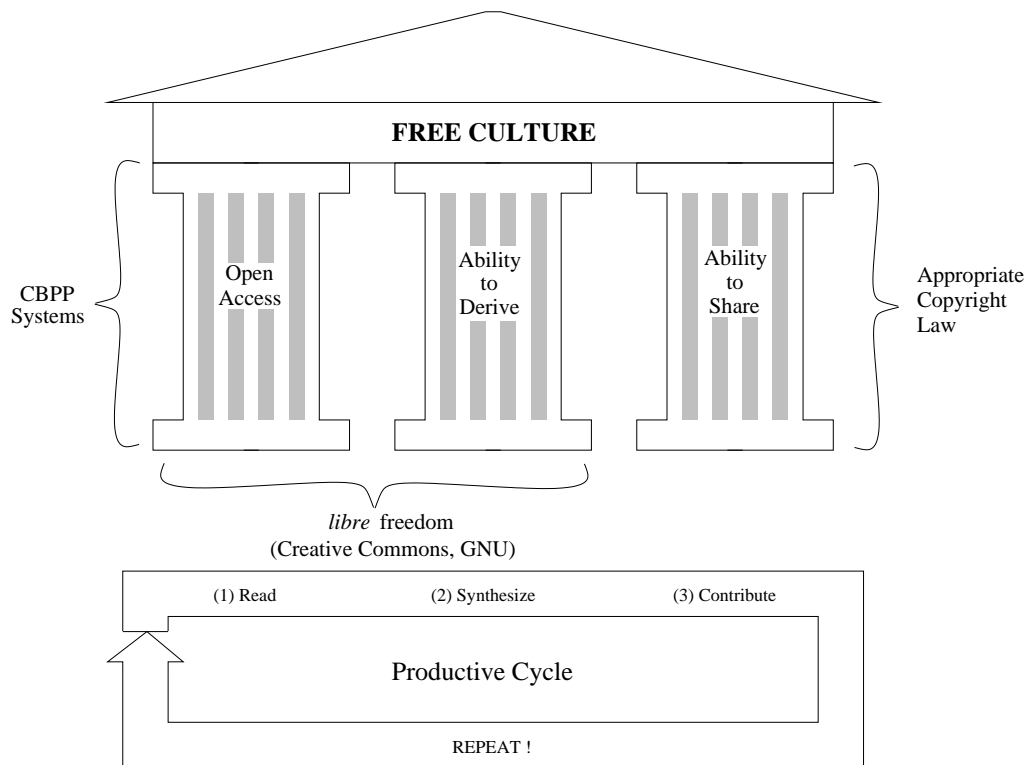


Figure 2: The “pillars” which support free culture. Each pillar is enabled both by CBPP systems and social conventions, as well as the proper copyright and license law.

4.2 Exemplary Information Systems

In this section I give a brief tour of digital library and information systems which integrate CBPP, either completely or to a large extent. The intent here is to demonstrate that CBPP can create or enrich this type of system, making them widely-used, engaging, useful, and sustainable.

- *Wikipedia* - <http://en.wikipedia.org/> - Wikipedia is a collaborative, general-information encyclopedia. It is made up of articles which are created and improved by volunteers (who can be anyone on the internet—a login is not even required to make edits). The English edition of Wikipedia, which is the largest, has over 600,000 articles. In essence, Wikipedia “scales up” the encyclopedia concept, distributing the production work, adding hyperlinking, removing limitations on topical coverage, and adding the typical benefits of digitization and internet-accessibility.
- *ArXiv* - <http://www.arxiv.org/> - The ArXiv was started at the beginning of the 90s by Paul Ginsparg, while he was at Los Alamos National Lab, as a pre-print server for physics research. This service allowed physicists to share each other’s work before journal publication, allowing more extensive feedback to be received earlier, bypassing the lengthy delays in the publication process (sometimes measured in years), and avoiding the suppression of legitimate research.¹⁰ Today, ArXiv (and Ginsparg) are at Cornell University, and the subject coverage of the service has spread into mathematics, computer science, and areas of chemistry and biology. In physics, the service is almost ubiquitous, with nearly every journal article (and then some) having an arxiv.org incarnation. Articles are often simply cited by their ArXiv URL.
- *PlanetMath* - <http://planetmath.org/> - PlanetMath is a mathematics community featuring a collaborative mathematics “encyclopedia” contributed by volunteers. It was started by myself and Nathan Egge, shortly before I began grad school (in 2001).¹¹ The concept is similar to

¹⁰Due to space constraints or “political” feuding.

¹¹Significantly, PlanetMath was started when a similar, earlier resource went offline due to being *gratis* free but not *libre* free. Therefore, the distinction is important, and this illustrates how free culture will fail to flourish without both senses of “free.”

Wikipedia, except that mathematics is taken as the focus, the authoring language is L^AT_EX (the de facto standard among the mathematics community), linking is automated, and a more academic authorship model is utilized (accounts are needed to edit and articles have *owners* who exert a high level of creative control over them).

- *Slashdot* - <http://slashdot.org/> - Slashdot is a news web log (or “blog”)—in fact one of the first (dating back to 1996). It chiefly serves the “geek” community—those interested in science and technology. News stories are submitted by the general public, vetted by editorial staff, and posted. However, the real “magic” begins after this point, as each article is also a discussion area. Hundreds or thousands of people comment on the typical article, and comments are scored and filtered collaboratively with Slashdot’s “karma” system. The result is that funny and insightful comments “float” to the top. Slashdot in essence filters news stories through the “hive mind,” often leading to surprisingly original and penetrating analyses the mainstream media tends to miss.
- *CiteSeer* - <http://citeseer.ist.psu.edu/> - CiteSeer is an autonomous scientific literature digital library. It is built through web crawlers (or “spiders”) which traverse the web looking for documents likely to be research papers (i.e. PDFs, PostScript files, etc.). Those that pass some basic machine learning filters are given metadata (also through machine learning—see Han et al. (2003)), posted, and interlinked with the rest of the collection. Although this digital library is automated, it is collaborative in the sense that contributors “post” to it simply by posting their research on the web (for instance, on their home pages, or at university technical reports archives). Further, users can suggest new locations for the system to crawl (and in essence, manually add papers), rate articles, and correct metadata. However, the most popular and effective services by far are the central acquisition, listing, and integration services.

How well are these largely-collaborative information systems doing? Some examples follow. As pointed out in Table 1, CiteSeer is more popular than its closed counterpart, ACM DL, at least in the lens of PageRank. ArXiv.org is just as popular as APS’s site, which hosts the massive complex of journals run by the American Physical Society. Slashdot is such an 800-pound gorilla

CBPP DL		Non-CBPP DL	
CiteSeer	8	7	ACM DL ^a
Wikipedia ^b	9	9	Britannica
PlanetMath	6	8	MathWorld ^c
ArXiv.org	9	9	APS

^aThe PageRank of the IEEE Computer Society’s DL was unavailable, possibly because of how their web site redirects and mangles URLs.

^bThe english edition of Wikipedia, at en.wikipedia.org, was used.

^cFor perspective, note that MathWorld is about three times the size of PlanetMath (in terms of content), and predates it by at least a decade.

Table 1: **Interesting PageRanks:** CBPP DLs overall score impressively in “PageRank” (Google’s derived metric of “importance” on the internet, see Brin and Page (1998); Page et al. (1998)), generally at least matching their Non-CBPP counterparts. Perhaps tellingly, the only fully closed-access DL on the list, ACM DL, was the only non-CBPP DL to achieve a PageRank strictly *lower* than its CBPP counterpart. Underscoring this, CiteSeer’s Alexa (<http://www.alexa.com/>) traffic rank can be found to be 1,378, versus ACM DL’s 8,591 (higher is more-visited).

in the “geek-o-sphere” that its news posts routinely bring the servers hosting the linked stories to a halt. There is even a verb for this—it is called getting “slashdotted.” Additionally, Wikipedia is clearly incredibly popular by a number of metrics, besides PageRank, as it has an Alexa ranking of 65, and has the distinction of being highlighted separately in Google searches and within Amazon’s A9. There have also been convincing (or at least thought-provoking) benchmark studies and qualitative arguments that Wikipedia is the “best” encyclopedia out there—and it took only four years to build.¹²

4.3 Research Results

Some encouraging formal research results are beginning to be reported, which suggest that making digital libraries more participatory fosters pedagogy, engagement, uptake, utility, and so forth. This goes a significant ways towards studying what happens in a collaborative, free culture information environment.

A few of such reports follow:

- In (Zhang and Quintana, 2005), a system called IdeaKeeper is evaluated. This system gives students a structured way to analyze digital library learning objects. For example, they can give feedback about whether a viewed item was related to their question, enter in the main idea of the presentation, list the supporting evidence, give feedback about bias and expertise, highlight the specific information that answers the initial question, and so forth. The research shows that uptake was very good and that there were significant positive effects on learning among students who used this tool over students who just used the learning objects without IdeaKeeper.
- In (Brusilovsky et al., 2005), the Knowledge Sea system is evaluated. This system focuses on *social navigation*, whereby users of the digital library give feedback (both implicit and explicit) which is used to facilitate the discovery process of others. Utilizing novel visualization techniques, the system provides cues for which resources are most popular,

¹²Strangely, it seems only the Germans have been interested in actually performing scientific studies on the quality of Wikipedia (perhaps McHenry (2004) was never translated to German). Two such studies have been done. See http://meta.wikimedia.org/wiki/Content_reviews for details.

active, and of the highest quality. Annotations, discussions, and ratings are supported and help provide the basis for the corresponding visualized indicators. The system's efficacy was tested with students, and it was found that there was a significant positive correlation between resources indicated as of high quality or of interest and the resources which were most utilized.

- In (Milson and Krowne, 2005) in this volume, we show that CBPP systems can be compatible with the formal education setting, while still yielding the extra benefit of creating useful learning objects. This was done with a small pilot study, whereby the Noösphere system was deployed for classroom support for a small mathematics graduate class. Students were loosely given assignments to produce mathematical articles, and were given credit for activity within the system. From the articles which were written by the students, a set of collaborative course notes were compiled. By conventional evaluation metrics, the performance of the students was found not to have been diminished, and they were additionally exposed to an aspect of scholarly work they otherwise would not have encountered.
- In (Efron and Sizemore, 2003), a pilot study of iBiblio is done. iBiblio is a large, public-access, collaborative archival digital library.¹³ In the study, the authors ask the question “do increased contributor efforts lead to increased collection popularity?” They find a strong answer in the affirmative: the more collaborative maintenance activity there is in a collection, the more popular it becomes.

Results like these should be no surprise, and I believe we will be hearing more like them. The emerging pattern seems to be that free culture through CBPP is beneficial in two ways: (1) for certain types of individuals (more in some settings than others), it turns them from passive consumers of information to engaged constructors of the shared knowledge environment, and (2) everyone benefits from the distribution of teaching and sharing away from small, central, often intellectually-homogeneous knowledge “oligarchies” to anyone who has the ability, motivation, and expertise to teach, help, and share.¹⁴

¹³<http://www.ibiblio.org/>.

¹⁴This can be considered to foster the kind of “sense-making” services that are called for in (Regazzi, 2004).

Much of the work towards these ends comes under the rubric of “personalization” research in the digital library research world. I would encourage the continuation of this thread of work, but would also add that researchers should consider enabling the *sharing* of the effects of personalization services whenever possible (if not making sharing the default). This is because each individual’s “sense-making” of the information in the library produces valuable secondary information which will likely be of use to many other users of the library. Sequences and groups, annotations, ratings, categorization, or even views and activation all provide generally useful information about the collection, even though the individual user may think of them primarily as means for customizing it.

5 How DLs Can Support Free Culture

The descriptions of the exemplary systems and published research above suggest a little bit about how CBPP can be employed by digital libraries. Some types of collaborative services that can be employed are:

- annotations/discussions
- list-making and categorizing (forming associations)
- ratings (for collaborative filtering/recommenders)
- reputations services
- reviews and moderation (i.e. include/exclude judgments)
- content authoring/creation
- correction/enhancement

DLs need not be *entirely* based on CBPP in order to provide compelling free culture value—they can pick and choose from the above menu of services (and surely beyond) to determine their overall constitution of collaborative, automated, and controlled labor.

It is worth briefly mentioning a few more CBPP projects which illustrate creative combinations of the above collaborative services. For example, the

Distributed Proofreaders¹⁵ project implements collaborative correction. Systems like Furl,¹⁶ CiteULike,¹⁷ and Delicio.us¹⁸ implement list-making, categorizing, and annotations.¹⁹ Amazon.com implements list-making, reviews, and ratings. eBay implements ratings and reputations services. Observe that all of these sites are very influential and successful, or at least are “upstarts” making a large splash. In fact, it was by integrating the latent social information of hyperlinks that Google leapfrogged the competition—in effect becoming a collaborative filtering service by exploiting millions of “endorsement” judgments on the web.

There is almost certainly a niche for DLs which have a “static” content base (either “frozen” due to being historical or slow-changing due to being centrally-vetted) yet which include a high-impact CBPP component. Figure 3 shows how CBPP services can still be layered on top of such a content base, being bound to the underlying artifacts via identifiers and standard internet interface conventions such as links and frames.

6 Roadblocks and Challenges

6.1 Perceptions of CBPP

There is an immense amount of consternation in certain quarters about the prospect of CBPP becoming a major force in the production of our information landscape (Biss, 2004; Gorman, 2005; McHenry, 2004). This, I think, is ultimately rooted in fear.

Popular culture in the west for the past century has not traditionally been free—nearly all aspects of it have been strictly controlled by corporations. Yet we now have a digital, networked landscape, where regular individuals suddenly have the authoring and publishing power once reserved only to powerful entities with extraordinary resources. This infrastructure for free culture—the internet—is now setting the stage for a struggle for realization of free culture, between the people and the powerful (Vaidhyanathan, 2004).

¹⁵<http://www.pgdp.net/>.

¹⁶<http://www.furl.net/>.

¹⁷<http://www.citeulike.org/>.

¹⁸<http://www.deliciou.us/>.

¹⁹The upshot of these kinds of sites is a collaborative filtering effect, whereby each user’s efforts to organize and categorize the web leads to an emergent aggregate organizational effect called a *folksonomy* (Mathes, 2004).

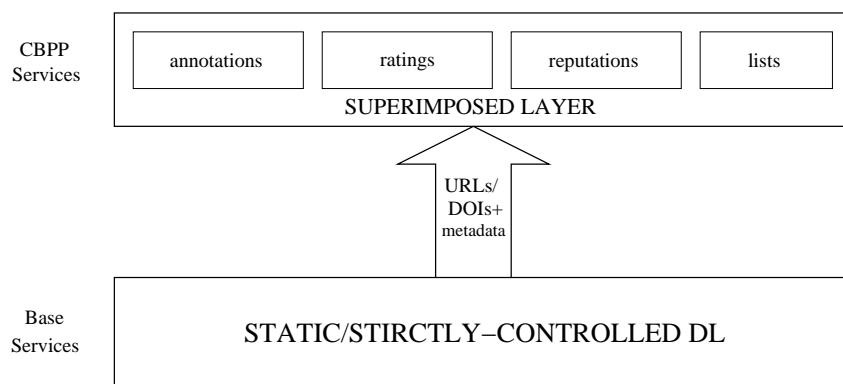


Figure 3: **Superimposed CBPP**: How digital libraries can employ CBPP without becoming completely collaborative. An underlying, central content base is “linked” to services and collaboratively-produced data at a superimposed layer. The example superimposed CBPP services given are annotations, ratings, reputations (systems), and user-created lists (which group resources in the DL).

The prospect of the “little guy” having an impact on culture is considered by some to be a scary thing. There are worries of confusion, information overload, inaccuracy, fraud, degeneracy, and vandalism. Yet, the systems toured in this article strongly suggest that these concerns are not terminal, if even significant.

In sum, opposition to CBPP and free culture is largely illegitimate: it is the result of fear of the new and unknown and/or an intuitive understanding from vested interests that they are being threatened. But the reality is that markets and firms are simply not enough to produce a vibrant, diverse, healthy culture, in a world where firms dominate the individual and the commons. This kind of world is an alien thing to the unchanging, social nature of humankind. A free culture which includes widespread CBPP systems offers a more natural alternative.

6.2 The Copyright Situation

As discussed earlier, for free culture to work through CBPP, people must be free to access, free to extend, and free to contribute. At any one of these steps, there commonly exist in the present day numerous, often insurmountable

legal pitfalls.²⁰

A rough (and probably incomplete) taxonomy of scenarios for these copyright pitfalls is:

- *Orphaned works* - Archives hold reams and volumes of digitized works, to which the copyright holder is unknown or unavailable. Due to how copyright terms have been extended, and how copyright defaults to “all rights reserved,” these works are therefore unavailable for sharing or making derived works.
- *Lock-up of significant cultural and knowledge works* - The well-known examples here are “free the mouse” (Economist, 2002) (i.e., how Disney has for three-quarters of a century retained iron-clad control over how their creations are used, despite their status as popular culture) and scholarly journals (which are typically closed to the general public).
- *No protection for CBPP efforts* - The “blessing” of distributed authorship of CBPP projects turns into a curse when current copyright law is applied. As the number of contributors increases (something which is in fact a reasonable and desirable goal), the probability of one of them intentionally or unintentionally causing the entire project to run afoul of copyright law approaches near-certainty. While I believe this is a small problem logistically, it is a huge problem legally.
- *Elimination of analogous fair use in the digital world* - Since every operation in the digital world is a “copy,” copyright technically forbids almost all operations. Thus, it is generally not permitted to “lend” an e-book to a friend, or to send to them some of the music on your iPod. Those who control information have exploited this unforeseen technicality to apply legal restrictions to what was once fair use, in order to ensure their own profit.
- *Re-capture of public domain works* - Related to the previous item, many information providers (even digital libraries) assert copyright which they have dubious claim to, over public domain materials which they

²⁰Copyright roadblocks are less a problem for purely-superimposed, “feedback” style CBPP services (such as a ratings) than for the contribution of extensive content objects (as in article-writing or review contribution) or any modification of primary content objects (as in metadata enhancement).

carry in some form. This has been called “thin copyright,” and it is unclear whether it is really allowed (Puzio, 2005).

- *Criminalization of “circumvention” (technicalization of copyright)* - As Vaidhyanathan and Lessig point out (Lessig, 2004; Vaidhyanathan, 2004), the DMCA has radically altered the nature of copyright by criminalizing any activity which circumvents copy protection technology (no matter how flimsy this technology is). Since this is now a criminal offense separate from normal copyright violation, and as aforementioned, every operation is a copy, the government has in effect outsourced thinking with respect to copyright to private companies, while giving them unlimited access to government power of reprisal.
- *Stillborn works* - As I realized when listening to Apple’s Bud Tribble deliver a speech (Tribble, 2005), some works are locked out of usage upon creation, due to the innate attributes of the author or the circumstances surrounding creation. As Tribble pointed out in his anecdote, kids who created valuable multimedia learning resources as a part of Apple’s school outreach programs were unable to share these works—because children are not “authorized” to assign usage license to their own creations. Consequently, one free educational digital library that could have existed did not.
- *Buried works* - Works that were once widely available for free (i.e., through libraries) often become *more scarce* after they are digitized (you read that correctly). Once paper copies of old works are tossed or sent to long-term storage, researchers must rely more heavily on their digitized representations. However, as discussed earlier, the odds of having a subscription to access these locked-down works can be quite poor (and worsening). For example, during the writing of this paper, a friend of mine doing his PhD work approached me for help finding a classic mathematics article from the 70s. The work was digitized and from a prominent journal, but he had been unable to access it through any of the *six* institutional libraries he tried!²¹

²¹Even if he eventually gets a copy of the work (in any form), he still will never get back the time wasted on searching and waiting. And if he was successful, he’d technically be violating terms-of-use of network access at any but his home institution. And further, if I found the article, I’d be running afoul of copyright law by sharing it with him. It is difficult to see how this model furthers the public good.

Most of these scenarios (except perhaps the last two) are well-covered in the rest of this volume. However, within this symposium and elsewhere, they have generally been seen as isolated copyright *problems*. This is understandable, given how we are all most familiar with the situations we've encountered, but it is not the case—they are all simply a result of the same underlying, outdated, flawed model of intellectual property, and a copyright regime built upon it. Thus, they should be properly taken as *scenarios*. Efforts to address them can then be focused on the common (and dysfunctional) underlying rules, which establish under what conditions permission (to disseminate or derive) is needed.²²

It is my hope that the confluence of minds and visibility of issues afforded by this symposium will accelerate progress in “solving copyright,” thus making all of the above “nightmare scenarios” go away.

6.3 Technical Challenges

Impediments to CBPP are not all perceptions and abstract legal conditions. There actually *are* technical challenges to making CBPP work—and critics latch onto these challenges as if they disprove the utility of the entire mode of production.

Previously, “information oligarchies” were naturally induced because the technical challenges of authoring (especially authoring *en masse*) and publishing were intractable otherwise (Vaidhyanathan, 2004). Now, we have powerful computer processing network architectures which enable solutions of most of these fundamental problems. So far, we've witnessed the birth of CBPP as a consequence of these advances. But more could be done to improve CBPP systems and increase their applicability, especially in the area of quality control.

Quality control, in fact, seems to be the last safe harbor for the CBPP naysayers. Even when CBPP resources (such as Wikipedia) are quite clearly high-quality, the uncertainty of how this was achieved seems to make the fact invisible to some. The “sleight of hand” is that quality in such systems is *emergent*, either through collaborative filtering, voting, or the sheer domination of good contributions over bad (Krowne, 2005). Making these emergent systems work is much more difficult than simply giving some authority the power to include or exclude portions and to give explicit indicators of quality.

²²See, especially, (Lessig, 2004) for more.

Further, explaining how this process happens in CBPP is more complicated than “because the editors (or moderators) say it is so.”

While real, I propose that quality control in CBPP is now more a problem of *degree* than of *kind*. The computer and network revolution is not over. New methods have yet to be developed and applied. Free toolkits for performing generalized CBPP quality-control functions will be developed and replicated many-fold.²³ CBPP systems can and will be taken farther, and will displace more and more information resources that could formerly only be produced centrally. Those who want the book to be closed now are setting themselves up for disappointment.

7 Conclusion

Perhaps the greatest success of the digital library community to date has in fact been Google. Yet, this paper has argued that even with this major (but essentially accidental) creation, the promise and potential of digital libraries has not been fully met. To remain as relevant and useful as Google and additionally fulfill this latent potential, digital libraries will need to support free culture by integrating commons-based peer production services.

While digital libraries that do not do this may not up and “vanish” overnight, in the near future, they may find their social impact shrinking relative to alternative resources. Technically superior resources, as Google has demonstrated, will not necessarily be libraries. And free culture resources will provide a compelling alternative to static, oligarchic, top-down-controlled information silos—a model which the digital library world has subscribed almost exclusively to thus far.

As there is much benefit to be had from the sustainable infrastructure of officially-funded efforts, it is my hope that digital libraries within such environments will embrace the benefits of CBPP and an ideal of free culture. By using pooled resources to provide a kind of knowledge to everyone which also considers worthwhile input *from* everyone, overhead efficiency as well as social impact would be maximized. In such a world, digital libraries could better-foster widespread equity in knowledge, empowerment to create and use it, and social harmony resulting from cooperatively doing so. Perhaps then the digital library truly could be the savior of civilization.

²³For example, see CoFE (<http://eecs.oregonstate.edu/iis/CoFE/>), an open source collaborative filtering system.

Acknowledgements

I would like to extend special thanks to those who reviewed and made suggestions to earlier versions of this article, in particular, Ross Singer, Joe Corneli, Raymond Puzio, and Robert Milson.

References

- Yochai Benkler. Coase's penguin, or, linux and the nature of the firm. *Yale Law Journal*, 112:369–446, 2002.
- Daniel K. Biss. The elephant in the internet. *Notices of the AMS*, November 2004.
- Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998. URL <http://citeseer.ist.psu.edu/brin98anatomy.html>.
- Peter Brusilovsky, Rosta Farzan, and Jae wook Ahn. Comprehensive personalized information access in an educational digital library. In *Proceedings of JCDL 2005*, June 2005.
- Vannevar Bush. As we may think. *The Atlantic Monthly*, 176(1):101–108, July 1945. URL <http://www.theatlantic.com/doc/prem/194507/bush>.
- Joe Corneli and Aaron Krowne. A scholia-based document model for commons-based peer production. In *Proceedings of the Symposium on Free Culture and the Digital Library*, October 2005.
- Economist. Free mickey mouse. *The Economist*, Oct. 10 2002. URL http://www.economist.com/printedition/displayStory.cfm?Story_ID=1378700.
- Miles Efron and Donald Sizemore. Link attachment (preferential and otherwise) in contributor-run digital libraries. In *JCDL*, pages 369–371, 2003.
- Gary Flake, Steve Lawrence, and C. Lee Giles. Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, August 20–23 2000. URL <http://citeseer.ist.psu.edu/flake00efficient.html>.

- Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans Coetzee. Self-organization of the web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002. URL <http://citeseer.ist.psu.edu/flake02selforganization.html>.
- Edward A. Fox, Neill Kipp, and Paul Mather. How digital libraries will save civilization. *Database Programming & Design*, 11(8):60–64, August 1998.
- David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234, 1998. URL <http://citeseer.ist.psu.edu/gibson98inferring.html>.
- Marcos André Gonçalves, Edward A. Fox, Layne T. Watson, and Neill A. Kipp. Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. *ACM Trans. Inf. Syst.*, 22(2):270–312, 2004.
- Michael Gorman. Google and God’s mind. *LA Times*, December 2004.
- Michael Gorman. Revenge of the blog people. *Library Journal*, February 2005. URL <http://www.libraryjournal.com/article/CA502009>.
- Hui Han, C. Lee Giles, Eren Manavoglu, Hongyuan Zha, Zhenyue Zhang, and Edward A. Fox. Automatic document metadata extraction using support vector machines. In *JCDL*, pages 37–48, 2003.
- F. Iannacci and E. Mitleton-kelly. Beyond markets and firms. *First Monday*, 10(5), May 2005. URL http://www.firstmonday.org/issues/issue10_5/iannacci/index.html.
- Aaron Krowne. The FUD-based encyclopedia. *Free Software Magazine*, (2), March 2005. URL http://www.freesoftwaremagazine.com/free_issues/issue_02/fud_based_encyclopedia/.
- Lawrence Lessig. *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*. Penguin Press, 2004.
- Sarah Lippincott and Martha Kyrillidou. How ARL university communities access information: Highlights from LibQUAL+™. *ARL Bimonthly Report*, 236, October 2004. URL <http://www.arl.org/newsltr/236/lqaccess.html>.

Xiaoming Liu, Kurt Maly, Mohammad Zubair, and Michael L. Nelson. DP9: an OAI gateway service for web crawlers. In *JCDL*, pages 283–284, 2002. URL http://www.cs.odu.edu/~liu_x/dp9/dp9.pdf.

Adam Mathes. Folksonomies - cooperative classification and communication through shared metadata, December 2004. URL <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>.

Robert D. McHenry. The faith-based encyclopedia. *Tech Central Station*, November 2004. URL <http://www.techcentralstation.com/111504A.html>.

Robert Milson and Aaron Krowne. Adapting CBPP platforms for instructional use. In *Proceedings of the Symposium on Free Culture and the Digital Library*, October 2005.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998. URL <http://citeseer.ist.psu.edu/page98pagerank.html>.

Raymond Puzio. On free math and copyright bottlenecks. In *Proceedings of the Symposium on Free Culture and the Digital Library*, October 2005.

P. Krishna Reddy and Masaru Kitsuregawa. An approach to relate the web communities through bipartite graphs. In *WISE*, pages 301–310, 2001. URL <http://citeseer.ist.psu.edu/reddy01approach.html>.

John J. Regazzi. The battle for mindshare: A battle beyond access and retrieval (miles conrad memorial lecture). In *46th NFAIS Annual Conference*, February 2004. URL http://www.nfais.org/publications/mc_lecture_2004.htm.

Ross Singer. Google scholar and the dawn of web localization, 2005. URL <http://rsinger.library.gatech.edu/papers/WebLocalizing.html>.

Bud Tribble. (keynote speech). In *JCDL*, June 2005.

Siva Vaidhyathan. *The Anarchist in the Library*. Basic Books, 2004.

Fred Vogelstein. Gates vs. Google: Search and destroy. *Fortune*, April 2005.
URL <http://www.fortune.com/fortune/print/0,15935,1050065,00.html>.

Meilan Zhang and Chris Quintana. Facilitating middle school students' sense making process in digital libraries. In *Proceedings of JCDL 2005*, June 2005.