Crowdsourcing and Open Access: Collaborative Techniques for Disseminating Legal Materials and Scholarship

Timothy K. Armstrong

Follow this and additional works at: http://digitalcommons.law.scu.edu/chtlj

Part of the Law Commons

Recommended Citation
Available at: http://digitalcommons.law.scu.edu/chtlj/vol26/iss4/4

This Article is brought to you for free and open access by the Journals at Santa Clara Law Digital Commons. It has been accepted for inclusion in Santa Clara High Technology Law Journal by an authorized administrator of Santa Clara Law Digital Commons. For more information, please contact sculawlibrarian@gmail.com.
CROWDSOURCING AND OPEN ACCESS: COLLABORATIVE TECHNIQUES FOR DISSEMINATING LEGAL MATERIALS AND SCHOLARSHIP

Timothy K. Armstrong†

Abstract

This short essay surveys the state of open access to primary legal source materials (statutes, judicial opinions and the like) and legal scholarship. The ongoing digitization phenomenon (illustrated, although by no means typified, by massive scanning endeavors such as the Google Books project and the Library of Congress’s efforts to digitize United States historical documents) has made a wealth of information, including legal information, freely available online, and a number of open-access collections of legal source materials have been created. Many of these collections, however, suffer from similar flaws: they devote too much effort to collecting case law rather than other authorities, they overemphasize recent works (especially those originally created in digital form), they do not adequately hyperlink between related documents in the collection, their citator functions are haphazard and rudimentary, and they do not enable easy user

† Associate Professor of Law, University of Cincinnati College of Law. B.A. 1989, M.P.Aff. 1993, J.D. 1993, The University of Texas at Austin; LL.M. 2005, Harvard Law School. This work began as a series of presentations which I delivered at the summer 2008 and 2009 CALI Conferences for Law School Computing, and an updated version was presented at the meeting of the Section of Internet and Computer Law at the 2010 Annual Meeting of the Association of American Law Schools (AALS). I appreciate the thoughtful comments of the attendees at each of these sessions. Research support from the Harold C. Schott Foundation is gratefully acknowledged, as is the research assistance of Ron Jones. Copyright © 2010, Timothy K. Armstrong.

This work is licensed under the Creative Commons Attribution-Share Alike 3.0 United States license. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/3.0/us/, or send a letter to Creative Commons, 171 2nd Street, Suite 300, San Francisco, California, 94105, USA. For purposes of Paragraph 4(c) of said license, proper attribution must include the name of the original author and the name of the Santa Clara Computer & High Technology Law Journal as publisher, the title of the Article, the Uniform Resource Identifier, as described in the license, and, if applicable, credit indicating that the Article has been used in a derivative work.
authentication against official reference sources.

The essay explores whether some of these problems might be alleviated by enlarging the pool of contributors who are working to bring paper records into the digital era. The same "peer production" process that has allowed far-flung communities of volunteers to build large-scale informational goods like the Wikipedia encyclopedia or the Linux operating system might be harnessed to build a digital library. The essay critically reviews two projects that have sought to "crowdsource" proofreading and archiving of texts: Distributed Proofreaders, a project frequently held up as a model in the academic literature on peer production; and Wikisource, a sister site of Wikipedia that improves on Distributed Proofreaders in a number of ways. The essay concludes by offering a few illustrations meant to show the potential for using Wikisource as an open-access repository for primary source materials and scholarship, and considers some possible drawbacks of the crowdsourced approach.

I. INTRODUCTION

The digital era has exposed the limitations of paper as an archival medium. Although paper (like other forms of hard-copy) makes an excellent tool for transmitting knowledge across lengthy spans of time, it makes a poor tool for transmitting knowledge across lengthy spans of distance. A wealth of knowledge, including legal knowledge, remains effectively trapped inside paper records, where it can be used only by those with access to the physical medium in which it is contained.

The movement to digitize paper records and make them freely available online promises to liberate information, including legal information, from these physical constraints and make it accessible around the globe. The scope of the task, however, is massive and daunting. Even the best organized (and best funded) efforts, such as the Google Books project (currently the subject of copyright litigation\(^1\)) and the Library of Congress’s efforts to scan American historical documents,\(^2\) can only scratch the surface. Indeed, the Library of Congress recently estimated that, at its present pace, it will

---

2. See infra note 69.
take *almost two thousand years* to digitize the nine billion text records it presently holds in its collection.  

Wikis and other collaborative tools change this picture in potentially important ways. Just as other informational projects have benefited by opening themselves to participation by a distributed community of volunteers, the means now exist to harness the efforts of legal professionals, students, and even interested members of the public at large to improve access to legal information, court decisions, statutes and regulations, and legal scholarship. In 2008, for example, the participants in one such project (initiated by the present author) succeeded in making crucial portions of the legislative history of the landmark Copyright Act of 1976 freely available online for the first time. The online version of the Copyright Act’s legislative history improves access not only by duplicating the text of the original report, but—perhaps more importantly—by making it possible for other online works that cite the report to hyperlink to it. This creates a seamless web of knowledge that improves upon the practical experience of using reference sources in paper form. If we multiply this isolated example by dozens, hundreds, or thousands of interested online users of legal texts, the possibility of a transformative moment in access to legal knowledge begins to appear ever closer.

This essay begins with a review of the open access imperative, which may be normatively grounded in considerations of transparency, democratic legitimacy, and the fulfillment of the university’s public service mandate. It then surveys the current status of a number of projects aimed at improving public access to legal

---


5. *See infra* note 142 and accompanying text.


7. As one writer put it: "Many nerds believe that a billion readers can reliably weave together the pages of old books, one hyperlink at a time. Those with a passion for a special subject, obscure author or favorite book will, over time, link up its important parts. Multiply that simple generous act by millions of readers, and the universal library can be integrated in full, by fans for fans." Kevin Kelly, *Scan This Book!*, N.Y. TIMES, May 14, 2006, § 6 (Magazine), at 42, 45.
materials and scholarship, and explores whether "crowdsourced,"
Wiki-centered efforts may achieve comparable results at lower cost. It
concludes with an assessment of some of the drawbacks and
limitations of the "crowdsourced" approach.

II. POLICY BACKGROUND: THE OPEN ACCESS IMPERATIVE

A. Open Access to Scholarship

"Open access," in the sense of making documentary materials
available over the Internet for reading and copying without charge, is
an emerging phenomenon in the legal academy. In the legal academic
community, the "open access" label is associated primarily with free
distribution of scholarly works. The discussion has revolved around
whether to improve access to faculty scholarship, how best to do so,
and what it might mean for the traditional legal publishing paradigm.

At one level, enlisting faculty support for scholarly open-access
initiatives consists merely of fostering personal and institutional self-
interest. Inaccessible scholarship is unpersuasive scholarship, and studies have tended to suggest that opening access to scholarly works correlates with greater scholarly impact (as measured by citation counts). Researchers' growing reliance on the Internet as a complement—and perhaps, one day, a successor—to proprietary databases or library hard copies feeds the demand for open access to scholarly works. Furthermore, the same technologies that enable open access to traditional legal scholarship also give scholars new forms to express themselves, creating forms of scholarly discourse that would have been uneconomical to produce in the pre-Internet era.

The movement to assure open access to scholarship is more advanced outside the legal academy. The difference is partly explained by differing market dynamics: University libraries, driven by eye-popping increases in subscription costs for specialized research journals, responded by dropping subscriptions, creating a risk that scholars working in those specialized fields would find it more difficult both to remain abreast of developments and to ensure dissemination of their own work to their peers. Open access

10. See Willinsky, supra note 8, at 22 ("[O]pen access is associated with increased citations for authors and journals, when compared to similar work that is not open access"); id. at 22–24 (summarizing research). For a look at some of the methodological pitfalls of studies of this type, as well as some possible solutions, see Bernard S. Black & Paul L. Caron, Ranking Law Schools: Using SSRN to Measure Scholarly Performance, 81 IND. L.J. 83, 92–95 (2006).

11. See Solum, supra note 8, at 859 ("There will come a day when the saying, 'If it isn't on the net, it doesn't exist,' is true. Open access legal scholarship will be the only legal scholarship that is actually read. Closed access legal scholarship will be the tree that falls with no one in the forest."); Carol A. Parker, Institutional Repositories and the Principle of Open Access: Changing the Way We Think About Legal Scholarship, 37 N.M. L. REV. 431, 431 (2007) (suggesting that "open access to legal scholarship will soon be adopted and implemented by every law school in the United States"); Richard A. Danner et al., The Twenty-First Century Law Library, 101 LAW LIBR. J. 143, 146 (2009) ("T[he] fact that young people are going to Google and to Wikipedia first is a call to arms in a way") (comments of Richard A. Danner).


13. See Willinsky, supra note 8, ch. 2; Dan Hunter, Walled Gardens; 62 WASH. & LEE L. REV. 607, 613–17 (2005). The effect of subscription costs as a driver of open access is surely greater in technical fields, where subscription rates for specialty publications may run into the thousands (or even tens of thousands) of dollars a year, than in law. But see Danner, supra note 8, at 377 ("[B]ecause they enjoy unlimited (and apparently cost-free) access to law journals and other information through Westlaw, LexisNexis, Hein Online, and other databases, it might be hard for law students and faculty to appreciate the impacts of access costs on researchers outside the U.S. legal education environment."); Solum, supra note 8, at 863 ("[A]s you move from major research universities to regional universities to local colleges, the access of faculty and students to closed electronic databases (Westlaw, LexisNexis, JSTOR, etc.) begins to become
publishing solves both problems by making current scholarship available worldwide at little expense. For that reason, faculty at several influential research institutions have voted to authorize archiving and distribution of their scholarship on open-access terms. Harvard University’s Faculty of Arts and Sciences did so (by unanimous vote) early in 2008, and the Harvard Law School faculty unanimously followed suit a few months later. The Massachusetts Institute of Technology adopted a university-wide open access mandate in early 2009, and similar measures are pending or have been adopted by other universities.

The adoption of open-access mandates by university faculty has led to the creation of institutional electronic repositories of scholarly works. Duke Law School’s faculty scholarship repository includes faculty papers dating back over half a century. Harvard’s new DASH repository may be unique in including student-authored papers alongside faculty scholarship. Nor is the push for scholarly open access confined to elite institutions: the Oklahoma City University School of Law, for example, maintains a repository of faculty scholarship extending back four decades. Cross-institutional repositories such as SSRN and BEPress hold even larger collections of faculty scholarship from universities worldwide.

Some law journals have also committed to publishing on an open-access model. The Science Commons organization (an affiliate

very sketchy. In the least-developed countries, such access is virtually nonexistent.”).


17. On the other hand, the news is not uniformly favorable. In April 2009, the faculty of the University of Maryland defeated a resolution encouraging (but not requiring) that faculty members make their scholarship available in open-access repositories.


19. The DASH repository is online at http://dash.harvard.edu/ (last visited Oct. 6, 2009).


of Creative Commons) has an Open Access Law Program ("OALP") that is intended to foster open access to legal scholarship by permitting authors to retain sufficient rights in their published works to enable those works to be hosted in open-access repositories.  

Several law journals have committed to honor the principles of open access in works published in their pages, and authors may obtain similar results even when publishing in journals that have not formally committed themselves to the OALP's principles.

B. Open Access to Primary Source Materials

As important as the movement to open access to legal (and other) scholarship is, the public may derive still greater benefit from making primary legal source materials—statutes, regulations, case law, and the like—more broadly accessible. Positive law directly regulates individual behavior, and partly for that reason, individual due process interests in access to the law have long been recognized. Principles of democratic legitimacy also favor open access by citizens to information necessary to police the functioning of government.

Although open access to primary legal source materials would appear to present an uncontroversial imperative, both legal and practical obstacles remain. The Copyright Act places federal statutes and judicial decisions in the public domain. But the absence of any comparable statutory provision regarding the copyright status of

24. See, e.g., infra note 149 and accompanying text.
26. See, e.g., id. at 746; Justin Hughes, Created Facts and the Flawed Ontology of Copyright Law, 83 NOTRE DAME L. REV. 43, 77–78 (2007) (considering several justifications for open access to court decisions); Nash v. Lathrop, 6 N.E. 559, 560 (Mass. 1886) ("it needs no argument to show that justice requires that all should have free access to the opinions, and that it is against sound public policy to prevent this, or to suppress and keep from the earliest knowledge of the public the statutes, or the decisions and opinions of the justices.").
primary legal source materials below the federal level has occasionally sparked controversy.\textsuperscript{29} Even at the federal level, a lack of consensus on the normative desirability of open access is evident in the competing bills now pending on the subject of open access to federally funded research.\textsuperscript{30}

\textbf{C. Roles of Universities and Public Institutions}

As prodigious producers and consumers of information, public and private universities stand among the most important institutional actors in the networked information infrastructure. Organized action by the academic community would go a long way towards making open access to information the norm rather than the still-developing exception. As of yet, however, progress towards inculcating open-access norms in the academic community has been halting and sporadic. The praiseworthy example shown by the prominent research institutions whose faculties have adopted open-access mandates\textsuperscript{31} only highlights the vastly greater number of institutions that have not.

Fostering open access naturally fits with the mission of the modern university along multiple dimensions. Faculty research benefits from a regime in which source materials are freely accessible,

\begin{itemize}
\item \textsuperscript{29} See, e.g., Veeck v. Southern Building Code Cong. Int'l, 293 F.3d 791 (5th Cir. 2002) (en banc) (reasoning that copyright protection in text of privately authored model building code evaporated when that code was enacted as positive law by two municipalities); L. Ray Patterson & Craig Joyce, \textit{Monopolizing the Law: The Scope of Copyright Protection for Law Reports and Statutory Compilations}, 36 UCLA L. REV. 719, 809–10 (1989) (reasoning that pagination and chapter or section numbering of primary legal source materials are insufficiently expressive to qualify for copyright protection, and decrying efforts to enforce such protections as “in effect impos[ing] a tax for the use of the law”); Katie Fortney, \textit{Ending Copyright Claims in State Primary Legal Materials: Towards an Open Source Legal Operating System}, at http://ssrn.com/abstract=1347158 (last visited Oct. 20, 2009).
\item \textsuperscript{30} Compare H.R. 801, 111th Cong. (2009) (forbidding federal agencies to condition research funding upon recipients' archiving of findings in open-access repositories) with S. 1373, 111th Cong. (2009) (encouraging agencies to adopt such policies).
\item \textsuperscript{31} See supra notes 14–17 and accompanying text.
\end{itemize}
and faculty scholarship that is made available in an open-access forum promises greater impact. Furthermore, the widespread adoption of open-access initiatives may yield substantial benefits even outside the directly involved university community. Most universities conceive of their missions as including substantial public service and public education components (indeed, for state-funded institutions, such mandates may be enacted as positive law), and it is not difficult to situate efforts to make information more widely available within the broad domain of public service.

In dealing with the open-access phenomenon, university libraries in particular confront imperatives that do not point uniformly in a single direction. To be sure, many librarians rightly see themselves as natural allies of the open-access movement and as well-positioned advocates for open-access policies because those policies best meet the needs of the library's core constituencies. On the other hand, the logic of open access enables disintermediation to occur simultaneously at many levels: just as open access may reduce the role and importance of publishers, so too may it diminish the historical status of libraries themselves as informational gatekeepers. It is no simple task for libraries and librarians to balance the conflicting incentives presented by the open-access phenomenon, which makes some of the pro-access steps that libraries and universities have been taking recently all the more remarkable.

32. See, e.g., OHIO REV. CODE ANN. § 3333.04(A), (E) (West 2009) (directing state board of regents to consider "the needs of the people," among other criteria, in identifying the "public services which should be offered" by state-supported higher education institutions).

33. See WILLINSKY, supra note 8, at 65, 227–32 (arguing for establishment of open-access publishing and archiving cooperative based partly on fulfillment of participating institutions' public service mandates).

34. See, e.g., James G. Neal, A Lay Perspective on the Copyright Wars: A Report from the Trenches of the Section 108 Study Group, 32 COLUM. J.L. & ARTS 193, 194 (2008) ("Universities and libraries are committed to openness—general and barrier-free access to information framed by the rhetoric of open source, open standards, open archives and open knowledge."); id. at 197–98 (discussing principles and policies developed by library community, several of which involve greater access to information).

35. See id. at 194 (noting Professor Clayton Christiansen's definition of "disruptive technologies" that "enable[] a larger population of less skilled people to do the things that historically only an expert could do") (footnote omitted). On the complex pattern of incentives that libraries may face to limit free access to digital collections, see Guy Pessach, The Role of Libraries in A2K: Taking Stock and Looking Ahead, 2007 MICH. ST. L. REV. 257, 261–62 (2007). The possibility that mass digitization projects may threaten adverse effects on the operation of libraries is an important factor behind the equivocal stance the library community has adopted towards the proposed settlement of the ongoing Google Book Search copyright litigation. See, e.g., Supplemental Library Association Comments on the Proposed Settlement, Authors Guild, Inc. v. Google Inc., No. 05-CV-8136-DC (S.D.N.Y.), at http://www.arl.org/bm-doc/library-associations-supp-filing-sept-2-09.pdf.
In early 2009, the directors of some of the largest and most prestigious law libraries in the United States issued the “Durham Statement on Open Access to Legal Scholarship,” a document that has been signed by over 50 librarians and other supporters nationwide.36 The Durham Statement envisions a wholesale restructuring of the endeavor of producing and disseminating legal scholarship: not only does it call for the increased deployment of stable digital repositories of faculty scholarship (as already exist in many forms), it goes further and calls upon law schools to rely solely on such digital repositories and to cease publishing law journals in printed form.37

On the question of open access to primary legal source materials, Professor Ian Gallacher believes law schools (and libraries) should take a leading role.38 Making primary source materials freely available would do much to serve the large number of law school graduates who practice outside large-firm settings.39 Gallacher goes on to argue that law schools’ institutional incentives may make them more desirable and effective custodians of primary source materials than private publishers (whose proprietary incentives may discourage widespread or free distribution) or even the government itself.40 Gallacher’s essay concludes by articulating a number of design standards that should be adopted by open-access archives of primary legal source materials, including: (1) universal openness and accessibility (whether with or without charge); (2) completeness; (3) flexibility in access and presentation; (4) flexibility in search and indexing methods; (5) speed; (6) reliability; (7) permanence; (8) vendor neutrality; (9) a citator or validator to identify doubtful precedents; and (10) community involvement in the development and maintenance of the archive.41 Some of the most robust nonproprietary legal databases already share many of these characteristics, with the

37. See id. (“We therefore urge every U.S. law school to commit to ending print publication of its journals and to making definitive versions of journals and other scholarship produced at the school immediately available upon publication in stable, open, digital formats, rather than in print.”).
39. See id. at 14–19.
40. See id. at 21–31.
41. See id. at 32–49.
last—community-centered development—perhaps most in need of strengthening.

III. TECHNICAL BACKGROUND: OPEN ACCESS INITIATIVES

Efforts to make legal information freely accessible online represent currents within a much broader stream that is the mass digitization movement. Exemplified by large-scale projects like Google Books, the Internet Archive, and Project Gutenberg, the digitization movement has long since succeeded in making the complete works of Shakespeare and Dickens, among countless others, freely accessible to a global audience. The all-encompassing ambitions of the largest digitization initiatives have, almost by happenstance, swept some legal materials within their ambit; it is possible, for example, to find early volumes of the Harvard Law Review on Google Books, or to read the United States Reports on the


44. Project Gutenberg’s home page is available at http://www.gutenberg.org (last visited Nov. 12, 2009). The project’s activities are considered in Zimmerman, supra note 43, at 995; Hannibal Travis, Building Universal Digital Libraries: An Agenda for Copyright Reform, 33 PEPP. L. REV. 761, 784 (2006). Project Gutenberg’s crowdsourced (but not Wiki-driven) proofreading engine, Distributed Proofreaders, is discussed infra at notes 86–106 and accompanying text.

45. Lists of Shakespeare’s works, with links to the text of each, are available at http://www.gutenberg.org/browse/authors/s (last visited Nov. 12, 2009); http://en.wikisource.org/wiki/Author:William_Shakespeare (last visited Nov. 12, 2009).

46. Lists of Dickens’s works, with links to the text of each, are available at http://www.gutenberg.org/browse/authors/d (accessed Nov. 12, 2009); http://en.wikisource.org/wiki/Author:Charles_Dickens (accessed Nov. 12, 2009). Were he alive today, Dickens himself would likely react unfavorably to the discovery that the natives of far-flung locales could easily read his great works without paying a penny in royalties. See, e.g., Larisa T. Castillo, Natural Authority in Charles Dickens’s Martin Chuzzlewit and the Copyright Act of 1842, 62 NINETEENTH-CENTURY LITERATURE 435, 436 (2008).
Internet Archive. The very scope of those projects, however, may diminish their utility: a search for "Harvard Law Review" on Google Books yields in excess of 10,000 "hits," most of which are mere mentions of the Review in otherwise unrelated publications.47 A search for "United States Reports" on the Internet Archive returns 164 results (far fewer than the actual number of published volumes) arranged seemingly at random; again, the results include mentions of the United States Reports in other publications arrayed indiscriminately alongside actual copies of the Reports themselves.48

Digitization initiatives focused specifically on legal materials, although less well developed overall, have already reached significant milestones in coverage and usability.

- Cornell Law School’s Legal Information Institute (LII)49 hosts a number of regularly updated federal law resources, including the United States Code50 and the Code of Federal Regulations.51 Although LII posts helpful information on each page about how

---

47. A Google Books search for “Harvard Law Review,” enclosed in quotation marks, returned 10,938 results when performed on November 13, 2009. Most of the top ten results were actual scans of early volumes of the journal, but after that the results quickly veered into unrelated publications—snippets of autobiographies whose authors mentioned their time on the Review, for example. Performing a narrower author search for "Harvard Law Review Association" yielded far fewer hits, but again, the results included a number of publications besides issues of the Review.

48. I performed a search for “United States Reports” limited to “media type: texts” on the Internet Archive on November 13, 2009 and received 164 results, most of which seemed to be scans of fairly recent volumes of the Reports. Again, the results list included many other publications also hosted at the Internet Archive that happened to mention the Reports.

My legal historian and librarian friends will no doubt be quick to remind me that “United States Reports” is a colloquialism of sorts; we now apply that name to a collection of materials that includes many early volumes never published with that title, which understandably might not turn up in a search for “United States Reports.” Nevertheless, given that more than 550 volumes of the Reports have been published (including over 400 since “United States Reports” became the official title of the series), a search that returns fewer than one-third that number of hits (and includes among the total many documents that are not themselves part of the United States Reports) still reveals something about the fragmentary and sporadic coverage of legal materials found in many of the general-purpose digitization projects.

49. The LII’s home page is available at http://www.law.cornell.edu/ (last visited Nov. 17, 2009). See also Gallacher, supra note 38, at 26 (praising LII as “the most visible example” of a law school’s “active[ly] engage[ment] in making the law accessible to everyone”).

50. LII’s United States Code portal is available at http://www.law.cornell.edu/uscode/ (last visited Nov. 17, 2009). Users may jump directly to a particular title and section, search for legislation by popular name, or browse the entire Code by following a hierarchical arrangement of links.

51. LII’s Code of Federal Regulations portal is available at http://www.law.cornell.edu/cfr/ (last visited Nov. 17, 2009). Users may jump directly to a particular title and section, or browse the full Code by following a set of links arranged hierarchically.
up-to-date that page’s content is, the site suffers most notably from the lack of a citator; that is to say, it is not possible to find cases construing any of the statutes or regulations hosted on LII.

- The student-organized AltLaw project\(^5\) hosts federal appellate case law, including nearly all Supreme Court cases and nearly six decades’ worth of lower federal appellate decisions.\(^5\) AltLaw’s cases are impressively, if incompletely, hyperlinked both forwards and backwards in time: most citations to other content available on AltLaw appear in the form of clickable hyperlinks, and the site includes a functional citator service via the “citations to/from this” link on each page. The site also offers a variety of search tools for sifting through its voluminous case law repository that, if perhaps not yet as fully developed as the query languages provided by proprietary database operators, nevertheless improve upon the basic functionality of searching for key words or phrases.\(^5\)

- The Justia project\(^5\) includes its own browseable copies of both federal and state legislation, along with a case law repository that, unlike most competing alternatives, also hosts district court decisions and dockets.\(^5\)

- Many of the foregoing projects draw data from the bulk collections maintained by public.resource.org, which has scanned thousands of pages of judicial and other records and made the results available for free download in a variety of formats.\(^5\)

---

\(^5\) AltLaw’s home page is available at http://www.altlaw.org/ (last visited Nov. 17, 2009). See also Gallacher, supra note 38, at 26 (singling out AltLaw as a particularly promising open-access resource).

\(^5\) See http://www.altlaw.org/v1/about/coverage (last visited Nov. 17, 2009).

\(^5\) See http://www.altlaw.org/v1/search/advanced (last visited Nov. 19, 2009); http://www.altlaw.org/v1/search/boolean (last visited Nov. 19, 2009). I do not mean to fault AltLaw’s designers for failing to match their impressive accomplishments in organizing and hyperlinking their hosted content with an equally sophisticated search engine. As David Weinberger has pointed out, one of the distinctive advantages of storing information digitally is that multiple overlapping organizational or searching schema may be adopted simultaneously without displacing or superseding other, equally valid, organizational schema for the same underlying content. See DAVID WEINBERGER, EVERYTHING IS MISCELLANEOUS: THE POWER OF THE NEW DIGITAL DISORDER 19-23 (2007). The important part of the process is the one at which AltLaw has excelled, namely, simply getting the content online and hyperlinked; higher-order indexing and search functions can follow later (or be developed by others) so long as they have an underlying pool of content upon which to work.

\(^5\) Justia’s home page is available at http://www.justia.com/ (last visited Nov. 19, 2009).

\(^5\) On the significance of Justia’s court document repository, see Peter W. Martin, Online Access to Court Records—From Documents to Data, Particulars to Patterns, 53 VILL. L. REV. 855, 885–87 (2008).

\(^5\) The project’s home page, unsurprisingly, is http://public.resource.org/ (last visited Nov. 19, 2009); see also John Markoff, A Quest to Get More Court Rulings Online, and Free,
The goal of making court records more accessible is shared by RECAP, a software extension that permits users to “liberate” records from the federal judiciary’s fee-based PACER service and make them available for free on the Internet Archive.\footnote{58}

The movement to make primary legal source materials freely accessible experienced what may one day prove to have been a transformative moment on November 17, 2009, when search-engine giant Google announced that it had made several recent decades’ worth of state and federal appellate, district, tax, and bankruptcy court decisions available for searching via its Google Scholar portal.\footnote{59} The site includes a rudimentary citator (accessible by clicking the “How cited” tab from each page) that usefully includes citations to secondary sources, such as articles and treatises online at Google Books.\footnote{60} Moreover, by harnessing the market-tested Google search engine, the addition of case law instantly made Google Scholar a “player” in the burgeoning market of alternatives to the traditional proprietary legal database publishers.\footnote{61} Google’s entry into the open-access case law world has drawn enthusiastic reactions from open-content advocates,\footnote{62} although the near-term effect may be simply that

\begin{footnotes}
\item[58] See \url{http://www.recapthelaw.org/} (last visited Nov. 19, 2009).
\item[60] The service, however, works in one direction only; that is, the Google Books treatises do not (yet?) link back to the cases they cite at Google Scholar. Nor is there a citator offered in either direction for statutes, none of which is included among the materials recently added to Google Scholar.
\item[61] See, e.g., John J. DiGilio, \textit{Bridging the Digital Divide: A New Vendor in Town? Google Scholar Now Includes Case Law}, LLRX.com, at \url{http://www.llrx.com/features/googlescholarcaselaw} (Nov. 18, 2009) (surveying pros and cons of Google’s new database); Mikhail Koulikov, \textit{Indexing and Full-Text Coverage of Law Review Articles in Nonlegal Databases: An Initial Study}, 102 LAW LIBR. J. 39, 52–37 (2010) (noting that “the emergence of search engines such as Google Scholar, which are not subscription-based, has presented an entirely new set of issues regarding the relationship between academics and published scholarship”); Eugene Volokh, \textit{The Future of Books Related to the Law?}, 108 MICH. L. REV. 823, 826 (2010) (predicting that advent of e-reader technologies, coupled with increasing availability of primary source materials in open-access repositories, will make it easier for reference sources commonly used in legal education, such as casebooks and treatises, to hyperlink directly to the cases and statutes cited therein).
Google crowds out other open-content projects—3—for example, in the wake of Google's announcement, the AltLaw project essentially declared that it was no longer relevant and would shut down. 4

Nevertheless, despite their noteworthy accomplishments in a comparatively short time, the many projects working to make legal information freely available online share some common drawbacks. Work focused on ameliorating these shared flaws would do a great deal to make open-access projects viable substitutes for proprietary legal databases.

• The bias towards collecting case law. First, nearly all the projects discussed above have focused on collecting the works of the judicial branch—a worthy endeavor, but one that risks bypassing the most important sources of governing authority in our "age of statutes." 5 Only the LII devotes substantial resources to collecting and updating federal executive and legislative materials such as the Code of Federal Regulations and United States Code. 6

• The bias towards contemporary sources. Legal history, if the extant online resources are any guide, began in the mid-20th century. 7 This is still a substantial improvement over the state of play in the open-access world as recently as three or four years ago, when history seemed to begin circa 1994. 8 But to a

---


64. See Lee Sims, In the Face of Google Assault, AltLaw Hangs It Up, at http://www.law.uconn.edu/content/face-google-assault-altlaw-hangs-it (last visited Dec. 1, 2009).


66. See supra notes 49-51 and accompanying text.

67. See, e.g., supra notes 53, 59 and references cited.

68. Vestiges of this earlier and more limited open-access world survive today. For example, the FindLaw web site maintains its own archive of federal appeals court decisions that date only from the mid-1990s. See Federal Courts of Appeal—Judicial Branch—Federal Resources, http://www.findlaw.com/10fedgov/judicial/appeals_courts.html (last visited Dec. 18, 2009), and the circuit-by-circuit archives linked from that page. The United States Supreme Court, to take another example, has posted bound volumes of its own decisions online, but the earliest available is Volume 502, which collects opinions from the October 1991 Term of the Court.
researcher seeking to illuminate the sources and development of doctrine, the absence of sources antedating 1950 or thereabouts represents a potentially serious impediment. 69

- Poor hyperlinking and citator functionality. Many open-access sites fail to take advantage of the improved capabilities that hypertext offers over publication of the exact same document in paper form. That is to say, although they reproduce the text of the courts’ opinions as published, they do nothing more than that: it is not possible to click and follow a citation that appears in the text of the court’s opinion and view the cited source, even if the other source is also online. Nor do most sites include any citator functionality whereby the opinion being displayed can be followed forward in time to locate references that cite it. 70

- Authentication against official referents. Reading an opinion hosted on the official site of the issuing court sends a powerful message that the text is authentic. Reading the same opinion on the site of a large proprietary publisher sends perhaps a different message, but not one that would cause most users to question the fidelity of the reproduction. Reading the same opinion on an open-access site, however, may prompt uncertainty as to the provenance


1995, roughly speaking, marks the point at which the courts of appeals began posting electronic copies of their own decisions online. Because these reported decisions were “born digital”—that is to say, created and disseminated initially in electronic form—storing and organizing them online entailed no digitization expense, which led to their rapid proliferation. In contrast, digitizing and archiving earlier works has frequently entailed substantial labor and expense. See, e.g., Markoff, supra note 57.

69. By way of an isolated counterexample, the Library of Congress’s American Memory Project has undertaken a massive and praiseworthy digitization initiative aimed at early American source texts. Of particular value is its collection entitled: A Century of Lawmaking for a New Nation: U.S. Congressional Documents and Debates, http://memory.loc.gov/ammem/amlaw/lawhome.html (last visited Jan. 4, 2010). Although this site hosts scanned images of a number of valuable texts not widely available elsewhere, such as documents produced by the Continental Congress and a number of works from the first years after the ratification of the Constitution, the Library of Congress has imposed a substantial technological impediment to easy access and use of the voluminous materials in its collection. In many instances, users may only view a single page image on screen at a time, and may only navigate forward and backward a page at a time, or jump to a specific, known, page number. Furthermore, for many of the most valuable works in its collection, the Library provides only page images, not text, making it impossible to (for example) copy-and-paste the language of early legislative enactments into another document. The present author’s Early United States Statutes project represents one effort to build upon the Library’s scanned document repository and make it more useful. http://homepages.uc.edu/~armstrty/statutes.html (last visited Jan. 4, 2010). For another such effort, see infra notes 143–148 and accompanying text.

70. The AltLaw and Google Scholar sites, as already noted, are exceptions to this general rule. See supra notes 52–54, 59–60 and accompanying text.
and authenticity of the text. The availability of mechanisms for verifying a text’s authenticity will be an important functionality in encouraging more widespread use of open-access alternatives to proprietary publishers.\(^{71}\)

In summary, although the legal open-access movement has attained some noteworthy successes, its shortcomings remain prominently visible. For recent federal appellate case law, multiple open-access alternatives exist, some of which have attained great sophistication. Once one moves beyond those types of legal materials, however, the situation becomes far murkier, with haphazard substantive coverage and an underdeveloped suite of tools available to deal with the posted content.

IV. CROWDSOURCING AS FORCE MULTIPLIER

A. Building an Informational Commons

A great deal has been written about the “commons-based peer production”\(^{72}\) phenomenon that began in the world of open-source software\(^{73}\) and has expanded in the past decade to support mass creation of a wide variety of expressive works.\(^{74}\) Open-content projects like Wikipedia harness the creative energies of a far-flung community of volunteers and enable them to collaborate asynchronously—their efforts mediated by the distributed architecture of the Internet\(^{75}\) and given legal stability through a family of specialized copyright licenses.\(^{76}\) Such mass collaboration—now frequently labeled “crowdsourcing”\(^{77}\)—enables the distributed

\(^{71}\) See Gallacher, supra note 38, at 40–41.


\(^{74}\) See, e.g., BENKLER, supra note 4.

\(^{75}\) This brief, descriptive summary surely understates the aspects of Wikipedia that are the most interesting and worthy of study; for a better assessment, see Jonathan Zittrain, The Future of the Internet—and How to Stop It ch. 6 (2008).

\(^{76}\) For peer-produced works, copyright licensing arrangements substitute for the legal authorizations that would otherwise be provided by the hierarchical structure of a private firm, enabling persons who are legal strangers to share, reuse, and expand one another’s expressive works. See Timothy K. Armstrong, Shrinking the Commons: Termination of Copyright Licenses and Transfers for the Benefit of the Public, 47 HARV. J. ON LEGIS. 359, 407 (2010).

\(^{77}\) Accessible overviews of the psychological and economic considerations that drive the crowdsourcing phenomenon are available in, e.g., JEFF HOWE, CROWDSOURCING: WHY THE
creation of works whose scope comfortably exceeds anything that an individual or small group of dedicated professionals could produce.\textsuperscript{78} Crowdsourcing can be viewed as a force multiplier: companies and other entities can sometimes get far more work done by opening their projects to collaborative input than they could have accomplished solely through the efforts of their own employees.\textsuperscript{79}

Certain types of works lend themselves more easily to crowdsourced production than others. The world still awaits the first peer-produced hit song, blockbuster film, or acclaimed novel. Nevertheless, the crowdsourced approach has proved its value in the creation of an informational commons: a wide variety of informational goods have been created through the internet-mediated efforts of a distributed community of volunteers.\textsuperscript{80} The success of peer-production projects in the information economy necessarily raises the question whether an archetypal informational good—the library—might be created through similar means.

\textbf{B. Crowdsourced Library-Building}

The Library of Congress presently holds a collection of nine billion texts that exist in paper-only form.\textsuperscript{81} The National Archives has estimated that, by working steadily at an expected pace of 500,000 texts a year, it can digitize those nine billion paper records in 1,800 years.\textsuperscript{82} The figure is, one suspects, purposefully outlandish; it is difficult to imagine any human endeavor (outside, perhaps, the realm of religion) that can be sustained over such a gulf of time. If the numerator—the number of texts to be digitized—cannot be changed, perhaps we can focus on the denominator—the number of texts digitized per year. If the National Archives, working alone, can digitize half a million texts a year, then perhaps it should not work alone. Enlarging the pool of contributors who are working to digitize

\begin{footnotesize}
\begin{enumerate}
\item See \textit{WEBER}, supra note 73, at 59 (noting complexities of software development process that tend to place relatively inflexible limits on what can be accomplished by one or two programmers).
\item The entry for “crowdsourcing” in the Wikipedia encyclopedia includes an interesting list of crowdsourced projects involving a wide variety of private and public entities. http://en.wikipedia.org/wiki/Crowdsourcing (last visited Jan. 5, 2010).
\item See \textit{Armstrong}, supra note 76, at 361 (collecting examples).
\item See \textit{Hafner}, supra note 3.
\end{enumerate}
\end{footnotesize}
historical texts and make them freely available online would appear to be one obvious solution.

1. Distributed Proofreaders and Project Gutenberg

Project Gutenberg is one of the oldest digital library projects, having been launched in 1971 at the University of Illinois. In the decades since, founder Michael Hart and many other Project Gutenberg contributors have made tens of thousands of books available for online browsing or download in a variety of formats.

In 2000, a group of Project Gutenberg contributors launched a companion Web site, Distributed Proofreaders ("DP"), with the aim of using collaborative techniques to help expand the library of texts available through Project Gutenberg. Their efforts have paid handsome dividends; indeed, as of the time of this writing, a majority of all texts available through Project Gutenberg were contributed via Distributed Proofreaders.

Distributed Proofreaders hosts scanned images—that is to say, pictures—of the pages of new texts that are candidates to be added to Project Gutenberg. Registered users of the site may select a text of interest to them from the listing of currently active proofreading projects. The site then displays to the user a split-screen window showing both a scanned image of the selected page of the work and the corresponding text that appears on that page (generated initially via optical character recognition (OCR) software). Because the uncorrected OCR output frequently contains errors, the text in the lower portion of the split-screen display may not exactly match the

---


84. The total exceeds 31,000 titles at the time of this writing. http://www.gutenberg.org/dirs/GUTINDEX-2010.txt (last visited Feb. 10, 2010).


86. The listing of texts that have achieved "Completed" status on Distributed Proofreaders (indicating that they have passed through all stages of the site's multi-step proofreading process) included over 17,000 titles at the time of this writing. See DP: Complete Gold E-Texts, http://www.pgdp.net/c/list_etexts.php?x=g&sort=5 (last visited Feb. 10, 2010).

87. See Figure 1, infra, at 31.
page image in the upper portion. Users of the site provide the human judgment that is necessary to match the text to the scanned page image and make any necessary corrections in the text box. The Distributed Proofreaders site provides special instructions concerning how users should mark punctuation and special characters that appear in the scanned page image.88

Proofreading at Distributed Proofreaders proceeds in multiple stages, with each stage representing progressively greater progress towards a completed text that multiple persons have verified against the scanned source images.89 First, each document goes through at least two, and optionally three, proofreading rounds, designated “P1” through “P3” in the nomenclature of the site.90 At the P1 round, users correct the raw OCR output to match the appearance of the corresponding scanned page image. The P2 and P3 rounds, in turn, take as their input the corrected text produced during the P1 and P2 rounds, respectively. After all the proofreading rounds have been completed, the document proceeds through two formatting rounds (“F1” and “F2”), where the goal is to check the proofread text to make sure that the visual appearance (not merely the text) mimics the scanned original. There is a final, optional, smooth reading (“SR”) round aimed at ensuring that the final digitized text has been correctly transcribed and formatted.91

Unlike wiki-based projects, not all users of Distributed Proofreaders may participate in each of the site’s activities. Distributed Proofreaders limits users’ eligibility to engage in various proofreading activities according to whether the user has created an account and the user’s prior history with the project.92 Access to all the later proofreading and formatting stages is granted only via application to the site’s administrators based upon certain eligibility criteria, as follows:

- Unregistered users of the site may participate only in SR rounds.93


89. A diagram illustrating the full workflow of a Distributed Proofreaders project (including preparation and post-processing activities that occur largely “behind the scenes,” invisible to ordinary users of the site) is available at http://www.pgdp.net/c/faq/DPflow.php (last visited Feb. 10, 2010).

90. See id.

91. See id.

92. See id.

93. See id.
- Registered DP users initially may participate only in P1 proofreading rounds. \[94\]

- To become eligible to participate in P2 proofreading rounds, a user must (1) complete 300 proofread pages at the P1 level, (2) be a member of the site for at least 21 days, and (3) pass a five-part proofreading quiz aimed at testing the user’s familiarity with the proofreading markup conventions used at the site. \[95\]

- To become eligible to participate in F1 formatting rounds, a user must (1) complete 300 proofread pages at the P1 level, (2) be a member of the site for at least 21 days, and (3) pass a five-part formatting quiz aimed at assessing the user’s familiarity with DP’s formatting markup. \[96\]

- To become eligible to participate in P3 proofreading rounds, a user must proofread a total of 400 pages, of which at least 50 must be from a “P2” round, (2) must complete a further 50 pages in an “F1” formatting round, and (3) must also pass a proofreading quiz. \[97\]

- Finally, to become eligible to participate in F2 formatting rounds, a user must (1) complete 400 pages in an “F1” formatting round, and (2) be a member of the site for at least 91 days. \[98\]

Regardless of the level of access they have attained, ordinary users of Distributed Proofreaders are unable to add new texts to the site. The ability to upload new scanned images and to initiate new proofreading projects is reserved for DP Project Managers, a status that must be granted separately by the site’s administrators. \[99\] The requirements for a DP user to be given Project Manager status include: (1) familiarity with the site’s workflow process and guidelines; (2) a minimum of 400 pages completed at the “F1” level; and (3) identification of an existing Project Manager who is willing to serve as a mentor. \[100\]

---

94. See id.
100. See id. The site also recommends, but does not require, DP membership for a period of 4–6 months as a condition of Project Manager status. Id.
Professor Yochai Benkler rightly celebrates Distributed Proofreaders as a paradigmatic crowdsourcing success story, and others have emphasized the site's potential value as a tool for crowdsourced library-building. The numbers are difficult to quarrel with. In the decade since its founding, volunteers coordinating their efforts through Distributed Proofreaders have proofread and released in electronic form (through Project Gutenberg) over 17,000 texts. DP's strengths include a large and supportive user community (with over 3,000 contributors active in the last 30 days at the time of this writing) and a rapid proofreading process, with completion times even for lengthy works measured in weeks (at least in the early rounds).

Nevertheless, a number of structural weaknesses may limit DP's utility as a tool for improving access to primary legal source materials. Unlike many of the most vibrant peer-produced informational projects, DP maintains a bureaucratic, hierarchical structure, with site administrators adjudicating users' compliance with the site's daunting criteria for promotion to higher levels of access, and all but the most senior users are disabled entirely from contributing new works. Furthermore, DP's selection of texts is driven by the philosophy underlying its senior partner, Project Gutenberg, which expressly aims to maximize the inclusion of texts popular with a mass audience. In consequence, Distributed Proofreaders and Project Gutenberg include comparatively few texts of interest to the legal community. Project Gutenberg's mission discourages the addition of such texts, and the DP architecture makes it difficult even for interested users inclined to do so. This hinders efforts to broaden the scope of the project's coverage.

101. See, e.g., BENKLER, supra note 4, at 81; Benkler, supra note 72, at 398--99.
102. See Travis, supra note 44, at 784("[c]ommons-based peer production has created what is arguably the largest and most successful digital library, and in a remarkably speedy, efficient, and user-friendly way.").
103. See supra note 86.
104. See supra notes 93--100 and accompanying text.
105. See supra notes 93--100 and accompanying text.
106. As Project Gutenberg's founder explained:
Project Gutenberg selects texts targeted a bit on the "bang for the buck" philosophy . . . we choose texts we hope extremely large portions of the audience will want and use frequently. We are constantly asked to prepare texts from out of print editions of esoteric materials, but this does not provide for usage by the audience we have targeted, 99% of the general public.
Hart, supra note 83.
2. Crowdsourcing the Wiki Way

Of the nine wiki projects operated by the nonprofit Wikimedia Foundation ("WMF"), one—Wikipedia—has garnered most of the scholarly praise and criticism. WMF’s other projects (Wikibooks, Wikinews, Wikiquote, Wikisource, Wikiversity, Wiktionary, and the Wikimedia Commons) have their own communities of dedicated users, who use a common set of wiki-based tools to contribute content within the scope of their respective missions. They have so far failed, however, to capture the academic imagination in quite the same way as Wikipedia—which, like an open flame, seems to have the power to draw all the oxygen out of academic discourse on law and wiki technologies. This is unfortunate, because WMF’s projects include another candidate that shares many of Wikipedia’s strengths, omits its most prominent weaknesses, and offers a natural fit with the interests and concerns of academics and others who study and value the public domain. That project is Wikisource.


111. At the 2010 Annual Meeting of the AALS Section on Internet and Computer Law where this paper (along with three others) was presented, Professor Paul Ohm observed that two of the four pieces presented by the members of the panel focused their analysis entirely on Wikipedia, signifying that perhaps the chosen theme of the Section’s panel (“Law and Wikis”) should have been revised to “Law and Wikipedia.” As discussed below, Wikipedia is, by a vast margin, the largest of WMF’s many projects (comfortably larger, indeed, than all the other WMF wikis combined), and may attract disproportionate attention for that reason alone. See infra Table 1, at 29.

Professor Ohm’s casual observation seems to have some empirical foundation. A search for “Wikipedia” in Westlaw’s JLR database on February 15, 2010 returned 2,258 “hits,” compared with just 30 for “Wikiquote,” 14 for “Wikitomy,” 10 for “Wikinews,” 8 for “Wikibooks,” 5 for “Wikimedia Commons,” 4 for “Wikiversity” (of which two appeared to be duplicates of one another), 3 for “Wikispecies” (with the same duplicates), and just 2 for “Wikisource.”

112. Like all of WMF’s wikis, Wikisource consists of not one project, but many, each serving the needs of speakers of a particular language. The home page of the English-language version of Wikisource is online. See generally Wikisource Homepage, http://en.wikisource.org/ (last visited Aug. 13, 2009). At present, the English-language Wikisource library is, by several
Wikisource is **a digital library of previously published free-content works**. The project's eligibility criteria are strict; only works that are in the public domain in the United States or are licensed under terms that allow free copying, modification, and reuse (including commercial use) are permitted to be hosted on the site.\(^\text{11}\) The requirement of prior publication is intended to ease verification (that is, to make it possible for the site's geographically far-flung users to confirm that the text posted at the site matches the published original) and to deter misuse of the site for self-publication.\(^\text{14}\)

Wikisource's mission differs from Wikipedia's in ways that tend to insulate it against some of the criticisms often aimed at its larger sibling. Wikipedia's stated goal is to describe the world from a *neutral point of view*\(^\text{15}\)—a goal that may be epistemologically unattainable,\(^\text{16}\) and at a minimum invites ongoing debate over the "neutrality" of articles published on the site.\(^\text{17}\) Wikisource's polestar, in contrast, is not neutrality, but *faithful reproduction of a source text as published*.\(^\text{18}\) It is easy to imagine users reasonably holding different units of measure, the largest; the top ten Wikisource libraries are listed *infra* Table 2, at 30.

\(^\text{11}\) For Wikisource's complete inclusion policy, see Wikisource: What Wikisource includes, http://en.wikisource.org/wiki/Wikisource:What_Wikisource_includes (last visited Feb. 10, 2010) [hereinafter "What Wikisource includes"]. The site also provides guidance as to which non-public domain works are sufficiently "free" to qualify for inclusion; for example, works issued under a simple Creative Commons Attribution license would qualify, but works issued under the more restrictive Attribution-Non Commercial-No Derivatives license would not. See Wikisource: Copyright Policy, http://en.wikisource.org/wiki/Wikisource:Copyright_policy (last visited Feb. 10, 2010).

\(^\text{14}\) See *What Wikisource includes*, *supra* note 113.


\(^\text{16}\) See, e.g., BENKLER, *supra* note 4, at 70-71 ("An effort to represent sympathetically all views on a subject, rather than to achieve objectivity, is the core operative characteristic of this effort."). Wikipedia is a uniquely self-critical work, and lengthy discussions of the practical difficulty of achieving the site's objective of substantive neutrality are easily located on the site itself. See generally *Wikipedia:Why Wikipedia is not so great*, http://en.wikipedia.org/wiki/Wikipedia:Why_Wikipedia_is_not_so_great (last visited Feb. 10, 2010).

\(^\text{17}\) At the time of this writing, over 6,500 Wikipedia articles have been flagged as possibly violating the site's neutrality principle, with most of them flagged as problematic for over a year. See *Category:NPOV disputes*, http://en.wikipedia.org/wiki/Category:NPOV_disputes (last visited Feb. 10, 2010).

\(^\text{18}\) Issues involving "neutrality" do sneak in through the back door at Wikisource under the heading of *completeness*; that is to say, a user's faithful reproduction of only a nonrepresentative or misleading excerpt from a work may prompt other users to add the rest of the work to provide necessary context. See *Wikisource:What is Wikisource?*, http://en.wikisource.org/wiki/Wikisource:What_is_Wikisource%3F
differing opinions about whether the Wikipedia biographies of Presidents Barack Obama\textsuperscript{119} or George W. Bush\textsuperscript{120} adhere to the stated standard of neutrality; it is less easy to imagine users reasonably adhering to different views about whether the text reproduced at Wikisource matches the content of the published source.\textsuperscript{121}

Like Distributed Proofreaders, Wikisource now draws most new content from users who proofread and correct the text extracted from scanned page images of a published source.\textsuperscript{122} Unlike Distributed Proofreaders, however, Wikisource was not originally engineered with proofreading of page scans in mind. This functionality has been in place only during the last two to three years of the project’s existence.\textsuperscript{123} Nevertheless, the site now offers a clean and well-organized user interface that at least rivals, and perhaps exceeds, the usefulness and intuitive functionality of Distributed Proofreaders.

First, each scanned volume image accessible at Wikisource\textsuperscript{124} (which typically, although not always, correspond to a separately bound hard copy volume of a work as originally published) has a so-called “Index page” that reproduces identifying information about that


\textsuperscript{121} This risk seems particularly low for works more recently added to Wikisource, many of which are assembled from scanned page images of the published original sources and which permit easy user verification of the text against the source image.

\textsuperscript{122} Statistical descriptions of Wikisource (or, indeed, any of the WMF projects) involve substantial risks of error due to the constant flux of additions and deletions to the project. With that caveat in mind, however, it is possible to make some very broad points to illustrate the relative magnitude of the works available at the English-language Wikisource. As of January 2010, Wikisource included approximately 321,000 individual pages of scanned text—a figure that may undercount the actual number of scanned images available at the site, not all of which have yet been used to produce a corresponding text page. See Wikisource Statistics—Tables—English—Database records per namespace, http://stats.wikimedia.org/wikisource/EN/TablesWikipediaEN.htm#namespaces (last visited Feb. 10, 2010) (the column heading “104” in this table corresponds to the “Page” namespace used on the project and marks the number of text records that match a scanned page image at Wikisource).


\textsuperscript{124} A few of these files are hosted at Wikisource itself, although the more common practice appears to be to host the files at Wikimedia Commons, where they are equally usable by all WMF projects.
volume as a whole—such as the title, author, publisher, year of publication, and possibly a table of contents. The volume’s Index page also includes links to each individual page contained within the volume. Each page link is color-coded using a standard schema that applies site-wide and reflects, in essence, the level of confidence of the project’s users that the text reproduced at that link accurately reflects the content of the corresponding scanned page. Thus, the Index page reveals at a glance how much progress the site’s users have made towards finalizing the proofreading and correction of the work. The color codes used on the site are:

- **Red** ("Not Proofread"): Signifies that the linked page contains text, but no user of the site has checked the text for accuracy. This color code is typically applied where the text included on the linked page consists entirely of the raw output of OCR software.

- **Yellow** ("Proofread"): Signifies that one user of the site has proofread and corrected the linked text so that it matches the content and formatting of the corresponding scanned page image.

- **Green** ("Validated"): Signifies that two or more users of the site have proofread and corrected the text of the linked page. This is the highest rating of page quality available on Wikisource.

In addition, there are three further color codes used on the site that convey additional information about the status of the corresponding linked page:

- **Purple** ("Problematic"): Signifies that the text on the linked page does not match the scanned original due to an error in the scanned image (such as a blurry or misaligned page), or because the content

---


127. See id.

128. See id. At the time of this writing, the number of pages that had reached each quality tier on the English Wikisource were: Not Proofread, 252,667; Proofread, 36,582; Validated, 15,190. The author of the present Essay is partly to blame for the predominance of pages consisting entirely of raw OCR output, having personally uploaded some 70,000 such pages to the site using automated scripts. See infra notes 143–148 and accompanying text (discussing project to host the United States Statutes at Large on Wikisource).
of the scanned page cannot be accurately reproduced for some other reason.\textsuperscript{129}

- **Gray** ("Unnecessary to Proofread"): Signifies that the corresponding page either is blank, or contains some content other than text (for example, an image or illustration).\textsuperscript{130}

- **White** ("Empty Page"): Signifies that, although the linked page includes a scanned image of the original source, no corresponding text of any kind exists yet on Wikisource. The site offers easy ways for users to fill in empty pages (generally upgrading the corresponding link from "Empty" to "Not Proofread" in the process), either by extracting text embedded in the image file, or by running an on-site OCR tool on the image.\textsuperscript{131}

The volume index page for a given work available for proofreading on Wikisource thus can appear, at any given moment, as an information-rich (and colorful) mosaic, instantly reflecting the validation level of each page included within that volume.\textsuperscript{132}

Clicking on any of the individual page links from the Index page opens a page-level view where proofing and correction of the text actually occurs.\textsuperscript{133} In an improvement over the Distributed Proofreaders interface, Wikisource displays the extracted text and the corresponding scanned page image side by side.\textsuperscript{134} Clicking on the "Edit" tab at the top of the page displays a scrolling text box side-by-side with the scanned image of the page.\textsuperscript{135} Users may enter any necessary corrections in the text window and update the contents of the page by clicking the "Save" button. If the user indicates a change in the page's overall proofreading level (by clicking an adjacent radio button for whichever color coding is appropriate), the color of the

\textsuperscript{129} See Help:Page Status, supra note 126.

\textsuperscript{130} See id.

\textsuperscript{131} See id. The availability of "Empty" pages on the site—for which an image, but no corresponding text, exists—explains why the existing Page statistics understate the true dimensions of Wikisource. See supra note 122.

\textsuperscript{132} See Figure 2, infra, at 32. The site’s architecture thus gives the proofreading process some aspects of a rudimentary video game, with users’ proofreading activities yielding progressive “rewards” in the form of perceptible changes in the appearance of the work’s index page. This feature is, if nothing else, an ingenious way of encouraging sustained user involvement, even if the makers of World of Warcraft or other actual video games probably have little to fear from the competition posed by Wikisource.

\textsuperscript{133} See Figure 3, infra, at 33.

\textsuperscript{134} See id.

\textsuperscript{135} See Figure 4, infra, at 34. Depending on the resolution of the scanned source image, the user may “zoom in” for a closer view of the image to ease proofreading of small text.
corresponding link from the volume’s index will automatically be updated to reflect the changed status of that page.136

As with Distributed Proofreaders, when the scanned pages of a work have been proofread to a satisfactory quality level, the proofread text of all (or some) of the pages within the work can be joined together to form a single electronic file of the complete proofread text. Unlike Distributed Proofreaders, however, this process resides entirely within the control of the users of the site and requires no additional software.137 Subject only to certain technical constraints imposed by the underlying architecture, the corrected text from dozens or hundreds of scanned original pages may be automatically joined together to form a single Web page with the complete text of the entire original work. The common practice on Wikisource is to keep the scanned page images available even after proofreading is complete in order to ease authentication138; for most works recently added to the site, users may verify for themselves (by clicking a small page number link that typically appears in the margin of the displayed text) that the text displayed at the site matches the content of the scanned page image.139

Wikisource, unlike Distributed Proofreaders, is a wiki: with the exception of a small number of “locked” pages, any user of the site may add or edit any work in the library.140 Thus, Wikisource removes some of the obstacles that make Distributed Proofreaders and Project Gutenberg unpromising candidates for hosting source materials of interest to the legal community. Indeed, the barriers to adding a new work to Wikisource are exceptionally low: so long as a user can locate (or create) an electronic file containing scanned images of the source work as originally published—and there are many scanned legal texts already available online at sites such as Google Books or the Internet Archive—the only indispensable step consists of uploading a set of suitable scans to Wikimedia Commons where it will be accessible by Wikisource.141 Every other step of the process—creating an index page, extracting (or creating) uncorrected OCR text from the scanned file, proofreading and correcting the text, and

136. See Figure 2, infra, at 32.
137. Cf. supra note 90.
138. See, e.g., infra note 149 and references cited therein.
139. See id.
141. See supra note 124.
joining the corrected text pages together to form a consolidated e-text—can be crowdsourced.

The Wikisource process has already been used to make some texts of interest to the legal community freely accessible online. Indeed, Wikisource now hosts some texts that are not yet freely available anywhere else, such as key portions of the legislative history for the landmark Copyright Act of 1976. In an ongoing experiment to use the site to expand the availability of historical materials, the present author made over 70,000 pages of scanned images (taken mostly from the Library of Congress’s outstanding American Memory project) and raw OCR text, representing the first forty-three volumes of the *United States Statutes at Large*, available for proofreading and correction on Wikisource. At the time of this writing, all the public and private laws and resolutions of the First United States Congress, which sat in three sessions from March 4, 1789 to March 3, 1791, have been proofread and made publicly available by users of the site. Other selected statutes and proclamations within the scanned collection have also been proofread by users with an interest in particular issues or periods in American legal history. The material proofread to date represents a very small fraction of the full dataset comprising the early volumes of the *Statutes at Large*. Nevertheless, sufficient progress has occurred so


143. See supra note 69.

144. This portion of the *Statutes at Large* represents nearly a century and a half of American statutory law (1789–1925), as well as early treaties, Presidential proclamations, and the first version of the *Revised Statutes* (which would grow in time to become the work we now know as the United States Code).


146. For example, some Wikisource editors have proofread all four of the so-called “Alien and Sedition Acts” passed by Congress in 1798 and made the proofread text available on Wikisource, with links to additional explanatory content hosted on Wikipedia. See *Alien and Sedition Acts*, http://en.wikisource.org/wiki/Alien_and_Sedition_Acts (last visited Apr. 17, 2010). Other Wikisource users have been proofreading and posting the proclamations of President Theodore Roosevelt in essentially chronological order. See United States Statutes at Large, Vol. 33, Part 2, http://en.wikisource.org/wiki/Index:United_States_Statutes_at_Large_Volume_33_Part_2.djvu (last visited Apr. 17, 2010) and pages linked therefrom.

147. The most complete volume at present is Volume 1, with approximately 25% of the pages proofread as of the time of this writing. See United States Statutes at Large Vol. 1,
far to at least demonstrate the viability of crowdsourced proofreading of legal texts, specifically works that would be unlikely to be included at Distributed Proofreaders. Wikisource can also serve as a repository for legal scholarship that meets the site’s inclusion criteria—thus potentially bringing together scholarship and primary source materials in a way not presently replicated by any other open-access repository.

By virtue of its design, Wikisource comports with many (although certainly not all) of Professor Ian Gallacher’s proposed design standards for open-access archives of primary legal source materials. Wikisource’s collection is universally accessible worldwide. It can be presented in a variety of formats (or downloaded freely and further processed to meet a user’s specific presentation needs), and its contents are open to indexing by Google or other standard search engines. The output format of any work hosted on Wikisource is an XHTML web page, an open vendor-neutral format that nevertheless enables preservation of a great deal of the original work’s formatting. The Wikimedia Foundation’s globally distributed server architecture yields adequate response speeds in ordinary use. The site offers permanence in the form of downloadable snapshots of the full database as it existed at various points in time; if

http://en.wikisource.org/wiki/United_States_Statutes_at_Large_Volume_1.djvu (last visited Apr. 17, 2010). Clicking the volume links for any of the other scanned volumes in the Statutes at Large (all of which are linked from the page for Volume 1) will reveal the overwhelming predominance of page links that appear against a red background, signifying “not proofread.” See supra note 126.

In addition to the sheer size of the dataset (the Statutes at Large scans alone presently account for over 20% of all the scanned pages available at Wikisource), the process of proofreading and correction is doubtless slowed by (1) the complex, multi-column page format employed in the original work; and (2) the poor quality of the raw OCR output from the software employed to date, which necessitates substantial human effort to proofread and correct a single page. There is nothing inevitable or irremediable about either of these problems; more technologically skilled users of the site may, in time, identify common OCR errors that may be auto-corrected across many pages at once using search-and-replace scripts, or may apply improved OCR software to the stored page images to yield a better baseline text that may be proofread more rapidly.

See Timothy K. Armstrong, Fair Circumvention, 74 BROOK. L. REV. 1 (2008), http://en.wikisource.org/wiki/Fair_Circumvention. In the version of the article online at Wikisource, many citations to key statutes or cases appear as clickable hyperlinks that take the user directly to the work referenced by the citation. Links to explanatory content available on Wikipedia or other WMF wikis also appear throughout the document.

When viewing any page within Wikisource (or any of the other Wikimedia Foundation wikis, such as Wikipedia), using the “View Page Source” function within one’s web browser will indicate, in the <DOCTYPE> declaration on the first line of the page source, the type of document being viewed.

See supra note 41 and accompanying text.
Wikisource itself were to ever go offline (for example, if WMF were ever to become insolvent), the content of the site (except for the most recent edits) could be swiftly restored by anyone with a mirror copy of the most recent database dump.\footnote{152} A citator of sorts is available from the “What Links Here” link on every page of any WMF wiki, and the entire project is open to public development and maintenance.\footnote{153}

For the purpose of assessing its potential value as a possible open-access repository for legal source texts, Wikisource’s strengths include: (1) a well-developed and stable architecture that enables contributions by any user familiar with the standardized MediaWiki editing syntax (which is much easier to learn than HTML); (2) the openness of its database, which any user may edit or expand; (3) the relative sophistication and user-friendliness of the site’s user interface; (4) the existence of a community of users within the site who are interested in legal topics and have already made several legal source texts available; and (5) the ease of authentication provided by the site’s use and preservation of scanned page images from the original published sources. Wikisource’s most evident weaknesses stem from the comparatively small community of users of the site: by any measure, Wikisource is a tiny project compared with Wikipedia or Distributed Proofreaders.\footnote{154} The smaller number of users at the site translates into substantially greater time required to complete any given proofreading project and has also limited the number of texts that have been added to the site. Thus, Wikisource remains very far from approaching Professor Gallacher’s ideal of \textit{completeness} for an open-access repository. Nevertheless, Wikisource offers an interesting alternative to Distributed Proofreaders as a platform for mass collaboration in making a variety of works freely available to the

\footnote{152. The contents of all WMF wikis, including Wikisource and Wikipedia, are available for download at http://download.wikimedia.org/ (last visited Feb. 12, 2010).}

\footnote{153. \textit{See supra} notes 140–141 and accompanying text.}

\footnote{154. \textit{See Table 1, infra, at 29; compare supra} note 104 and accompanying text. Measured by the number of users who have participated at each project during the last thirty days, Wikisource is approximately one-tenth the size of Distributed Proofreaders, and barely one five-hundredth the size of Wikipedia. Wikisource’s small size and the relatively recent redesign of the site’s architecture to facilitate proofreading have also meant restricted throughput of works. As noted above, Distributed Proofreaders has completed over 17,000 texts, while the comparable statistic for Wikisource (consisting of those works that have achieved a quality level of “Validated” on all their pages) is only slightly over 100 texts at the time of this writing. \textit{See Category:Index Validated, at} http://en.wikisource.org/wiki/Category:Index_Validated (last visited Feb. 12, 2010).}
public, and the successes the site’s users have achieved to date offer a hint of its promise.

C. Pros and Cons

Crowdsourcing methods represent territory mostly unexplored by the various projects currently working to provide open access to legal source materials. The “peer production” approach, which has been ably used to create a wide variety of other informational goods, holds at least some promise as a tool for making legal and historical materials available more widely and without restriction.

Most fundamentally, crowdsourcing techniques alleviate resource constraints that otherwise limit the scope and operations of typical open-access efforts. Many of the organizations that have launched legal open-access sites are arms of educational or nonprofit institutions, and their reach is constrained by available resources. Inviting interested members of the legal community and the public to collaborate in building a free commons of legal source materials removes the resource constraints of any single organizing entity as a limiting factor.

Reduced organizational overhead is a second identifiable benefit of crowdsourcing. As the example of the Statutes at Large illustrates, launching a new crowdsourced open-access initiative is a project within the means of a dedicated individual acting on his or her own initiative. The need to build committee structures or to lobby for consensus-based decision-making is not an impediment; texts within any single user’s areas of interest and expertise may be added to a project almost effortlessly, with other users of the site free to contribute as their own interest and curiosity dictates.

The wiki-based architecture of a project like Wikisource offers another potential benefit in the form of reducing barriers to participation. Wikisource’s approach differs markedly from Distributed Proofreaders*: the open wiki-based architecture invites and facilitates participation by users of widely varying expertise. Some users may be competent in proofreading the scanned OCR text and marking rudimentary corrections, others may be knowledgeable about the MediaWiki formatting markup used across all the WMF sites, others may excel at categorization and indexing, and still others may have the type of skills that are necessary to program templates or

155. See supra notes 49–58 and accompanying text. The Google Scholar case-law service may be that rare open-access project developed with minimal constraints as to resources. See supra notes 59–64 and accompanying text.
scripts for managing more complex tasks. The architecture of the site permits users to contribute according to their respective expertise.\textsuperscript{156}

Like any other organizational tool, however, crowdsourced methods have weaknesses as well as strengths. Goal-setting and prioritization of work is a recurring issue for projects situated outside any formal organizational structure. For example, Wikisource, like Wikipedia or any number of similar open-content projects, has no "benevolent dictator" assigning tasks or ensuring that user effort flows to where it is most needed.\textsuperscript{157} User contributions are largely self-directed towards those areas where their interests happen to gravitate.\textsuperscript{158} Although some users diligently perform work (such as archiving past discussions and rationalizing the site's frequently confusing categorization structures) that improves the quality and usefulness of the site overall, most users appear to be focused on expanding the library by adding new content. In consequence, Wikisource is an unruly patchwork, with comparatively stable and well-organized content existing alongside fragmentary works organized only according to the idiosyncratic whim of a particular contributor.

The problem of sustaining user engagement over time in the absence of traditional incentives (such as the payment of a salary) is also endemic to many crowdsourced projects. Users of peer-produced projects are free to come and go, and there is no guarantee that a user who launches any given project will see it through to completion.\textsuperscript{159} Although Wikipedia seems not to be in any imminent danger of

\begin{footnotes}
\item[156] Of course, making it possible for a wide variety of users to contribute irrespective of expertise may not represent an unalloyed blessing. A certain portion of the editing work on a site like Wikisource necessarily involves correcting erroneous contributions made by inexpert users of the site, although the benefit to allowing such users to participate and thereby to acquire greater familiarity with the site's tools and culture surely outweighs the occasional need to undo mistaken or malicious edits.

\item[157] For examples of open-source development projects that do employ management structures guided by a "benevolent dictator," at least to help make final decisions about which contributions will be accepted into the project, see ERIC S. RAYMOND, Homesteading the Noosphere and The Magic Cauldron, in THE CATHEDRAL AND THE BAZAAR, supra note 73, at 79, 124-26.

\item[158] This characteristic is typical of the open-source approach to development of expressive content. See ERIC S. RAYMOND, Homesteading the Noosphere, in THE CATHEDRAL AND THE BAZAAR, supra note 73, at 100-02 (explaining open-source software development as driven, at least in part, by the satisfaction users derive from practicing skills in areas of personal interest to them); Weber, supra note 73, at 62 ("The key element of the open source process, as an ideal type, is voluntary participation and voluntary selection of tasks.").

\end{footnotes}
failing,\textsuperscript{160} it is hardly difficult to locate examples of essentially moribund projects on Wikisource or any of the smaller WMF sites. In contrast, the Distributed Proofreaders architecture channels user participation by requiring users to select from a small number of currently ongoing projects if they wish to participate, and may actually provide some structural advantages here.\textsuperscript{161}

V. CONCLUSION

Despite remarkable successes in the past fifteen years or so, no open-access project for primary legal source materials approaches the size and sophistication of the large proprietary legal databases. Proprietary database publishers benefit from an inflow of subscriber revenues that no open-access project can hope to match;\textsuperscript{162} indeed, the whole point of the open-access movement is to provide an alternative to the proprietary subscriber-access paradigm and make information freely accessible to all.

A variety of high-quality informational goods have been produced using nonproprietary production processes that aggregate the individual contributions of a wide community of volunteers. As a matter of principle, there is no reason why such a crowdsourced production process might not be employed to extend access to legal materials and scholarship. The technological architecture for building new open-access projects in the legal arena is already in place; all that is missing is a sufficiently large pool of contributors willing to assist in building the informational commons as their interests and abilities permit.

To maximize the overall benefit to the information commons, crowdsourced projects should aim to supplement rather than to supplant existing open-access repositories for legal works. New projects should aim at building strength in areas where existing repositories are weak: they should focus more on legislative and executive materials rather than case law, and more on historical rather


\textsuperscript{161} It is difficult to know whether the larger active user base at Distributed Proofreaders reflects superior architecture, or simply longer existence; after all, Distributed Proofreaders has been around for a decade, and may also draw contributions from users interested in Project Gutenberg, a nearly four-decade-old project.

than contemporary works. Adding contextual richness with hyperlinking and verifiability against official sources will make such projects more attractive for everyday use and provide a practical alternative to proprietary legal databases.
Table 1. Relative Size of All English-Language Wikimedia Foundation Wikis as of January 2010, ordered by size of database

<table>
<thead>
<tr>
<th>Site</th>
<th>Content Pages</th>
<th>Registered Users</th>
<th>Active Users</th>
<th>Database Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>3,160,178</td>
<td>11,416,483</td>
<td>143,781</td>
<td>10.0 GB</td>
</tr>
<tr>
<td>Wikimedia Commons</td>
<td>94,283</td>
<td>984,709</td>
<td>21,127</td>
<td>1.5 GB</td>
</tr>
<tr>
<td>Wikisource</td>
<td>138,336</td>
<td>179,090</td>
<td>319</td>
<td>875.6 MB</td>
</tr>
<tr>
<td>Wiktionary</td>
<td>1,557,899</td>
<td>312,418</td>
<td>1,068</td>
<td>441.8 MB</td>
</tr>
<tr>
<td>Meta-Wiki</td>
<td>17,017</td>
<td>412,387</td>
<td>3,471</td>
<td>346.3 MB</td>
</tr>
<tr>
<td>Wikibooks</td>
<td>38,628</td>
<td>313,018</td>
<td>623</td>
<td>188.1 MB</td>
</tr>
<tr>
<td>Wikiquote</td>
<td>17,587</td>
<td>219,124</td>
<td>557</td>
<td>125.8 MB</td>
</tr>
<tr>
<td>Wikinews</td>
<td>16,197</td>
<td>183,048</td>
<td>286</td>
<td>79.7 MB</td>
</tr>
<tr>
<td>Wikiversity</td>
<td>11,808</td>
<td>194,352</td>
<td>427</td>
<td>56.0 MB</td>
</tr>
<tr>
<td>Wikispecies</td>
<td>208,839</td>
<td>144,062</td>
<td>175</td>
<td>45.8 MB</td>
</tr>
</tbody>
</table>

NOTES TO TABLE 1:

a All references are to the English-language versions of the listed sites (so, the statistics for Wikipedia, for example, are those of en.wikipedia.org), with the exception of Wikimedia Commons. The Commons is a cross-language repository used by all Wikimedia Foundation sites to store graphic images, audio or video clips, and scanned page images that are intended for use at any Wikimedia Foundation site.

b The figures presented in the next three columns are taken from the Statistics page of each indicated site, which may be accessed by typing Special:Statistics into the Search box on each site’s home page. The data in these columns are current as of January 15, 2010.

c The figures in this column also come from the Statistics page of the indicated site and are current as of January 15, 2010. Although all the WMF sites permit editing by users who have not registered and created an account, certain practical advantages accrue from registration. The figures listed in this column represent the total
number of users who have registered and created an account on the indicated site.

d The figures in this column also come from the Statistics page of the indicated site and are current as of January 15, 2010. The figures represent the number of registered users who have edited the site within the preceding 30 days.

e Snapshots of the complete database of each of the listed sites are made available for download periodically at download.wikimedia.org. Snapshots are not prepared for each site according to the same schedule; thus, it is not possible to compare the size of each of the listed sites as of a single common date. The relative sizes of the download archives, however, is generally reflective of the quantity of content available at each site listed. The sizes listed represent the size of the complete database (with all editing history intact) as a single compressed file, in gigabytes (GB) or megabytes (MB), as indicated. The quoted figures are taken from the following snapshot dates: Oct. 22, 2009 (Wikipedia); Jan. 8, 2010 (Wikimedia Commons); Jan. 10, 2010 (Wikinews and Meta-Wiki); Jan. 11, 2010 (Wikisource); Jan. 12, 2010 (Wikiversity); and Jan. 13, 2010 (Wiktionary, Wikibooks, Wikispecies, and Wikiquote).
Table 2. List of Ten Largest Wikisource Libraries, as of January 2010, ordered by number of content pages

<table>
<thead>
<tr>
<th>Language</th>
<th>Abbr</th>
<th>Content Pages</th>
<th>Registered Users</th>
<th>Active Users</th>
<th>Database Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>en</td>
<td>137,999</td>
<td>178,593</td>
<td>317</td>
<td>875.6 MB</td>
</tr>
<tr>
<td>Chinese</td>
<td>zh</td>
<td>99,529</td>
<td>12,806</td>
<td>83</td>
<td>316.1 MB</td>
</tr>
<tr>
<td>Portuguese</td>
<td>pt</td>
<td>84,248</td>
<td>4,429</td>
<td>29</td>
<td>48.1 MB</td>
</tr>
<tr>
<td>Russian</td>
<td>ru</td>
<td>76,169</td>
<td>10,240</td>
<td>104</td>
<td>151.5 MB</td>
</tr>
<tr>
<td>French</td>
<td>fr</td>
<td>53,125</td>
<td>11,596</td>
<td>153</td>
<td>637.1 MB</td>
</tr>
<tr>
<td>Spanish</td>
<td>es</td>
<td>45,800</td>
<td>10,885</td>
<td>60</td>
<td>111.0 MB</td>
</tr>
<tr>
<td>German</td>
<td>de</td>
<td>42,423</td>
<td>8,976</td>
<td>188</td>
<td>322.0 MB</td>
</tr>
<tr>
<td>Hebrew</td>
<td>he</td>
<td>30,823</td>
<td>2,477</td>
<td>48</td>
<td>77.3 MB</td>
</tr>
<tr>
<td>Italian</td>
<td>it</td>
<td>21,190</td>
<td>5,550</td>
<td>47</td>
<td>93.6 MB</td>
</tr>
<tr>
<td>Arabic</td>
<td>ar</td>
<td>20,013</td>
<td>4,341</td>
<td>22</td>
<td>153.7 MB</td>
</tr>
</tbody>
</table>

NOTES TO TABLE 2:

- The home page of each of the listed sites is accessible by prepending the two-character language abbreviation to the common suffix wikisource.org—thus, en.wikisource.org, zh.wikisource.org, and so forth.
- Database size figures are taken from download.wikimedia.org as explained in Note e to Table 1. The quoted sizes for each downloadable snapshot are taken from the following dates: Jan. 8, 2010 (fr, ar); Jan. 10, 2010 (pt, ru, de, he); Jan. 11, 2010 (en, zh); Jan. 12, 2010 (es); and Jan. 14, 2010 (it).
THE SOLITARY FARM

Terrified, as though dreadfully what she was to say and do. The room was filled with shadows, even though three lamps were lighted, and the silence became quite oppressive. Mrs. Vand was a healthy animal, and not in the least imaginative, but after a time she felt that some evil influence was in the room, and tightly gripped her husband's hand. The perspiration broke out on her forehead. Henry gave her no comfort, not even by pressing her hand. His eye were fixed on the perfectly expressionless and still face of Granny Tumple. The scene had all the elements of terror about it.

Figure 1. Proofreading User Interface at Distributed Proofreaders

Figure 2. A sample Wikisource index page
necessary to be raised by direct taxation is to be appropriated to the several states, according to the number of their inhabitants, and although Congress may raise the money by their own authority, if necessary, yet that authority need not be exercised, if such taxes shall furnish the means.

The expenses on the maintenance of the several states necessarily arising out of debts, paying any thing but money a debt is a pledge of faith, or impounding the obligations of contracts, by an act of Congress, was thought necessary as a security to commerce, in the manner of foreigners, as well as of the citizens of different states, may be.