2012

Orphan Works as Grist for the Data Mill

Matthew Sag

ORPHAN WORKS AS GRIST FOR THE DATA MILL

Matthew Sag[†]

ABSTRACT

The phenomenon of library digitization in general, and the digitization of so-called "orphan works" in particular, raises many important copyright law questions. However, as this Article explains, correctly understood, there is no orphan works problem for certain kinds of library digitization.

The distinction between expressive and non-expressive works is already well recognized in copyright law as the gatekeeper to copyright protection—novels are protected by copyright, while telephone books and other uncreative compilations of data are not. The same distinction should generally be made in relation to potential acts of infringement. Preserving the functional force of the idea-expression distinction in the digital context requires that copying for purely non-expressive purposes (also referred to as non-consumptive use), such as the automated extraction of data, should not be regarded as infringing.

The non-expressive use of copyrighted works has tremendous potential social value by making search engines possible, and by providing an important data source for research in computational linguistics, automated translation, and natural language processing. Furthermore, the macro-analysis of text is being increasingly used in fields such as the study of literature itself. So long as digitization is confined to data processing applications that do not result in infringing expressive or consumptive uses of individual works, there is no orphan works problem because the exclusive rights of the copyright owner are limited to the expressive elements of their works and the expressive uses of their works.

---

† Associate Professor of Law at Loyola University Chicago. Thanks to Pamela Samuelson for encouraging me to pursue this line of inquiry. Thanks also to Jerome Reichman, Matthew Jockers, and the participants at the 2012 Orphan Works and Mass Digitization conference at U.C. Berkeley School of Law. Please address comments to Matthewsag@gmail.com.

TABLE OF CONTENTS

I. **INTRODUCTION** 1504

II. **UNRAVELING THE DIGITIZATION DEBATE** 1506

A. LIBRARY DIGITIZATION-PRESERVATION 1507
B. LIBRARY DIGITIZATION-DISTRIBUTION 1508
C. LIBRARY DIGITIZATION-SEARCH 1511

III. **NON-EXPRESSIVE USE** 1512

A. COPYRIGHT, BALANCE, AND THE DISTINCTION BETWEEN IDEAS AND EXPRESSION 1512
B. NON-EXPRESSIVE USE 1517
1. *Coming to Grips with the Concept of Non-Expressive Use* 1517
2. *Examples of the Non-Expressive Use of Expressive Works* 1523
a) Internet Search Engines 1523
b) Plagiarism Detection Software 1524
c) Non-Expressive Use and Library Digitization 1525
C. THE SCOPE OF COPYRIGHT WITH RESPECT TO THE NON-EXPRESSIVE USE OF EXPRESSIVE WORKS 1528
1. *Substantial Similarity* 1528
2. *Intermediate Copying* 1530
3. *The Implications of Computer Software and Other Functional Works Protected by Copyright Law* 1532
D. ACTIVATING THE PRINCIPLE OF NON-EXPRESSIVE USE THROUGH FAIR USE 1533
1. *Why Fair Use* 1533
2. *Application to Fair Use* 1535
a) The "Purpose and Character" of Non-Expressive Uses 1535
b) Non-Expressive Use and Commercial Fair Use 1537
c) Non-Expressive Use and "Amount and Substantiality" 1538
d) The Market Effect of Non-Expressive Uses 1539

IV. **CONCLUSION: UNLEASH THE MACHINES** 1542

APPENDIX 1550

I. INTRODUCTION

Modern technology makes it possible for libraries to scan their paper collections and render them in digital form, making them more useful and

more available than ever before. Modern copyright law ensures that this scanning and digitization process is ensnared in a host of thorny issues.[1] Library digitization[2] has been rendered thornier still by Google's bold entry into the field in 2004, the ensuing litigation authors and publishers instigated, and the audacity of the class action settlement negotiated in 2008 (and revised in 2009) attempting to resolve that litigation.[3]

One of the main issues confronting libraries and others with respect to digitization is whether and how to clear rights with respect to works whose copyright owners are not easily found. The existence of these so-called orphan works is one of the most vexing issues in U.S. copyright law today.[4] One of the main benefits of the class action settlement proposed in relation to *Authors Guild v. Google* was that it would have constituted an expeditious resolution of the orphan works problem standing in the way of library digitization.[5] However, the treatment of orphan works proposed in the settlement was also one of the primary reasons that the court ultimately rejected it.[6]

This Article aims to untangle the orphan works thicket as it relates to library digitization and show that, correctly understood, *there is no orphan works problem for certain kinds of library digitization*. So long as digitization is confined to data processing applications that do not result in infringing expressive or

1. There is a large literature on library digitization in general and the Google book search litigation. *See, e.g.*, Emily Anne Proskine, *Google's Technicolor Dreamcoat: A Copyright Analysis of the Google Book Search Library Project*, 21 BERKELEY TECH. L.J. 213 (2006); Jonathan Band, *The Long and Winding Road to the Google Books Settlement*, 27 J. MARSHALL REV. INTELL. PROP. L. 227 (2009); James Grimmelmann, *D is For Digitize Symposium: An Introduction*, 55 N.Y.L. SCH. L. REV. 11, 12 (2010) (introducing the symposium issue of the *New York Law School Law Review* on the Google Books lawsuit and settlement); Pamela Samuelson, *The Google Book Settlement as Copyright Reform*, 2011 WIS. L. REV. 479 (2011).

2. Library digitization is the process whereby print-based library collections are converted to digital form using scanning and optical character recognition.

3. Authors Guild et al. v. Google Inc., 770 F. Supp. 2d 666, 669–72 (S.D.N.Y. 2011) (reviewing procedural history and rejecting proposed settlement).

4. The U.S. Copyright Office defines "orphan works" as works that are subject to copyright but whose copyright owners "cannot be identified and located by someone who wishes to make use of the work in a manner that requires permission of the copyright owner." U.S. COPYRIGHT OFFICE, REPORT ON ORPHAN WORKS 1 (2006) *available at* http://www.copyright.gov/orphan/orphan-report-full.pdf [hereinafter REPORT ON ORPHAN WORKS].

5. *Authors Guild*, 770 F. Supp. 2d at 670 (noting that "Older books—particularly out-of-print books, many of which are falling apart buried in library stacks—will be preserved and given new life" (citing Matthew Sag, *The Google Book Settlement & the Fair Use Counterfactual*, 55 N.Y.L. SCH. L. REV. 19, 73 (2010))).

6. *Id.* at 673–86 (rejecting proposed settlement under Rule 23 of the Federal Rules of Civil Procedure).

consumptive uses of individual works, there is no orphan works problem. This conclusion is supported by the fact that copyright owner's exclusive rights are generally limited to the expressive elements of their works and the expressive uses of their works.[7]

II. UNRAVELING THE DIGITIZATION DEBATE

Google entered the world of library digitization in 2004 when it began scanning and digitizing the collections of a number of prestigious private and public academic libraries to make their contents searchable in the same way it makes Internet websites searchable. In many cases, Google also displayed three-line "snippets" of the contents of those books to the general public—just enough to indicate to the searcher whether the text was really responsive to their search term.[8] Google has been mired in copyright litigation regarding its library digitization project since 2005 when the Authors Guild, along with a group of publishers, sued Google in a class action on behalf of all authors.[9] Google does not need permission to digitize works in the public domain and the company has also obtained permission from several publishers to include their works in the Google book search engine under agreed terms.[10] However, the company is also digitizing millions of in-copyright works without prior authorization from the relevant copyright owners, and therein lays the core of the dispute.[11]

The first step towards unraveling the digitization debate is to distinguish between different types of library digitization projects. Google's aspirations for book searches have shifted in a way that complicates the library digitization debate. Initially, the Google Library Project ("GLP") focused on data processing and search; however, on October 28, 2008, Google, the Authors Guild, and a group of leading publishers proposed a class action settlement that, among other things, would have transformed the GLP into a

7. I first made this argument in an article addressing the significance of copy-reliant technology more generally. This Article refines and extends my earlier analysis. *See* Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. 1607 (2009).

8. *Id.* at 1620–22 (describing the operation of the Google book search engine).

9. *Authors Guild*, 770 F. Supp. 2d at 670.

10. Google's Partner Program enables rights owners to opt into book search and allows them to control how their works are searched and displayed. Google has signed up over 20,000 rights holders to this Partner Program. *See Information for Authors and Publishers*, GOOGLE BOOKS, http://www.google.com/googlebooks/publishers.html.

11. As of March 2011, Google had scanned at least twelve million books. *See Authors Guild*, 770 F. Supp. 2d at 670.

general distribution platform for electronic versions of books.[12] For the sake of clarity, this Article will refer to the former as "GLP-search" and the latter as "GLP-distribution." The distinction is important because, although GLP-search has a strong claim to legality under current U.S. copyright law, GLP-distribution does not.[13]

Looking beyond Google, it is useful to think of all library digitization initiatives in three conceptually distinct genres corresponding to the three objectives of library digitization: (1) preserving existing volumes ("library digitization-preservation"); (2) facilitating data analysis and digital searching ("library digitization-search"); and (3) facilitating access to electronic versions of books ("library digitization-distribution"). The legal issues relating to each of these genres must be considered separately.

A. LIBRARY DIGITIZATION-PRESERVATION

Although libraries have certain privileges under the Copyright Act, nothing in the statute expressly allows wholesale library digitization with the exclusive aim of preserving existing volumes. Section 108 of the Copyright Act allows libraries to reproduce and distribute works "for purposes of preservation and security or for deposit for research use in another library" or to replace copies that are "damaged, deteriorating, lost, or stolen," or for which the existing format has become obsolete.[14] The scope of § 108 is very narrowly tailored and the provision does not authorize a general digitization program for preservation purposes.[15] For example, § 108(b) allows a library to make three copies of any unpublished work in its collection for preservation and security purposes.[16] Section 108(c) also permits a library to

12. Plaintiff's Motion for Preliminary Settlement Approval, Authors Guild v. Google Inc., 93 U.S.P.Q.2d 1159 (S.D.N.Y. Oct. 28, 2008) (No. 05 Civ. 8136) [hereinafter Motion for Preliminary Settlement Approval]. Settlement negotiations apparently began in the fall of 2006. In response to significant public criticism, including from the Department of Justice, the parties proposed an Amended Settlement Agreement on November 13, 2009. *Authors Guild*, 770 F. Supp. 2d at 671–72.

13. *See infra* Section II.C.

14. 17 U.S.C. § 108(a)–(c) (2010).

15. *See* Lois F. Wasoff, *If Mass Digitization Is the Problem, Is Legislation the Solution? Some Practical Considerations Related to Copyright*, 34 COLUM. J.L. & ARTS 731, 738 (2011). Wasoff noted:

> Current U.S. copyright law has no provision permitting libraries to make preservation copies of published works. Preservation copies are limited to unpublished works; replacement copies can be made of published works if the work is damaged or lost, but only if an unused copy cannot be located at a fair price.

Id.

16. 17 U.S.C. § 108(b).

make three copies of published works to replace a work in the library's collection that is (or was) damaged, deteriorating, lost, or stolen—but only if the library is unable to obtain a new copy at a fair price.[17]

The recommendations of the § 108 Study Group and the Copyright Principles Project to expand and clarify the scope of § 108, with respect to preservation, have much to recommend. However, the legal status of digitization aimed solely at preservation is an issue at the periphery of the debate.[18] Even if library digitization-preservation were clearly protected under the Copyright Act, there would still be considerable pressure to address the issues of search and distribution.

B. LIBRARY DIGITIZATION-DISTRIBUTION

In general, the digitization of library books to enable substantial display and/or distribution of e-books clearly implicates the copyright owner's rights. To scan a book is to reproduce the work in a digital copy,[19] and substantial textual displays and distribution of further copies clearly have the potential to substitute for the copyright owner's authorized copies and would not generally be protected by fair use. It is certainly arguable that fair use would protect the display of works that are out-of-print and whose copyright owner or owners cannot be located with reasonable efforts.[20] But putting the orphan works issue to one side for the time being, without additional facts, there is nothing to indicate that merely making a work more available is a transformative use that imbues the original "with a further purpose or different character, altering the first with new expression, meaning, or message."[21] There may be specific instances where such display or distribution would be justified as fair use, or would be protected by the § 108

17. 17 U.S.C. § 108(c). *See also* Laura N. Gasaway, *Values Conflict in the Digital Environment: Librarians Versus Copyright Holders*, 24 COLUM. J.L. & ARTS 115, 121–23 (2001); Samuelson, *Google Books*, *supra* note 1.

18. SECTION 108 STUDY GROUP, U.S. COPYRIGHT OFFICE, THE SECTION 108 STUDY GROUP REPORT iii–x (2008) (recommending numerous changes to library and archival privilege), *available at* http://www.section108.gov/docs/Sec108StudyGroupReport.pdf; *see also* Pamela Samuelson et al., *The Copyright Principles Project: Directions for Reform*, 25 BERKELEY TECH. L.J. 1175 (2010) (recommending some updates to library and archival privilege).

19. *See* 17 U.S.C. § 106(1) (providing that the copyright owner has the exclusive right to reproduce the work in copies).

20. To expand upon my view of this issue would be distracting in the context of this Article.

21. *See* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994).

library privilege or some other exception under the Copyright Act—but these would be exceptions to the usual rule.[22]

Consider, for example, two features of the GLP provided for in the Amended Settlement Agreement ("ASA"). Unless the rights holder opted out, the ASA would have allowed Google to sell online access to entire books as consumer purchases or "institutional subscriptions."[23] The ASA also envisaged a default book display of up to 20% of a book, not just a three-line snippet.[24] Such extensive displays may well benefit copyright owners by stimulating interest in the entire work, but they also potentially substitute for the original works.[25] The ASA would have allowed Google to sell access to copyrighted works in a format and to an extent that substituted for purchase of copyright owner authorized copies. Such an action is well beyond the conceivable parameters of the idea-expression distinction or fair use.[26]

To many, the legal obstacles confronting a full-fledged e-distribution model of library digitization highlight the failure of the law to adapt to new technology. GLP-distribution, as proposed under the ASA, has been described as "one of the most important applications of digital information technology in the information age."[27] Many out-of-print books are currently available only to those with access to large research libraries. Furthermore, library digitization has the potential to democratize access to these works and create an important sphere of equality of opportunity. If digitization were linked to some kind of payment mechanism it would help authors "breathe

22. A recent empirical study of fair use concludes that transformative use by the defendant is a robust predictor of a finding of fair use; the amount and substantiality of the defendant's unauthorized use of the plaintiff's work is a significant factor in litigated fair use cases; but also notes that there is "no evidence that commercial use (in contrast to direct commercial use) reduces the defendant's chance of maintaining a fair use defense." Matthew Sag, *Predicting Fair Use*, 73 OHIO ST. L.J. 47, 85 (2012).

23. Amended Settlement Agreement § 4.1, Authors Guild et al. v. Google Inc., 770 F. Supp. 2d 666 (S.D.N.Y. 2011) (No. 05 Civ. 8136) [hereinafter ASA].

24. *Id.* § 4.3(b)(1).

25. A&M Records v. Napster, Inc., 239 F.3d 1004, 1018 (9th Cir. 2001) (holding that increased sales of copyrighted material attributable to unauthorized use should not deprive the copyright holder of the right to license the material).

26. *See Authors Guild*, 770 F. Supp. 2d at 678 ("Google did not scan the books to make them available for purchase, and, indeed, Google would have no colorable defense to a claim of infringement based on the unauthorized copying and selling or other exploitation of entire copyrighted books.").

27. *See* Lateef Mtima & Steven D. Jamar, *Fulfilling the Copyright Social Justice Promise: Digitizing Textual Information*, 55 N.Y.L. SCH. L. REV. 77, 104 (2010).

new life into older, out-of-print books that are generally inaccessible to the public and have stopped generating revenue."[28]

Copyright law poses an obstacle to the electronic distribution of out-of-print books because of the high costs of proactively clearing rights with copyright owners. As time progresses, things like assignments, deaths, bankruptcies, mergers, spin-offs, asset sales, reversion clauses in publishing contracts, poor private record keeping, and poor public record keeping by the Copyright Office can complicate the question of who owns the work.[29] The more time elapses, the higher the likelihood that the public record no longer provides enough information to know whom to ask for permission to use the copyrighted material.[30] Changes in copyright law over the years have exacerbated this problem by making the vesting (and continuation) of copyright automatic and by increasing the term of protection to the author's life plus seventy years, or ninety-five years from first publication for works made for hire.[31] To the extent that digitization is infringing, libraries and technology developers cannot afford to ignore the fact that these works may be subject to copyright because, even in the absence of actual harm or malicious intent, copyright owners may recover both statutory damages (up to $150,000 per work infringed) and attorney's fees.[32]

The scale of the orphan works issue is potentially vast. One estimate finds that only 2.3% of books published in the United States between 1927 and 1946 are still in print.[33] Another reports that five out of every seven books Google scanned were not commercially available.[34] The Authors Guild estimates that approximately 75% of books in U.S. libraries are out-of-print and have ceased earning any income at all for their rights holders.[35] As the

28. Motion for Preliminary Settlement Approval, *supra* note 12, at 4.

29. *See generally* REPORT ON ORPHAN WORKS, *supra* note 4, at 23–29.

30. *Id.* at 26–29.

31. 17 U.S.C. § 302(b)–(c) (2010). As Pamela Samuelson notes:

> Had copyright terms not been repeatedly extended by Congress, all books published before 1953 would now be in the public domain, as would most of the books published before 1978 insofar as their rights holders did not renew the copyright. Because of copyright term extensions, books first published in 1960 are, however, unlikely to be out of copyright until 2055.

Pamela Samuelson, *Google Book Search and the Future of Books in Cyberspace*, 94 MINN. L. REV. 1308, 1313 (2010); *see also* Christopher Sprigman, *Reform(aliz)ing Copyright Law*, 57 STAN. L. REV. 485 (2004); *see generally* REPORT ON ORPHAN WORKS, *supra* note 4, at 41–44.

32. 17 U.S.C. §§ 504(c) (statutory damages), 505 (attorney's fees).

33. *See* Jason Schultz, *The Myth of the 1976 Copyright "Chaos" Theory*, LESSIG 2.0, http://www.lessig.org/blog/archives/jasonfinal.pdf (last visited June 14, 2012).

34. *See* Motoko Rich, *Google Hopes to Open a Trove of Little-Seen Books*, N.Y. TIMES, Jan. 5, 2009, at B1.

35. Motion for Preliminary Settlement Approval, *supra* note 12, at 27.

Copyright Office report on orphan works notes, this problem is particularly severe for institutions, such as libraries and museums, whose mission is to preserve and make available large archives of historical works.[36]

To the extent that rights clearance is truly uneconomic, copyright is failing both orphan works owners and the public at large. Copyright exists to enable authors to set a price on access, not to frustrate access for its own sake. Library digitization's enormous potential (whether it be economic, educational, social, or democratic), and the copyright law's current failure in relation to orphan works, have led to many proposals for reform.[37]

C. LIBRARY DIGITIZATION-SEARCH

The GLP version proposed under the ASA requires either judicial approval of the ASA (which will not be forthcoming) or legislative intervention. But what if Google were to scale back its ambitions to its initial proposal where unauthorized digitization was only incident to search? In the pure search scenario, the legality question of library digitization initiatives takes on a different complexion. Stated briefly, the argument favoring the legality of scanning, processing, and making fractional displays of books involved in GLP-search has three significant parts.

First, search results consisting of bibliographic information and relevance to a particular search query are facts not subject to copyright protection.[38] This is textbook copyright law in the United States and beyond serious dispute.[39]

Second, the very brief snippets or quotations that Google displays in its search results are either (a) too brief, fragmented, and insubstantial to constitute a reproduction of an entire copyrighted work[40] or (b) used in a transformative manor to indicate the relevance of search results and not to substitute for the actual text, as such these snippets serve a different function

36. REPORT ON ORPHAN WORKS, *supra* note 4, at 37–38.

37. *See, e.g.*, Shawn Bentley Orphan Works Act of 2008, S. 2913, 110th Cong. (2008).

38. 17 U.S.C. § 102(b) (providing that "[i]n no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such a work"). *See also* Feist Publ'g, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 363–64 (1991) (noting that copyright distinguishes between facts and their expression).

39. *See infra* Section III.A for further discussion of the idea-expression distinction.

40. Even if one took the view that Google's actual three line snippets were too long, there must be *some* length of snippet—whether it be three lines, two lines, one line, or ten words—that would be non-infringing.

than the original work and are thus fair use.[41] Either conclusion renders the search results displayed in GLP-search non-infringing. However, even if Google never showed a single book to anyone, the fact remains that it has been technically copying entire works to create its searchable database.

Third, copying entire expressive works for non-expressive (and otherwise non-infringing) purposes is itself fair use.[42] Notice here that although orphan works may raise distinct issues in some contexts, the legitimacy of scanning and digitizing orphan works for library digitization-search is largely folded into the broader question of the scope of the copyright owner's rights in relation to non-expressive use. However, we should not lose sight of the importance of orphan works to the underlying policy debate. The intractable licensing problems that create orphan works mean extending the rights of copyright owners to include non-expressive use that would create a substantial market failure. Going forward, it is conceivable that publishers will get the rights they need from authors and agree to license these rights to those seeking to make non-expressive use of covered works, but, for the reasons canvassed above, the rights with respect to the majority of orphan works held in libraries will never be cleared. Put simply, if a court establishes a new right that gives copyright owners a veto over non-expressive use of their works, those rights will never be cleared for millions of orphan works.

III. NON-EXPRESSIVE USE

A. COPYRIGHT, BALANCE, AND THE DISTINCTION BETWEEN IDEAS AND EXPRESSION

As expressed in the U.S. Constitution, copyright's motivating purpose is "to promote the Progress of Science and useful Arts."[43] Copyright law in the United States does not exist primarily to recognize or validate the natural rights of authors vis-à-vis their creations. Instead, its purpose is to encourage the authors' creativity and to promote the creation and dissemination of

41. *See* Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1165 (9th Cir. 2007) (holding that "[e]ven making an exact copy of a work may be transformative so long as the copy serves a different function than the original work.") (citing Kelly v. Arriba Soft Corp., 336 F.3d 811, 818–19 (9th Cir. 2003)); *see also* Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 609–10 (2d Cir. 2006); Field v. Google Inc., 412 F. Supp. 2d 1106, 1117–19 (D. Nev. 2006).

42. There are at least three search engine cases indicating as much. *See Perfect 10*, 508 F.3d at 1167–68; *Kelly*, 336 F.3d at 815; *Field*, 412 F. Supp. 2d at 1117–19. For a discussion of the fair use implications of non-expressive use generally, see Matthew Sag, *Copy-Reliant Technology*, *supra* note 7.

43. U.S. CONST. art. I, § 8, cl. 8.

works of authorship.[44] As the Supreme Court has noted on a number of occasions, the promotion of science and the useful arts requires a balance between "the interests of authors and inventors in the control and exploitation of their writings and discoveries on the one hand, and society's competing interest in the free flow of ideas, information and commerce on the other hand."[45] Where the law strikes that balance dictates what the public can copy and what authors can control. Just as importantly, it also mediates relationships between different generations of authors: initial authors and those who build upon their works.[46] Thus, while copyright aims to give authors an incentive to create and share their works, it also strives to provide subsequent authors with sufficient "breathing room" to make their own additive contributions.[47] The copyright system is predicated both on the existence of certain rights to protect authors from unfair competition, and on significant gaps in those rights that give others freedom to create and freedom to interact with existing works.

The distinction between ideas and expression is an important part of the balance of copyright law.[48] Copyright in an expressive work does not confer any exclusive rights in the facts, ideas, concepts, or discoveries contained in that work, regardless of the form in which the work describes, explains, or illustrates them.[49] This principle, often simply abbreviated to the "idea-

44. Eldred v. Ashcroft, 537 U.S. 186, 219 (2003).

45. Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 580 (1985) (quoting Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 429 (1984)).

46. *See generally* Mark A. Lemley, *The Economics of Improvement in Intellectual Property Law*, 75 TEX. L. REV. 989 (1997) (discussing sequential innovation in copyright and patent law).

47. *See, e.g.*, Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 933 (2005). The court in *Sony* noted:

> The fair use doctrine must strike a balance between the dual risks created by the copyright system: on the one hand, that depriving authors of their monopoly will reduce their incentive to create, and, on the other, that granting authors a complete monopoly will reduce the creative ability of others.

Sony, 464 U.S. at 479.

48. *Eldred*, 537 U.S. at 219 (stressing that the idea-expression distinction is one of copyright's "built-in First Amendment accommodations" and that "[d]ue to this distinction, every idea, theory, and fact in a copyrighted work becomes instantly available for public exploitation at the moment of publication."); *Harper & Row*, 471 U.S. at 556 (noting that the idea-expression distinction "strikes a definitional balance between the First Amendment and the Copyright Act by permitting free communication of facts while still protecting an author's expression.").

49. 17 U.S.C. § 102(b) (2010); *Harper & Row*, 471 U.S. at 547 (1985) (holding that "no author may copyright facts or ideas").

expression distinction," is longstanding at common law and was expressly incorporated into the 1976 revision of the Copyright Act.[50]

At least since *Baker v. Selden* in 1879, courts have recognized that "there is a clear distinction between the book, as such, and the art that it is intended to illustrate."[51] The distinction holds even in those unusual cases where the work's true value lies in the methods, systems, and ideas it discloses, rather than in the author's expression of those concepts.[52] In *Selden*, for example, the plaintiff had developed a novel and useful bookkeeping method, the practice that created value regardless of how and from what source a bookkeeper learned the method.[53] Nonetheless, the plaintiff's copyright in his instructional material was limited to the expression of his useful methods and did not encompass those methods themselves.[54] Of course, in most cases, protecting the unique expression of an idea is sufficient to ensure that the author will be able to appropriate a return on her investment.[55]

Copyright law also distinguishes between facts and the expression of facts, providing no protection for the former and only limited protection for the latter.[56] In *Feist Publications, Inc. v. Rural Telephone Service Co.*, the Supreme Court ruled that copying listings from a telephone directory did not infringe the copyright in that directory because the information itself was not copyrightable.[57] As the Court explained, facts—whether they are telephone numbers and addresses or the details of historical occurrences—are not

50. 17 U.S.C. § 102(b) provides: "In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such a work."

51. Baker v. Selden, 101 U.S. 99, 102 (1879). "Art" and "illustrate" are not meant in the aesthetic sense in this context.

52. *Id.*

53. *Id.* at 99–100. Selden's system may well have been patentable under today's standards. *See* State Street Bank & Trust Co. v. Signature Fin. Group, Inc., 149 F.3d 1368, 1373 (Fed. Cir. 1998) (holding that a patent on a data processing system is valid). *But see* Lab. Corp. of Am. Holdings v. Metabolite Labs., Inc., 548 U.S. 124, 136 (2006) (per curiam) (Breyer, J., dissenting) (noting that the Supreme Court has never endorsed the Federal Circuit's "useful, concrete, and tangible result" test for patentable processes). *See generally* Pamela Samuelson, *Why Copyright Law Excludes Systems and Processes from the Scope of Its Protection*, 85 TEX. L. REV. 1921, 1924 (2007) (arguing that thin copyright protection for computer programs is especially appropriate given the availability of patent protection for program innovations).

54. *Baker*, 101 U.S. at 103–04.

55. *Cf.* WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY 91–108 (2003).

56. *See* Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 349–50 (1991) (holding that facts are not copyrightable and that the copyright in a factual compilation is thin).

57. *Id.* at 362–63.

"original" to the author.[58] The author's copyright, therefore, did not cover the facts themselves.[59] The *Feist* Court further held that the expression of those facts was not protectable, because the selection and alphabetical arrangement of those facts in the telephone directory was "so mechanical or routine as to require no creativity whatsoever."[60] The rule in *Feist* even extends to "false facts."[61]

Through the idea-expression distinction, copyright law protects the expressive elements of the author's work while guaranteeing subsequent authors the necessary breathing space to make their own contributions by adding to, reusing, or reinterpreting the facts and ideas embodied in the original work. Subsequent authors may not compete with the copyright owner by offering her original expression to the public as a substitute for the copyright owner's work, but they are free to compete with their own expression of the same facts, concepts, and ideas. Thus, the idea-expression distinction is a central element of the balance between the interests of authors in preventing the exploitation of their writings and society's competing interest in the free flow of ideas, information, and commerce.[62]

Demarcating the precise boundary between ideas and expression is no easy task. The famous 1930 case of *Nichols v. Universal Pictures Corp.* dealt with a play about lovers from different religious backgrounds and a motion picture with the same motif.[63] The playwright, whose work came first, alleged

58. *Id.* at 348 ("[C]opyright protection may extend only to those components of a work that are original to the author.").

59. *Harper & Row*, 471 U.S. at 556 ("No author may copyright his ideas or the facts he narrates.").

60. *See Feist*, 499 U.S. at 362 (holding that the selection, coordination, and arrangement of Rural's white pages did not satisfy the minimum constitutional standards for copyright protection); *see also* Matthew Bender & Co. v. West Publ'g Co., 158 F.3d 674, 676 (2d Cir. 1998) (holding that West's factual enhancements to judicial opinions could be reasonably viewed as obvious, typical, and lacking even minimal creativity).

61. False facts are denied protection under a theory of "copyright estoppel." Skinder-Strauss Assocs. v. Mass. Continuing Legal Educ., Inc., 914 F. Supp. 665, 675–76 (D. Mass. 1995); Houts v. Universal City Studios, Inc., 603 F. Supp. 26, 28 (C.D. Cal. 1984) ("once a plaintiff's work has been held out to the public as factual the author-plaintiff cannot then claim that the book is, in actuality, fiction and thus entitled to the higher protection allowed to fictional works."). Some courts have been willing to grant de facto database protection to individual facts brought into being as a result of creative choices, such as bluebook valuations, and price guides to rare coins. *See, e.g.*, CDN, Inc. v. Kapes, 197 F.3d 1256 (9th Cir. 1999).

62. Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 429 (1984); s*ee also* Warner Bros. v. Am. Broad. Cos., 720 F.2d 231, 240 (2d Cir. 1983) (describing the idea-expression distinction as "an effort to enable courts to adjust the tension between these competing effects of copyright protection").

63. Nichols v. Universal Pictures Corp., 45 F.2d 119, 120–21 (2d Cir. 1930).

that the movie infringed his rights.[64] Ruling for the defendant, Judge Learned Hand observed that although copyright must extend beyond the exact literal text of a work, similarities between two works at a high level of generality cannot violate the author's rights because a playwright can not "prevent the use of his 'ideas,' . . . apart from their expression . . . ."[65] Having described the idea-expression distinction, the learned judge immediately observed that "[n]obody has ever been able to fix that boundary, and nobody ever can."[66] Although the precise point of departure between protectable expression on the one hand and unprotectable fact and ideas on the other may be elusive,[67] unstable[68] and somewhat subjective,[69] no one doubts that it exists.[70]

The distinction between expressive and non-expressive *works* is already well recognized in copyright law as the gatekeeper to copyright protection—novels are protected by copyright, telephone books and other uncreative compilations of data are not.[71] The position of this Article is that the same distinction should generally be made in relation to potential *acts* of infringement. Preserving the functional force of the idea-expression distinction in the digital context requires courts to conclude that copying for purely non-expressive purposes, such as the automated data extraction, should not be regarded as infringing. As this Article will explain in Section III.C, *infra*, courts are already tacitly implementing the principle of non-expressive use in the case law. The principle, however, needs to be brought to the surface.

---

64. *Id.*

65. *Id.* at 121.

66. *Id. See also* Peter Pan Fabrics, Inc. v. Martin Weiner Corp., 274 F.2d 487, 489 (2d Cir. 1960) ("Obviously, no principle can be stated as to when an imitator has gone beyond copying the 'idea,' and has borrowed its 'expression.' Decisions must therefore inevitably be ad hoc.").

67. Professor Chafee's proposed "pattern" test for determining the line between an idea and its expression is as good as any, but it essentially reframes the question rather than answering it. Zechariah Chafee, *Reflections on the Law of Copyright: I*, 45 COLUM. L. REV. 503, 513–14 (1945).

68. *See, e.g.*, Peter Jaszi, *The Metamorphoses Of "Authorship,"* 41 DUKE L.J. 455, 465 (1991).

69. *See* Amy B. Cohen, *Copyright Law and the Myth of Objectivity: The Idea-Expression Dichotomy and the Inevitability of Artistic Value Judgments*, 66 IND. L.J. 175, 228 (1990) (reviewing the application of the idea-expression distinction in case law and concluding that where the line is drawn "reflects the judge's view of the artistic value of the works at issue based on what the judge knows about and values in literary works on that subject.").

70. *But see* Edward C. Wilde, *Replacing the Idea/Expression Metaphor With a Market-Based Analysis in Copyright Infringement Actions*, 16 WHITTIER L. REV. 793 (1995).

71. Feist Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991).

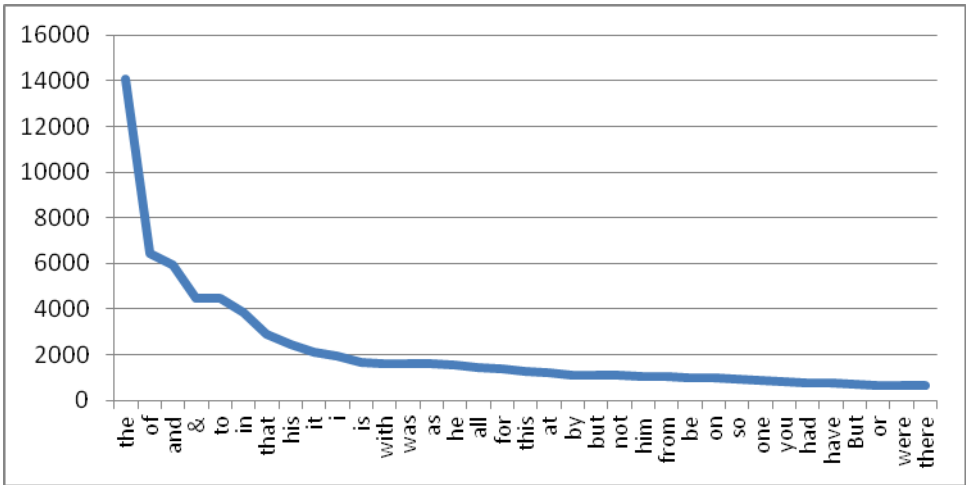
B. NON-EXPRESSIVE USE

1. Coming to Grips with the Concept of Non-Expressive Use

To understand what non-expressive use means, consider the following thought experiment.

Brian has a perfect memory for the written word: he can recite every book he has ever read perfectly from start to finish. He can, if pushed, write out frequency tables that report the number of times any given word or punctuation mark appears in any work.[72] Brian might, for example, produce the following word frequency graphs for Herman Melville's *Moby Dick*. Figure 1 illustrates the frequency of common English words in Moby Dick based on a list of words that is not sensitive to context such as "the," "of," "and," "have," etc. Figure 2 illustrates word frequencies in Moby Dick once the words in Figure 1 have been excluded.

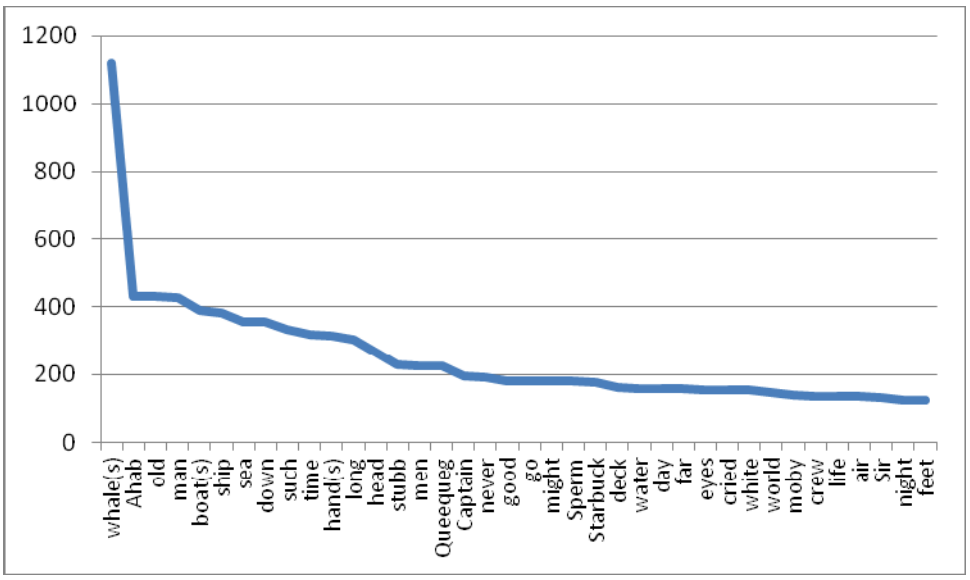
Figure 1: Frequency of Common English Words in Moby Dick[73]



72. ALEKSANDR ROMANOVICH LURIA, THE MIND OF A MNEMONIST: A LITTLE BOOK ABOUT A VAST MEMORY (1987) (an account of a Russian man with a limitless memory).

73. HERMAN MELVILLE, MOBY DICK; OR, THE WHALE, *available at* http://www.princeton.edu/~batke/moby/moby.html. Word frequency obtained using Wordle.net java applet. WORDLE, http://www.wordle.net/(java applet for obtaining word frequency). Words selected by the author.

Figure 2: Frequency of Selected Words in Moby Dick[74]



The same information can be represented in a more whimsical visual style using a word cloud as follows in Figure 3 and Figure 4.

74. HERMAN MELVILLE, MOBY DICK; OR, THE WHALE, *available at* http://www.princeton.edu/~batke/moby/moby.html. Word frequency obtained using Wordle.net java applet. WORDLE, http://www.wordle.net/ Words selected by the author.

Figure 3: Moby Dick Word Cloud (Common)[75]

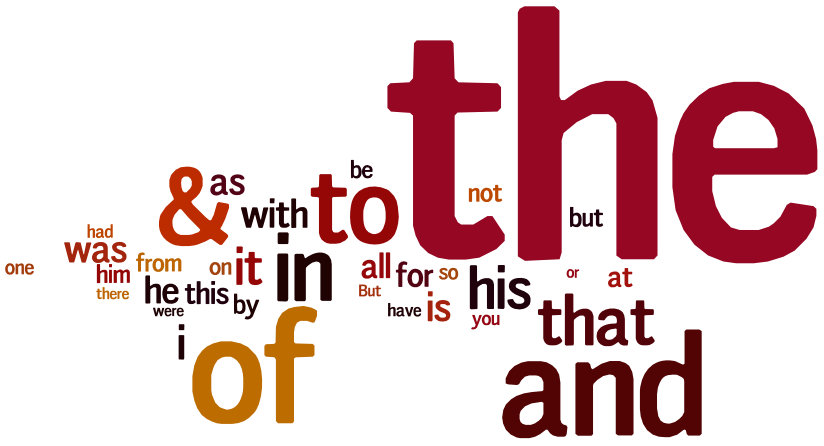


Figure 4: Moby Dick Word Cloud (Uncommon)[76]

Assume for the moment that *Moby Dick* is still protected by copyright.[77] If Brian were to simply transcribe the novel and sell his transcription in bookstores or on street corners, he would undoubtedly infringe the author's

75. HERMAN MELVILLE, MOBY DICK; OR, THE WHALE, *available at* http://www.princeton.edu/~batke/moby/moby.html. Word cloud produced using Wordle.net. WORDLE, http://www.wordle.net/.

76. HERMAN MELVILLE, MOBY DICK; OR, THE WHALE, *available at* http://www.princeton.edu/~batke/moby/moby.html. Word cloud produced using Wordle.net. WORDLE, http://www.wordle.net/. Words selected by the author.

77. First published in 1851, Moby Dick is no longer protected by copyright.

copyright.[78] But would the frequency table infringe the author's copyright? The frequency table itself is metadata, data about the work that is entirely independent of the expressive value of the work. True enough, the data relies on the underlying work, but it has no similarity to the work in terms of plot, structure, character (other than the names of characters) or theme. This data, by itself, does not infringe the copyright owner's rights.

Is there a point at which an analytical work explains so much of the content of its expressive subject that the author's rights have been infringed? Perhaps. In *Castle Rock Entertainment v. Carol Publishing Group, Inc.*,[79] the Second Circuit held that a quiz book based on the characters and events of the popular television series *Seinfeld* violated the show's copyright. The court acknowledged that the substantially similar standard depends on "the copying of expression, rather than ideas" and that the quiz reproduced none of the plot, sequence, pace, or setting of the show.[80] The defendant's quiz focused on "facts" internal to the *Seinfeld* universe, such as the reason that Kramer enjoys going to the airport (because he is hypnotized by the baggage carousels) or what it was that Jerry placed on Elaine's leg during a piano recital (a Pez dispenser), and not facts about the show.[81] The court of appeals took the view that "[b]ecause these characters and events spring from the imagination of *Seinfeld's* authors, the [quiz] plainly copies copyrightable, creative expression."[82] Of course, the court cannot really mean that any work that refers to the characters and events in a creative work is infringing. Furthermore, there are volumes of guide-books and analytical works that do not interfere with the copyright owner's exclusive rights, and it is well established that "ownership of copyright does not confer a legal right to control public evaluation of the copyrighted work."[83] The real problem with the defendant's quiz in *Castle Rock Entertainment* was that it sought to "repackage *Seinfeld* to entertain *Seinfeld* viewers" and that the quiz itself was in no way analytical.[84] If the *Seinfeld* quiz infringed the copyright owner's rights at all, it was because it essentially recast the series' copyrightable characters

78. 17 U.S.C. § 106(1) (reproduction), (3) (distribution).
79. 150 F.3d 132 (2d Cir. 1998).
80. *Id.* at 138.
81. *Id.* at 139.
82. *Id.*
83. Ty, Inc. v. Publ'ns Int'l. Ltd., 292 F.3d 512, 521 (7th Cir. 2002).
84. *Castle Rock Entertainment*, 150 F.3d at 140–43.

into a new format, much the same as if the defendant had made miniature dolls of the show's characters.[85]

The recent *Harry Potter Lexicon* case is also on point.[86] In *Warner Brothers Entertainment Inc. v. RDR Books*, the court found that a guidebook to the famed *Harry Potter* series violated the author's copyright.[87] The court found that the Lexicon was substantially comprised of direct quotations (often without quotation marks) and close paraphrases of vivid passages in the Harry Potter books.[88] Like the *Seinfeld* quiz, the Lexicon related "fictional facts" the author, J.K. Rowling, had created. In line with *Castle Rock*, the court concluded "such invented facts constitute creative expression protected by copyright because characters and events spring from the imagination of the original authors."[89] One interpretation of the court's opinion in the *Harry Potter Lexicon* case is that if the guidebook had not borrowed so extensively from the original author's expression, it would not have been found to infringe.[90] The Lexicon's purpose was to "give the reader a ready understanding of individual elements in the elaborate world of Harry Potter that appear in voluminous and diverse sources."[91] The district court in the *Harry Potter Lexicon* case held that the Lexicon did not infringe the copyright owner's right to make derivative works because it no longer represented the original work of authorship and did not fall under any example of derivative works listed in the statute.[92] The court followed the Seventh Circuit's holding that a collector's guide to stuffed toys is not a derivative work because "guides don't recast, transform, or adapt the things to which they are guides."[93] If the Lexicon had been drafted with more care, it need not have infringed the copyright owner's rights.

---

85. *See, e.g.*, Hasbro Bradley, Inc. v. Sparkle Toys, Inc., 780 F.2d 189 (2d Cir. 1985) (upholding copyrightability of "Transformer" changeable robotic action figures as sculptural works).

86. Warner Bros. Entm't Inc. v. RDR Books, 575 F. Supp. 2d 513 (S.D.N.Y. 2008).

87. *Id.*

88. *Id.* at 527 ("the Lexicon indeed contains at least a troubling amount of direct quotation or close paraphrasing of Rowling's original language"); *id.* at 530 ("The Lexicon's close paraphrasing is not limited to the seven *Harry Potter* novels, but can be found in entries drawn from the companion books as well."); *id.* at 531 ("Instances of such verbatim copying or close paraphrasing of language in the *Harry Potter* works occur throughout the Lexicon.").

89. *Id.* at 536 (citations and quotations omitted).

90. The decision could be clearer as to the relationship between findings of fact and legal conclusions.

91. *Id.* at 539.

92. *Warner Bros.*, 585 F. Supp. 2d at 539 ("Under these circumstances, and because the Lexicon does not fall under any example of derivative works listed in the statute, Plaintiffs have failed to show that the Lexicon is a derivative work.").

93. Ty, Inc. v. Publ'ns Int'l. Ltd., 292 F.3d 512, 520 (7th Cir. 2002).

The automated data analysis of text that this Article addresses is a far cry from the fragmented expression copying in the *Harry Potter Lexicon* case and other similar "fictional facts" cases. Copyright does not protect individual words, even in the rare instances where they are in fact a creation of the author.[94] For example, an author such as J.K. Rowling can be said to originate the following twenty-word string of text: "[g]oblin-made armour does not require cleaning, simple girl. Goblins' silver repels mundane dirt, imbibing only that which strengthens it."[95] But none of these individual words originates with Rowling. The corresponding entry in the Lexicon reads "[a]ccording to Phineas Nigellus, goblin-made armor does not require cleaning, because goblins' silver repels mundane dirt, imbibing only that which strengthens it, such as basilisk venom."[96] Moreover, the observation that no word other than "goblin" is repeated in either sentence originates, not with Rowling, but with the author of this Article.[97] Likewise, if some anti-plagiarism software were to identify a high level of similarity between the two quotes—as it surely would—that data could not be said to originate with either the author of Harry Potter or the author of the Lexicon. It is a fact about the works and is in no sense a reproduction of either work or a substantial part of the original expression therein. In summary, metadata of the sort described here infringes only as much as a landscape painting inspired by a novel, or a musical composition inspired by a film would—i.e., not at all.[98]

Returning to our thought experiment, would Brian infringe the copyright owners rights by simply memorizing *Moby Dick* as part of the process of making the table? If Brian is a human being, it seems absurd to suggest that the perfect storage of information in his brain violates the copyright owner's exclusive right to "reproduce the work in copies . . ." under § 106(1) of the Copyright Act. Even if scientists told us that Brian's brain stored and could recall the information with perfect accuracy,[99] it is inconceivable that human

94. The Copyright Office has a long-standing rule that "words and short phrases such as names, titles, and slogans" are not copyrightable. 37 C.F.R. § 202.1(a) (2004). *See* Justin Hughes, *Size Matters (Or Should) In Copyright Law*, 74 FORDHAM L. REV. 575 (2005).

95. *Warner Bros.*, 575 F. Supp. 2d at 527 (quoting J.K. ROWLING, HARRY POTTER AND THE DEATHLY HALLOWS 303 (2007)).

96. *Id.*

97. Admittedly, this is not a profound observation.

98. *See* Robert Kastenmeier, Copyright Law Revision, H.R. Rep. No. 94-1476 (2d sess. 1976) (noting a programmatic musical composition inspired by a novel).

99. Looking closely at the definition of copies in § 101 of the Act it is not immediately clear that the human brain cannot be a copy. To amount to a copy under the Act, the medium storage must simply be a "material object . . . in which a work is fixed . . . and from which the work can be perceived, reproduced, or otherwise communicated, either directly or

thought or human memory could be a form of copyright infringement.[100] Now suppose that Brian is a computer; should the answer really be any different?

2. Examples of the Non-Expressive Use of Expressive Works

Ordinarily, the direct or indirect purpose of reproducing an expressive work relates to human appreciation of the expressive qualities of that work. We might, for example, download a film to watch it, or photocopy a magazine article to read it. The examples that follow illustrate a very different kind of motivation for copying text: reproduction as part of a process of data analysis that does not enable human enjoyment, appreciation, or comprehension of the text. These examples demonstrate the utility of automated non-expressive uses. They also demonstrate that such uses are no threat to the interests of copyright owners. This Section begins with two of the more obvious examples unrelated to library digitization—Internet search engines and plagiarism detection software—before turning to the role of non-expressive use in library digitization.

a) Internet Search Engines[101]

Internet search engines provide the most obvious example of the importance of the non-expressive use of copyrighted works. Internet search engines are a form of copy-reliant technology in that they require the routine and indiscriminate copying of html web pages.[102] Search engines use automated software agents that continuously "crawl" across the Internet copying web pages. These copies form the raw data underpinning these

with the aid of a machine or device." 17 U.S.C. § 101 (2010). Most people's brains do not store information with the stability and fidelity required to meet this definition, but what of those that do? One answer is to posit, as David Nimmer has in relation to tattoos, that a human is not a "material object," and while this may be a sound policy-based exclusion, it does not supply its own rationalization. *See* Declaration of David Nimmer, Whitmill v. Warner Bros. Entm't, No. 4:11CV752 (E.D. Mo. May 20, 2011) (declaration in support of the defendant in copyright litigation regarding the use of Mike Tyson's facial tattoo in the motion picture THE HANGOVER II (Warner Brothers 2011)).

100. Likewise in patent law, "[a] principle, in the abstract, is a fundamental truth; an original cause; a motive; these cannot be patented, as no one can claim in either of them an exclusive right." Le Roy v. Tatham, 55 U.S. (14 How.) 156, 175 (1853); Mayo Collaborative Servs. v. Prometheus Labs., Inc., 132 S. Ct. 1289 (2012) (noting that natural phenomena, mental processes, and abstract intellectual concepts are not patentable). *See generally* Kevin Emerson Collins, *Propertizing Thought*, 60 SMU L. REV. 317 (2007).

101. *See* Sag, *Copy-Reliant Technology*, *supra* note 7 for a more detailed account the operation of Internet search engines and plagiarism detection software.

102. Sergey Brin & Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, 30 COMPUTER NETWORKS AND ISDN SYS. 107 (1998), *available at* http://infolab.stanford.edu/~backrub/google.html.

search engines, which are subsequently analyzed and cataloged. As part of this process, search engines both copy and index each web page they find. The search engine directs the user to particular websites based on the relationship of her search term to the index of pages maintained by the search engine provider.[103] The search engine's use is non-expressive because the software copies expressive works in order to apply certain mathematical functions to their contents, not to comprehend or enjoy copyrighted works in the way that humans do. Of course, at the end of the day, search engines are mostly useful because they lead people to particular websites. But the search engine itself does not copy the website for the end user. Instead, this process is performed separately by the user's browser at the direction of the user.[104]

b) Plagiarism Detection Software

Plagiarism detection software is another illustration of the copying of expressive works for non-expressive ends. In the educational context, automated plagiarism services rely on access to entire copies of student term papers and any works from which a student might have copied them, yet the services do not necessarily display any of the copyrighted content they process to the end users.[105] The software works by comparing strings of text in new works to strings of text in existing works.[106] The similarities between two works can be assessed by looking for common strings of words. However, there are also various algorithms that can be applied to a document to create a digital fingerprint that captures other characteristics of the work. These digital fingerprints allow a document to be characterized by its structure, vocabulary, and content. Furthermore, they are essentially abstractions of the original documents and allow for faster comparisons that will not be as easily deceived by minor text alterations.[107] If the software finds

103. *See, e.g.*, U.S. Patent No. 6,285,999 (filed Jan. 9, 1998) ("Method for Node Ranking in a Linked Database").

104. Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1161 (2007).

105. S*ee* Sag, *Copy-Reliant Technology*, *supra* note 7 for a more detailed account the operation of Internet search engines and plagiarism detection software.

106. *See* Amy Argetsinger, *Technology Snares Cheaters at U-Va.; Physics Professor's Computer Search Triggers Investigation of 122 Students*, WASH. POST, May 9, 2001, at A1.

107. *See, e.g.*, Khair Eddin M. Sabri & Jubair J. Al-Ja'afer, *The JK System to Detect Plagiarism*, 6(2) J. COMPUTER SCI. & TECH. 66 (2006). The Turnitin software at issue in *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630 (4th Cir. 2009) used statistical techniques originally designed to analyze brain waves to compare the fingerprints of student papers to more than a billion documents that have been fingerprinted in a similar fashion. *See Plagiarise. Let No One Else's Work Evade Your Eyes*, THE ECONOMIST, Mar. 14, 2002, *available at* http://www.economist.com/node/1033832.

a match, it indicates as much. By itself, the report that a new work is similar to another work already in the database in no way reproduces or communicates the expressive qualities of either work.[108]

c) Non-Expressive Use and Library Digitization

Library digitization raises many novel issues, but one should not lose sight of the fact that some of the relevant issues are not at all new. The fundamental issue with respect to the legality of copying to build a search engine is the same for web pages as it is for library books. In point of fact, there are some interesting differences. To start, library digitization also raises interesting questions about the scope of the § 108 library privilege.[109] Non-profit libraries that undertake digitization initiatives might have additional arguments to make with respect to fair use. Likewise, the automated copying of html pages may also be protected by an implied license in many cases.[110] But these are distractions; the key question remains whether automated and systematic copying of text to enable a search engine (but not a display engine) or other data-processing function violates the rights of the copyright owner.

In addition to book searches, there are many non-expressive uses for library digitization. Researchers could use a digitized collective (referred to in the trade as the "corpus") to test and refine search algorithms more generally.[111] Other researchers could use the resulting data field to improve automated translation software and to develop and test theories in linguistics. Some of the most interesting illustrations of the kind of non-expressive use that library digitization enables relate to the meta-analysis of literature.

In the world of books, a non-expressive use is any use that, while it may literately involve reproduction of the work, does not involve any human

108. Of course, in practice most plagiarism software is also programmed to display the source file from which the work being scrutinized was allegedly copied. This optional feature is an expressive use, although it is almost certainly protected by fair use because the purpose of the display is to provide evidence of a claim of cheating. A.V. *ex rel.* Vanderhye, 562 F.3d at 641–42 (finding that the defendant's use of the works as part of a digitized database from which to compare the similarity of typewritten characters used in other student works was unrelated to any creative component of the work).

109. *See, e.g.*, Peter B. Hirtle, *Digital Access to Archival Works: Could 108(b) Be the Solution?*, COPYRIGHT & FAIR USE: STANFORD UNIV. LIBRARIES (Sept. 24, 2006), http://fairuse.stanford.edu/commentary_and_analysis/2006_08_hirtle.html.

110. Field v. Google Inc., 412 F. Supp. 2d 1106, 1115 (D. Nev. 2006).

111. *See* ASA, *supra* note 23, § 1.93 (defining non-consumptive use to include Image Analysis and Text Extraction, Textual Analysis and Information Extraction, Linguistic Analysis, Automated Translation, and Indexing and Search (research on different techniques for indexing and search of textual content).

reading the digitized copy of the book. If the data extracted does not allow for the work to be reconstructed, there is no substitution of expressive value. Extracting factual information about a work in terms of its linguistic structure or the frequency of the occurrence of certain words, phrases, or grammatical features is a non-expressive use.[112]

To start with a simple example, merely reporting the fact that the word "whale" or "whales" appears 1,119 times in Herman Melville's *Moby Dick* does not infringe any copyright in the book because this information about the work is entirely independent of the expressive value of the work.[113] There is no copyright in such basic information as the names of characters in a novel or a list of places they have been.[114] Nor is copyright infringed by the simple observation that Melville writes a great deal about whales, old men, the sea, boats, water, and ships. To preserve the force of the idea-expression distinction in the age of reading machines, one must recognize that copyright law does not prevent the automated extraction of such features by machine applications, even if those machines reproduce the text as a step in the analytical process. In this context, so long as the output is non-infringing, the machine is non-infringing.

Consider, for example, Franco Moretti's fascinating map of protagonists in Parisian Novels and the objects of their desire.[115] Aggregating information across many books allows us to see not only that the heroes of this particular genre are clustered in the Latin Quarter, but also that they are invariably separated from their heart's true desire by the River Seine and distributed in a convex arc as if held from the Latin Quarter by a constant unseen force. Moretti and a team of graduate students constructed this map by hand, but there is no obvious reason why a similar process on a grander scale could not be automated.

112. The ASA uses the awkward term "Non-Consumptive Research" to express the same concept. The ASA defines Non-Consumptive Research as "research in which computational analysis is performed on one or more Books, but not research in which a researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book." *Id.*

113. *See supra* Figure 2.

114. For a literary character to be protected as such by copyright it must, at a minimum, be distinctively delineated such that it represents a specific incarnation and not a general archetype. *Warner Bros. Pictures, Inc. v. Columbia Broadcasting Systems, Inc.*, 216 F.2d 945, 950 (9th Cir. 1954) sets a higher standard, that the "the character really constitutes the story being told" and is not merely a "chess man in the game of telling the story."

115. FRANCO MORETTI, GRAPHS, MAPS, TREES 55 (2005).

Literature scholars have traditionally focused on a close reading of canonical texts as the core of their discipline.[116] Even those who venture further afield do not travel that far. For example, literary historian Ian Watt's seminal 1957 work on the origins of the novel[117] is undoubtedly a brilliant synthesis of modern literature, and yet his entire scope of analysis is confined to three authors.[118] Three! Close reading of the literary cannon or of a few dozen exemplar works in a particular literary genre is no doubt an impressive skill, but the limits of this analysis are striking.[119] Empirically, the cannon itself cannot tell you anything about the characteristics that propel works into the cannon in the first place. To make that judgment, you must compare the cannon to the mass of other works all vying for that status but now largely forgotten.

Data-mining and macro-analysis of literature offers broad possibilities. Computer assisted text-analysis not only stores, searches, and retrieves text efficiently, it also automates the process of measuring and classifying natural-language documents to identify patterns that may be associated with author, subject, and genre or type.[120] Macro-analysis does not replace reading altogether, but it offers scholars a way to empirically test intuitions that are in fact quantitative or comparative in nature.[121] To take a rudimentary example, the notion that female characters are underrepresented in a particular period may be intuited from a small selection of prominent works. As such, macro-analysis of all the books from that period would allow that intuition to be tested empirically and potentially confirmed or falsified.[122] In his forthcoming book, *Macroanalysis: Methods for Digital Literary History*, Matthew Jockers uses various empirical techniques to identify the dominant themes in two of the most famous American novels of the nineteenth century—*The Last of the Mohicans* (1826) and *Moby Dick* (1851)—and contrast them against the nineteenth century corpus as a whole. Jockers does not read all 10,000 novels of the era,[123] but instead undertakes this investigation using word frequency

116. *See generally* MATTHEW JOCKERS, MACROANALYSIS: DIGITAL METHODS AND LITERARY HISTORY (forthcoming 2013).

117. IAN WATT, THE RISE OF THE NOVEL (1957).

118. Daniel Defoe, Samuel Richardson and Henry Fielding. *See id.* at 7.

119. There is an obvious parallel here with the rationale for conducting empirical legal studies. *See, e.g.*, Matthew Sag, Tonja Jacobi & Maxim Sytch, *Ideology and Exceptionalism in Intellectual Property: An Empirical Study*, 97 CALIF. L. REV. 801 (2009).

120. *See* Geoffrey Rockwell, *Why Bother With Computer-Assisted Text Analysis? A Short Answer*, TEXT ANALYSIS DEVELOPERS ALLIANCE (Apr. 30, 2005), http://tada.mcmaster.ca/Main/WhatTA.

121. *Id.*

122. *Id.*; JOCKERS, *supra* note 116.

123. 10,000 is a very rough guess.

analysis and computer generated topic modeling that identifies patterns based on the frequency with which words are combined. Jockers is an English professor, but he borrows techniques developed in computational linguistics and natural language processing to take account of grammatical structure and idiomatic usage in this analysis. No doubt, this is just the beginning of an exciting new field. The question for lawyers, judges, and legal academics is whether this type of analysis must be limited to public domain works and those licensed by publishers.

C. THE SCOPE OF COPYRIGHT WITH RESPECT TO THE NON-EXPRESSIVE USE OF EXPRESSIVE WORKS

The prescription in this Article, that copyright law should not stand in the way of the automated reproduction of text for non-expressive purposes, rests on the view that, in general, the copyright owner's exclusive rights are limited to the right to communicate the expressive aspects of her work to the public. To put it another way, copyright typically only concerns itself with the threat of expressive substitution. As already noted, the idea-expression distinction itself establishes that the copyright owner cannot prevent an ordinary reader from extracting and reproducing the facts or ideas embodied in her work. But the principle goes much deeper than this.

Copyright consists of a bundle of discrete exclusive rights, such as the reproduction right, the derivative right, and the public performance and display rights.[124] These rights are defined, articulated, and limited by a number of initially judge-made doctrines, such as the idea-expression distinction, the threshold of substantial similarity, and the fair use doctrine.[125] In my earlier work, I have explained in some detail that these doctrines typically limit copyright protection to the expressive aspects of original works of authorship in a way that confirms the place of public communication at the heart of copyright.[126] This Article will expand and clarify just a few of these arguments.

1. Substantial Similarity

The tests courts apply to determine the threshold of infringement—i.e., when some copying is *too much* copying—strongly suggest that the statutory rights of the author are limited to the communication of original expression

124. 17 U.S.C. § 106(1)–(6) (2010).

125. The Copyright Act of 1976 also reflects the idea-expression distinction and the fair use doctrine. *See* 17 U.S.C. §§ 102(b), 107. But these doctrines remain essentially common law features of the copyright system.

126. Sag, *Copy-Reliant Technology*, *supra* note 7.

to the public. The copyright owner's exclusive right to "reproduce the copyrighted work in copies" extends to both exact and inexact reproductions.[127] In both cases, however, the Copyright Act leaves the threshold of reproduction undefined. In cases of nonliteral infringement—where the accused work is not an exact copy of the copyright owner's work—courts assess whether the allegedly infringing work possesses a substantial similarity to the copyrighted work.[128]

Courts often define the threshold of substantial similarity from the perspective of the ordinary observer.[129] Infringement is defined in reference to the perspective of the consuming public because the copyright owner's "legally protected interest is not, as such, his reputation . . . but his interest in the potential financial returns from his [work] which derive from the lay public's approbation of his efforts."[130] Thus, the determination of whether work "B" borrowed too much from work "A" hinges upon how the public would regard the similarities between the works. But this is not the end of the analysis. Even when two works are similar taken as a whole, any similarities based on overlapping ideas or expressions that were not the plaintiff's to begin with "are by definition unprotected . . . ."[131] A plaintiff in a copyright case "must show that defendants' works are substantially similar to elements of plaintiff's work that are *copyrightable* or protected by the copyright."[132]

In cases of fragmented literal similarity, courts determine whether the copying amounts to infringement "by considering the qualitative and quantitative significance of the copied portion in relation to the plaintiff's work as a whole."[133] This focus on the qualitative and quantitative significance of the copied portion in the plaintiff's work is consistent with

127. 17 U.S.C. § 106(1); Nichols v. Universal Pictures Corp., 45 F.2d 119, 121 (2d Cir. 1930) ("[T]he question is whether the part so taken is substantial" (citing Marks v. Feist, 290 F. 959, 960 (2d Cir. 1923))) (internal quotation marks omitted).

128. *See* Tufenkian Imp./Exp. Ventures, Inc. v. Einstein Moomjy, Inc., 338 F.3d 127, 131 (2d Cir. 2003); Laureyssens v. Idea Grp., Inc., 964 F.2d 131, 140 (2d Cir. 1992).

129. This is especially true in the Second Circuit. *See* Shine v. Childs, 382 F. Supp. 2d 602, 614 (S.D.N.Y. 2005) (summarizing authorities). For a survey of other approaches, see MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 13.03 (2012).

130. Arnstein v. Porter, 154 F.2d 464, 473 (2d Cir. 1946) (footnotes omitted); *see also* Warner Bros. v. Am. Broad. Cos., 720 F.2d 231, 240 (2d Cir. 1983).

131. NIMMER *supra* note 129, § 13.03[2].

132. Whitehead v. Paramount Pictures Corp., 53 F. Supp. 2d 38, 46 (D.D.C. 1999) (emphasis in original) (citing NIMMER, *supra* note 129, § 13.03[2].).

133. Newton v. Diamond, 388 F.3d 1189, 1195 (citing Worth v. Selchow & Righter Co., 827 F.2d 569, 570 n.1 (9th Cir. 1987)); *see also* Jarvis v. A&M Records, 827 F. Supp. 282, 289–90 (D.N.J. 1993); NIMMER, *supra* note 129, § 13.03[A][2][a].

the prohibition against expressive substitution. Even where some of the copyright owner's original expression has been copied directly, such copying does not rise to the level of infringement unless the expression was significant, in either quantity or quality, in the author's original work.[134] Just as copyright law does not prevent the copying of facts and ideas, it also does not prevent the copying of trivial expressive elements from an existing work, because to do so does not unfairly compete with the copyright owner.[135] In other words, trivial copying of expressive elements is not copyright infringement because it does not interfere with the copyright owner's exclusive right to communicate her work to the public.

In summary, the very mechanics of assessing whether the threshold of substantial similarity has been met provide further evidence that copyright primarily protects the author against expressive substitution.

2. Intermediate Copying

For those in Hollywood, facing dubious claims of copyright infringement is a recognized cost of doing business.[136] Presumably, some of these claims are opportunistic, while others are the product of self-delusion. The attraction of a substantial payday combined with passing similarities based on title, theme, or subject matter can be enough to trigger a suit. What is significant for the purposes of this Article is that when confronted with motions for summary judgment based on an objective lack of similarity between their own work and that of the defendant, plaintiffs in a number of cases have turned to allegations of intermediate copying.[137] Typically, plaintiffs in this situation will urge the courts to allow scrutiny of every single

134. *Newton*, 388 F.3d at 1195 (9th Cir. 2004). The court noted:

> Fragmented literal similarity exists where the defendant copies a portion of the plaintiff's work exactly or nearly exactly, without appropriating the work's overall essence or structure. Because the degree of similarity is high in such cases, the dispositive question is whether the copying goes to trivial or substantial elements. Substantiality is measured by considering the qualitative and quantitative significance of the copied portion in relation to the plaintiff's work as a whole.

Id. (internal citations omitted).

135. *Id.* at 1193, 1195 ("The principle that trivial copying does not constitute actionable infringement has long been a part of copyright law. . . . [T]he dispositive question is whether the copying goes to trivial or substantial elements.").

136. Meritorious cases tend to be settled in private through Writers Guild arbitration.

137. *See, e.g.*, Stromback v. New Line Cinema, 384 F.3d 283, 299 (6th Cir. 2004); Flaherty, v. Filardi, No. 03 Civ. 2167, 2007 U.S. Dist. LEXIS 69202, at *8–9 (S.D.N.Y. Sept. 19, 2007) (dismissing copyright claim to interim drafts of a published non-infringing final work as a matter of law); Walker v. Time Life Films, Inc., 615 F. Supp. 430, 434–35 (S.D.N.Y. 1985) (denying request to discover drafts).

draft of the defendant's screenplay, in the hope that some earlier version of the work will disclose a greater resemblance to their own copyrighted work than the finished film does. Courts invariably deny these requests.[138] The reasons behind the denials provide an important insight into the structure of copyright law.

Courts refuse to entertain discovery with respect to early drafts of a non-infringing final work precisely because infringement requires at least some potential interference with the copyright owner's expectation of exclusivity. As noted in *Davis v. United Artists*, "the ultimate test of infringement must be the film as produced and broadcast, we do not consider the preliminary scripts."[139] Courts do not refuse to examine interim drafts merely because of judicial economy. As the Second Circuit noted in *Warner Bros., Inc. v. American Broadcasting Cos.*, "a defendant may legitimately avoid infringement by intentionally making sufficient changes in a work which would otherwise be regarded as substantially similar to that of the plaintiff's."[140] Likewise, in *See v. Durang*, the Ninth Circuit held "[t]he only discovery plaintiff suggests is the production of early drafts of defendant's play on the theory they might reflect copying from plaintiff's play that was disguised or deleted in later drafts. Copying deleted or so disguised as to be unrecognizable is not copying."[141]

The refusal of courts to entertain copyright infringement allegations in relation to unpublished drafts and preliminary scripts demonstrates the practical importance of a focus on expressive substitution. Because the copyright owner's rights are generally limited to the communication of their original expression to the public, even if it were not in the public domain, a filmmaker would be perfectly entitled to start with Jane Austen's *Emma* and

138. *See id.* at 435 (noting that courts routinely reject requests to consider earlier drafts of screenplays).

139. Davis v. United Artists, Inc., 547 F. Supp. 722, 724 n.9 (S.D.N.Y. 1982) (citing Fuld v. Nat'l Broad. Co., 390 F. Supp. 877, 882 n.4 (S.D.N.Y. 1975)); *see also Stromback*, 384 F.3d at 299 ("In deciding infringement claims, courts have held that only the version of the alleged infringing work presented to the public should be considered."); Madrid v. Chronicle Books, 209 F. Supp. 2d 1227, 1234 (D. Wyo. 2002) ("Since a court considers the works as they were presented to the public, discovery in this case . . . would be pointless.") (internal quotation marks omitted); *Walker*, 615 F. Supp. at 434 ("The Court considers the works as they were presented to the public.").

140. Warner Bros. v. Am. Broad. Cos., 720 F.2d 231, 241 (2d Cir. 1983) (citing 3 NIMMER ON COPYRIGHT § 13.03[B] at 13-38.1 to 38.2; Eden Toys, Inc. v. Marshall Field & Co., 675 F.2d 498, 501 (2d Cir. 1982); Durham Indus. v. Tomy Corp., 630 F.2d 904, 913 & n.11 (2d Cir. 1980)). Courts addressing the question of intermediate copying in the software context have seen the matter slightly differently. *See infra*, note 172 and accompanying text.

141. See v. Durang, 711 F.2d 141 (9th Cir. 1983).

rework the plot over and over again until she comes out with *Clueless*.[142] Intermediate scripts that never see the light of day do not communicate the author's original expression to the public and thus cannot constitute copyright infringement.

3. The Implications of Computer Software and Other Functional Works Protected by Copyright Law

Copyright protection for computer software has long been a source of controversy and disquiet.[143] Although the statutory definition of "literary works" in the Copyright Act is broad enough to include computer programs,[144] treating software as a work of literature presents something of a contradiction. The 1976 Copyright Act clearly states that copyright protection does not extend to any "process, system, [or] method of operation . . . ."[145] And yet, as made clear by a 1980 amendment to the Act, Congress intended that copyright protection would extend to computer programs.[146] The amendment defines a computer program as "a set of statements or instructions to be used directly or indirectly in a computer in order to bring about a certain result."[147] A "set of instructions" used "in order to bring about a certain result" appears to be the very essence of the "process, system, method of operation" exclusion under § 102(b).

With this contradiction in mind, it is hardly surprising that the general theory of copyright advanced in this Article—the centrality of expressive substitution—does not fit perfectly to software.[148] Users do not typically

142. CLUELESS (Paramount 1995). *See* Suzanne Ferriss, *Emma Becomes Clueless*, *in* JANE AUSTEN IN HOLLYWOOD 122 (Linda Troost & Sayre Greenfield eds. 2000).

143. *See* Jane C. Ginsburg, *Four Reasons and a Paradox: The Manifest Superiority of Copyright over Sui Generis Protection of Computer Software*, 94 COLUM. L. REV. 2559 (1994); Peter S. Menell, *An Analysis of the Scope of Copyright Protection for Application Programs*, 41 STAN. L. REV. 1045 (1989); Pamela Samuelson et al., *A Manifesto Concerning the Legal Protection of Computer Programs*, 94 COLUM. L. REV 2308 (1994).

144. 17 U.S.C. § 101 (2010) ("literary works" includes works "expressed in words, numbers, or other verbal or numerical symbols or indicia").

145. 17 U.S.C. § 102(b). Exclusive rights in processes and methods of operation are generally left to the patent system. *See* 35 U.S.C. § 101 (2010).

146. S*ee* Apple Computer, Inc. v. Franklin Computer Corp., 714 F.2d 1240, 1247–49 (3d Cir. 1983) (reviewing legislative history); *but see* Pamela Samuelson, *CONTU Revisited: The Case Against Copyright Protection for Computer Programs in Machine-Readable Form*, 1984 DUKE L.J. 663.

147. Computer Software Copyright Act of 1980, Pub. L. No. 96-517, § 101, 94 Stat. 3028 (1980).

148. The same objections could be raised with respect to the copyright protection of architectural plans and the following discussion applies *mutatis mutandis* to that subject matter. The Berne Convention Implementation Act (1988) and the Architectural Works Copyright Protection Act (1990) recognize two separate forms of protection for architectural

copy copyrighted computer programs so that they can imbibe the artistry of the programmer's expression. Even if computer programs are to some extent expressive, they are predominantly functional.[149] The distinction between expressive and non-expressive uses is not intended to eviscerate copyright protection for computer software. As the preceding discussion makes clear, the rational justification for copyright is generally that it protects the author against expressive substitution. But the anomalous nature of computer software points to a different basis for attaching copyright protection and thus does not admit a defense of non-expressive use to the same extent. In sum, computer software (and other functional works that have been grafted onto copyright) should continue to be treated as exceptional—non-expressive use should not be regarded as a defense to ordinary acts of software piracy.[150]

Combined with the idea-expression distinction, this brief review of the application of the tests for substantial similarity and fragmented non-literal similarity, and the refusal of courts to apply the author's reproduction right to intermediate drafts that never see the light of day, all point in the same direction: the copyright owner's exclusive rights are limited to the right to communicate the expressive aspects of her work to the public. This point is important because once it is understood that copyright's primary function is to protect the author from the threat of expressive substitution, the case in favor of non-expressive uses becomes almost self-evident. Standing alone, a non-expressive use carries no threat of expressive substitution and such uses should thus fall outside the scope of an author's entitlement.

D. ACTIVATING THE PRINCIPLE OF NON-EXPRESSIVE USE THROUGH FAIR USE

1. Why Fair Use

The preceding discussion concentrates on *why* we should recognize a general principle that non-expressive use is non-infringing; this Section turns

works, one for architectural plans and the other for structures based on such plans. For an overview, see 1 NIMMER ON COPYRIGHT § 2.08 (2012).

149. Pamela Samuelson, *supra* note 143, at 2315–18 (explaining that "the primary source of value in a program is its behavior, not its text"); Dennis S. Karjala, *Copyright Protection Of Computer Program Structure*, 64 BROOK. L. REV. 519, 532 (1998) (arguing that computer programs "are not like dictionaries or maps, which are useful only insofar as they supply information to human beings. A computer program is not intended to be 'read' or 'understood' by its target audience, let alone appeal to a user's sense of esthetics.").

150. However, as noted below, the non-expressive use analysis still provides a useful framework for understanding software reverse engineering cases. *See infra* notes 173–176 and accompanying text.

to the prescriptive implications of that principle, i.e., the question of *how* it should be recognized. The answer, in short, is that the reproduction of expressive copyrighted works for non-expressive uses requires context-specific review under the fair use doctrine for three reasons.

The first reason is simply that to hold otherwise would contradict the Copyright Act's plain language. Section 106(1) of the Act gives copyright owners the exclusive rights "to reproduce the copyrighted work in copies."[151] Copies are defined as "material objects . . . in which a work is fixed . . . and from which the work *can* be perceived, reproduced or otherwise communicated."[152] Thus, to make a prima facie infringing reproduction, one need only reproduce the work in a stable format such that it is *capable* of being perceived and used expressively. There is no express requirement in the Act that anyone actually perceives the work or uses it expressively.

The second reason, as already noted, is that blanket exclusion for non-expressive use would substantially undermine the legal protection of copyright's more irregular subject matter, such as computer software and architectural plans. Applying the principle of non-expressive use to anomalous copyright subject matter must be considered carefully. Rightly or wrongly, Congress has extended copyright protection to computer software and architectural plans to provide incentives for the development of these primarily functional objects.[153] Although computer programs are treated as expressive literary works, their expressive elements are secondary to the functional output of the program—i.e., what it actually does. In consequence, the everyday use of a computer program is non-expressive, but that does not suggest that copyright protection for software should be effectively dismantled. Instead, courts must exercise caution when dealing with anomalous copyright subject matter so as not to negate the very protection Congress intended.

The third reason not to adopt a per se rule with respect to non-expressive use is that in many contexts the concept is ambiguous. Like its subject matter equivalent, the idea-expression distinction, the line between expressive use and non-expressive use may often turn out to be a matter of context and degree. Where the validity of a defendant's claim that a particular

151. 17 U.S.C. § 106(1) (2010).

152. *Id.* § 101 (emphasis added).

153. *See* Apple Computer, Inc. v. Franklin Computer Corp., 714 F.2d 1240, 1247 (3d Cir. 1983) (noting that "[a]lthough section 102(a) does not expressly list computer programs as works of authorship, the legislative history suggests that programs were considered copyrightable as literary works."); Architectural Works Copyright Protection Act, Pub. L. No 101-650, 104 Stat. 5089, 5133 (1990).

use is non-expressive is contestable, courts may find that adopting a categorical rule that non-expressive uses are non-infringing simply shifts the argument's focus from substantive issues to questions of category definition.

For these three reasons, it is submitted that the principle of non-expressive use should be applied in the context of copyright's fair use doctrine and not as a freestanding defense to copyright infringement.

2. Application to Fair Use

This Section explores how the principle of non-expressive use should be (and, implicitly, is being) applied to the traditional four-factor fair use inquiry required under § 107 of the Copyright Act.[154]

a) The "Purpose and Character" of Non-Expressive Uses

The non-expressive nature of the defendant's use is perhaps most relevant under the first fair use factor, "the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes."[155] Recognizing the copyright owner's exclusive rights as implicitly defined and limited in reference to expressive communication to the public makes sense of both expressive and non-expressive fair uses. Indeed, recognition of this overarching principle may be the key to rescuing the concept of transformative use from elastic imprecision.

According to the Supreme Court's most recent fair use decision, *Campbell v. Acuff-Rose*, the first factor turns primarily on:

> [W]hether the new use merely supersedes the objects of the original creation . . . or instead adds something new, with a further purpose or different character, altering the first with new expression, meaning, or message; it asks, in other words, whether and to what extent the new work is "transformative." . . . Although such transformative use is not absolutely necessary for a finding of fair

154. 17 U.S.C. § 107. The factors are:

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work.

Id.

155. *Id.*

> use . . . the goal of copyright, to promote science and the arts, is generally furthered by the creation of transformative works.[156]

Traditionally, courts apply the concept of transformative use to new expressive uses that "provide social benefit, by shedding light on an earlier work, and, in the process, creat[e] a new one."[157] Transformative use is most obvious when the work is itself transformed; however, in many cases courts have held that the mere recontextualization of a copyrighted work from one expressive context to another is sufficient to sustain a finding of fair use—the work itself need not be altered.[158]

Understanding the rationale for transformative use is the key to grasping the link between transformative use and non-expressive use. The privileged status of transformative uses under the fair use doctrine allows for the creation of new works from old. This is not a sufficient explanation, however, because other doctrinal levers, such as a narrower understanding of the author's exclusive right to make derivative works, could achieve the same effect.[159] Beyond a simple enthusiasm for new works based on the copyrighted work, courts accord special status to transformative uses because they do not substitute for the author's original expression—they do not merely supersede the objects of the original creation.[160] Because of this special status, the greater the extent of transformation, the less significant other factors weighing against fair use will become.[161]

Cognizant of the Supreme Court's focus on transformative uses, some courts have simply equated non-expressive with transformative. In *Perfect 10, Inc. v. Amazon, Inc.*, the court held that Google's use of thumbnails in its Internet search engine "may be more transformative than a parody because a search engine provides an entirely new use for the original work, while a

156. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994) (citations and internal quotation marks omitted); *see also* Pierre N. Leval, Commentary, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990) ("I believe the answer to the question of justification turns primarily on whether, and to what extent, the challenged use is transformative.").

157. *Campbell*, 510 U.S. at 579.

158. *See, e.g.*, Bill Graham Archives v. Dorling Kindersley Ltd., 448 F.3d 605, 609–10 (2d Cir. 2006) (holding that use of promotional posters in a rock biography was "a purpose separate and distinct from the original artistic and promotional purpose for which the images were created"); Mattel, Inc. v. Walking Mountain Prods., 353 F.3d 792, 796–98, 800–06 (9th Cir. 2003) (concluding that photos parodying Barbie by depicting "nude Barbie dolls juxtaposed with vintage kitchen appliances" was a fair use).

159. *See, e.g.*, 17 U.S.C. § 106(2).

160. *See, e.g.*, *Campbell*, 510 U.S. at 579.

161. *Id.*

parody typically has the same entertainment purpose as the original work."[162] This scenario seems to be stretching the concept of transformation beyond its natural utility. It would be better to recognize uses that do not relate to the expressive appeal of a work may find favor under the first fair use factor—whether they qualify as transformative in the expressive sense or not.

By construction, the more non-expressive the use of a copyrighted work is, the less it substitutes for the author's original expression.[163] As such, courts should regard primarily non-expressive uses as equivalent (but not identical) to highly transformative uses—their "purpose and character" is such that they do not merely supersede the objects of the original creation.[164] In addition, the same logic that dictates that the more transformative a work is, the less significant the other factors become, also applies to non-expressive uses.[165]

b) Non-Expressive Use and Commercial Fair Use

While considering the "purpose and character of the use" under the first factor, courts are instructed to consider "whether such use is of a commercial nature or is for nonprofit educational purposes."[166] The status of commercial fair use has proved to be confusing in the fair use case law, in part because it is so closely linked with the question of market substitution under the fourth factor.[167] Even if commercial entities develop and maintain copy-reliant technologies such as search engines, this does not weaken their claim to fair use.[168] If a use is non-expressive, its commercial or noncommercial nature is

162. Perfect 10, Inc. v. Amazon, Inc., 508 F.3d 1146, 1165 (9th Cir. 2007) (holding further that "even making an exact copy of a work may be transformative so long as the copy serves a different function than the original work" (citing Kelly v. Arriba Soft Corp., 336 F.3d 811, 818–19 (9th Cir. 2003)).

163. The analysis in this Section is subject to the caveat regarding computer software and other quasi-functional works discussed in Section III.C.3, *supra*.

164. *See Campbell*, 510 U.S. at 583.

165. *See id.* at 579.

166. 17 U.S.C. § 107(1) (2010).

167. Sag, *Predicting Fair Use*, *supra* note 22, at 58–61. Indeed, the Ninth Circuit's approach to commerciality in *Napster* defines the concept exclusively in terms of market substitution. *See* A&M Records v. Napster, 239 F.3d 1004, 1015 (9th Cir. 2001) (holding that "commercial use is demonstrated by a showing that repeated and exploitative unauthorized copies of copyrighted works were made to save the expense of purchasing authorized copies").

168. This assessment is reinforced by recent empirical analysis of fair use cases in U.S. district courts, Sag, *Predicting Fair Use*, *supra* note 22, at 77 (finding that there is no evidence that commercial use plays any objectively ascertainable role in determining the outcome of fair use cases). Non-commercial entities such as universities may have an especially strong claim to fair use for reasons *related to* their non-commercial status, but *not because of the status itself*. For example, copying by a university for the purposes of research or education may be

irrelevant because non-expressive uses do not substitute for the author's original expression.[169]

c) Non-Expressive Use and "Amount and Substantiality"

The degree that a use is non-expressive is also significant in terms of the third fair use factor, "the amount and substantiality of the portion used in relation to the copyrighted work as a whole."[170] Far from being linear or arithmetic in nature, proper application of the third factor is contingent upon the purpose and the effect of the defendant's use.[171] Instead of relying on a mechanical quantification of the *amount* of the original work used, the third factor asks courts to assess how much of the *value* of the original work is present in the allegedly infringing work.[172] Accordingly, the extent to which a use is non-expressive plays a vital role in the assessment of the third fair use factor. A non-expressive use does not generally substitute for the expressive value of the author's original expression, and therefore courts should view it as qualitatively insignificant under the third factor, even if it involves literal copying of an entire work.

This insight helps us make sense of the superficial conflict between Hollywood cases alleging intermediate copying and analogous Silicon Valley cases.[173] In cases involving motion pictures, courts have refused to apply the author's reproduction right to allegedly infringing intermediate drafts of screenplays. However, courts addressing the question of intermediate copying in the software context have seen the matter slightly differently.[174] In software reverse engineering cases, courts appear to take the allegation of infringement via intermediate copying seriously as a potential basis for infringement.[175] This difference is best explained by the anomalous nature of

less likely to have a market effect or may generate positive externalities, which make efficient licensing less likely.

169. *See supra* note 140 and accompanying text (noting the caveat relating to anomalous copyright subject matter such as computer software).

170. 17 U.S.C. § 107(3). This inquiry can be traced back to Justice Story's original formulation of the fair use doctrine in *Folsom v. Marsh*, 9 F. Cas. 342 (C.C. Mass. 1841) (No. 4,901). In that case, Justice Story was concerned to protect the "chief value of the original work" against the extraction of its "essential parts" through the mere "facile use of scissors" or its intellectual equivalent. *Id.* at 345.

171. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 586–87 (1994) ("the extent of permissible copying varies with the purpose and character of the use").

172. *See* Matthew Sag, *God in the Machine, A New Structural Analysis of Copyright's Fair Use Doctrine*, 11 MICH. TELECOMM. & TECH. L. REV. 381, 391 (2005).

173. The terms Hollywood and Silicon Valley are used representationally.

174. *See, e.g.*, Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1518–19 (9th Cir. 1992).

175. *Id.* at 1519. The *Sega* court found that:

computer software itself. Although software is protected under the expressive regime of copyright law, the value of software that the law is protecting relates to the function or behavior of the code, not to its expression. In contrast, a screenplay has no behavioral value beyond the communication of the author's expression to the public. Thus it makes sense that in film cases allegations of intermediate copying would be categorically dismissed, whereas in software cases the courts would take a more contextual approach and review the allegation as a question of fair use.[176]

Returning to the third factor itself, the reverse engineering cases nicely illustrate the contention that non-expressive uses do not generally substitute for the value of the work. In *Sony v. Connectix*, for example, the court acknowledged that Connectix had copied an entire section of Sony's software multiple times; however, it concluded "in a case of intermediate infringement when the final product does not itself contain infringing material, this factor is of very little weight."[177]

d) The Market Effect of Non-Expressive Uses

The fourth fair use factor is "the effect of the use upon the potential market for or value of the copyrighted work."[178] Of course, the question of market effect risks collapsing into tautology because every use by a defendant represents something that could, in theory, be licensed to the defendant if the court rules that such use is not fair use. But courts avoid this circular reasoning by limiting the abstract market to a market that is cognizable under copyright. The market harms that courts refuse to recognize illustrate again that the copyright owner's exclusive rights are limited to the communication of their original expression to the public. The case law indicates that courts exclude consideration of market effects that do not arise from expressive substitution.

> [I]ntermediate copying . . . may infringe the exclusive rights granted to the copyright owner in section 106 of the Copyright Act regardless of whether the end product of the copying also infringes those rights. If intermediate copying is permissible under the Act, authority for such copying must be found in one of the statutory provisions to which the rights granted in section 106 are subject.

Id.

176. However, the reverse engineering cases all find that the practice is fair use, suggesting that future courts might invoke a per se analysis for the sake of judicial economy. *See infra* note 179.

177. Sony Computer Entm't, Inc. v. Connectix Corp., 203 F.3d 596, 606 (9th Cir. 2000) (internal quotation marks omitted).

178. 17 U.S.C. § 107(4) (2010).

In *Campbell*, the Supreme Court quite plainly differentiated the copyright owner's general economic interests from the limited protection afforded by copyright:

> [W]hen a lethal parody, like a scathing theater review, kills demand for the original, it does not produce a harm cognizable under the Copyright Act. Because parody may quite legitimately aim at garroting the original, destroying it commercially as well as artistically, the role of the courts is to distinguish between biting criticism that merely suppresses demand and copyright infringement, which usurps it.[179]

Just as *Campbell* recognizes that criticism is outside of the copyright owner's protectable sphere of interest, the reverse engineering cases recognize that the copyright owner has no protectable interest in preventing the copying of unprotectable expression and ideas buried within its object code. Courts have consistently held that making unauthorized copies of a computer program, as a necessary step in reverse engineering, is fair use.[180] For example, in *Sony v. Connectix*, the Ninth Circuit held that although the defendant's Virtual Game Station console directly competed with Sony in the market for platforms capable of playing Sony Playstation games, the Virtual Game Station was a "legitimate competitor" in that market.[181] The court concluded that Sony's desire to control the market for gaming platforms was understandable but that "copyright law . . . *does not confer such a monopoly*."[182]

The treatment of parody and reverse engineering illustrates the exclusion of market effects that do not arise from expressive substitution. This rationale is implicit in *Campbell* where the Court notes "[p]eople ask for

179. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 591–92 (1994) (internal quotation marks and citations omitted).

180. *See, e.g.*, *Sony Computer Entm't*, 203 F.3d at 606, *cert. denied*, 531 U.S. 871 (2000) (holding that Connectix's copying of Sony's copyrighted basic input-output system (BIOS) during reverse engineering, used by Connectix to develop a software program that emulates the functioning of the Sony PlayStation console for regular computers, was fair use); Atari Games Corp. v. Nintendo of Am., Inc., 975 F.2d 832, 842–43 (Fed. Cir. 1992) (observing that Atari's reverse engineering of Nintendo's 10NES program would have been a fair use of the program, except that Atari did not possess an authorized copy of the work); Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1520 (9th Cir. 1992) (holding that Accolade's reverse engineering of Sega's video game programs in order to figure out how to make its own games compatible with Sega's Genesis system is a fair use); *see also* David A. Rice, *Copyright and Contract: Preemption After Bowers v. Baystate*, 9 ROGER WILLIAMS U. L. REV. 595, 601 n.19 (2004) (collecting cases). Circumventing encryption for the purpose of reverse engineering is also allowed under the safe harbor provisions of the DMCA. *See* 17 U.S.C. § 1201(f).

181. *Sony Computer Entm't*, 203 F.3d at 607; *see also Sega*, 977 F.2d at 1522–23.

182. *Sony Computer Entm't*, 203 F.3d at 607 (emphasis added); *see also Sega*, 977 F.2d at 1523–24.

criticism, but they only want praise."[183] Thus, "the unlikelihood that creators of imaginative works will license critical reviews or lampoons of their own productions removes such uses from the very notion of a potential licensing market."[184] This rationale is explicit in the reverse engineering cases. From the beginning of its decision in *Sony v. Connectix*, the court emphasized the importance of the idea-expression distinction: "[w]e are called upon once again to apply the principles of copyright law to computers and their software, to determine what must be protected as expression and what must be made accessible to the public as function."[185] Consistent with its decision in *Sega Enterprises v. Accolade, Inc.*,[186] the Ninth Circuit held in *Sony v. Connectix* that intermediate copying of software is fair use if the copying was necessary to gain access to the software's functional elements.[187] The court based its ruling firmly on the importance of maintaining the idea-expression distinction: "[w]e drew this distinction because the Copyright Act protects expression only, not ideas or the functional aspects of a software program . . . . Thus, the fair use doctrine preserves public access to the ideas and functional elements embedded in copyrighted computer software programs."[188]

In the case of expressive uses such as parody, and non-expressive uses such as reverse engineering, courts have consistently held that the protection that copyright affords is limited to certain cognizable markets. Transformative expressive uses do not usually affect the market in any relevant sense because the second author's expression does not substitute for that of the original author. The absence of any cognizable market effect is even more apparent in cases of non-expressive use because, to the degree that a particular use is non-expressive, it has no potential substitution effect on a cognizable copyright market.

As established earlier in this Part, the copyright owner's exclusive rights typically hinge upon the communication of original expression to the public. Acts of copying that do not communicate the author's original expression to the public should not generally be held to constitute copyright infringement. The most appropriate method of doctrinal incorporation of the principle of non-expressive use is through the fair use doctrine. The role of expressive

183. *Campbell*, 510 U.S. at 592 (quoting SUMMERSET MAUGHAM, OF HUMAN BONDAGE 241 (Penguin ed. 1992)).

184. *Id.*

185. *Sony Computer Entm't*, 203 F.3d at 598.

186. *Sega*, 977 F.2d at 1510.

187. *Sony Computer Entm't*, 203 F.3d at 607.

188. *Id.* at 603 (citing *Sega*, 997 F.2d at 1510).

substitution is not merely compatible with the fair use doctrine; more accurately, expressive substitution is necessary to make sense of much existing case law. It is unrealistic to attempt to reduce the entirety of fair use jurisprudence into any one coherent principle. Nonetheless, the general proposition that the doctrine favors acts of copying unlikely to substitute for the copyright owner's original expression explains the majority of cases. Like transformative expressive uses, primarily non-expressive uses should generally be classified as fair uses because, by their very nature, they do not substitute for the author's original expression. Accordingly, like transformative use, non-expressive use should be favored under the first, third, and fourth factors—such uses are non-substitutive in "purpose and character," appropriate a qualitatively insignificant proportion of the value of the copyright owner's original expression, and produce no cognizable market effect under the fourth factor.[189]

IV. CONCLUSION: UNLEASH THE MACHINES

Digital technology offers powerful tools for organizing, analyzing, and searching through an otherwise overwhelming sea of information. The legality of these tools has generally been accepted in the purely online context of text-based and visual search engines and the context of software enabled plagiarism detection systems.[190] The library digitization debate brings the same issue to a new context: printed books.

The Authors Guild's campaign against the Google book search initiative came to an abrupt halt with the proposal of a class action settlement in 2008, followed by an Amended Settlement Agreement in 2009. That agreement has since been rejected by the supervising court and the legality of Google's initiative is still disputed by many authors and publishers. Google has provided electronic versions of millions of library books to the university libraries that made the paper copies initially available. Those universities must now determine how, if at all, they should use this resource. In 2008, several

189. As is so often the case, the second statutory factor does not appear to have much bite in the context of non-expressive uses, and thus does little to "separat[e] the fair use sheep from the infringing goats." Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 586 (1994). *See* Sag, *Predicting Fair Use*, *supra* note 22.

190. A.V. *ex rel.* Vanderhye v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009); Perfect 10, Inc. v. Amazon, Inc., 508 F.3d 1146 (9th Cir. 2007); Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003); Field v. Google, Inc., 412 F. Supp. 2d 1106, 1117–19 (D. Nev. 2006).

universities agreed to combine their digital collections in a shared repository called the HathiTrust.[191]

In September 2011, the Authors Guild announced that it was suing five universities and the HathiTrust for the "systematic, concerted, widespread and unauthorized reproduction and distribution of millions of copyrighted books . . . ."[192] The Guild objects to the universities' plan to distribute works for which they have been unable to locate the copyright owner, i.e., orphan works.[193] Implausibly, the Guild stakes the claim that libraries are not entitled to fair use under § 107 of the Copyright Act because libraries are the beneficiary of a more limited exemption under § 108.[194] Whether the limited reproduction and distribution of orphan works is permitted by fair use is a crucial question, but it is not the subject of this Article. Instead, this Article addresses the other aspect of the Guild's claim, the assertion that even library digitization restricted to enabling data analysis constitutes "one of the largest copyright infringements in history . . . ."[195]

The HathiTrust aims to develop and facilitate the development of data mining and analysis of its digital collection.[196] This activity would have qualified as "non-consumptive research" under the now defunct Amended Settlement Agreement ("ASA").[197] "Non-consumptive research" as defined in the ASA is a form of non-expressive use as the term is used in this Article. According to the Authors Guild, in the absence of a class action settlement

191. The HathiTrust includes material provided by Google, the Internet Archive, Microsoft, and the universities themselves. *See Our Digital Library*, HATHITRUST DIGITAL LIBRARY, http://www.hathitrust.org/digital_library (July 17, 2012).

192. Authors Guild, *Australian Society of Authors, Quebec Writers Union Sue Five U.S. Universities*, AUTHORS GUILD BLOG (Sept. 12, 2011), http://blog.authorsguild.org/2011/09/12/authors-guild-australian-society-of-authors-quebec-writers-union-sue-five-u-s-universities-2/. In October 2012, the district court released its opinion in *Authors Guild, Inc. v. HathiTrust*, 11 CV 6351 HB, 2012 WL 4808939 (S.D.N.Y. Oct. 10, 2012).

193. First Amended Complaint, Authors Guild v. HathiTrust, No. 11 Civ. 6351, (S.D.N.Y. Sept. 12, 2011).

194. *Id.*

195. *Id.* ¶ 7. Paragraph 68 of the Amended Complaint also states:

> [U]sers may search and identify bibliographic information (title, author, subject, ISBN, publisher, and year of publication) for the works contained in the HDL. HathiTrust also permits all users to search the entire text of all works in the HDL (including public domain and in-copyright works) to determine the number of times and page location(s) of any keyword or phrase found in a book.

Id. ¶ 68.

196. *Functional Objectives*, HATHITRUST DIGITAL LIBRARY (Nov. 5, 2010), http://www.hathitrust.org/objectives.

197. ASA, *supra* note 23, § 1.93.

or some express authorization by copyright owners, the creation of systems for the automated analysis of library books constitutes copyright infringement.[198] If this is correct, then the non-expressive use of copyrighted works will be impeded: the large number of permissions required and the difficulty of locating and identifying the relevant interests makes right-clearance on the scale of millions of works implausible.[199]

Where large-scale electronic text collections are available, advances in computational power and a proliferation of new text mining and visualization tools offer scholars of the humanities the chance to do what biologists, physicists, and economists have been doing for decades—analyze data.[200]

Scholars in the "Digital Humanities" believe that text mining and the computational analysis of text are vital to the progress of human knowledge in the "Information Age." The potential of these non-expressive uses of text has already been made apparent in the life sciences where researchers use a variety of text-mining tools to accelerate the identification of relevant research across disparate fields and to suggest hitherto unseen correlations or associations such as protein-protein interactions and gene-disease associations.[201]

Similar breakthroughs are on the horizon in the humanities. Traditionally, literary scholars have relied upon the close and often anecdotal study of select works. Modern computing power and the mass-digitization of texts now permits investigation of the larger literary record.

Literary analyses of digitized collections are at the core of Digital Humanities research. Large scale quantitative projects such as those being undertaken at the Stanford Literary Lab are unearthing previously unknowable information about individual works, genres, and even entire eras.[202] Digitization enhances our ability to process, mine, and ultimately

198. First Amended Complaint, *supra* note 193.

199. Imagine someone other than a phone company trying to write a new telephone book and having to ask every household for permission.

200. This paragraph and remainder of the text were written in parallel with two amicus briefs, one in *Authors Guild v. HathiTrust* and one in *Authors Guild v. Google*. Matthew Jockers, Jason Schultz, and I jointly authored these briefs. We were assisted by many people, most notably, David Hansen and Ana Enriquez.

201. Sophia Ananiadou, Douglas B. Kell & Jun-ichi Tsujii, *Text mining and its potential applications in systems biology*, 24:12 TRENDS IN BIOTECHNOLOGY 571 (2006) (citing Christian Blaschke et al., *Information extraction in molecular biology*, 3 BRIEF. BIOINFORM. 154 (2002); Toshihide Ono et al., *Automated extraction of information on protein-protein interactions from the biological literature*, 12 BIOINFORMATICS 155 (2001)).

202. The Stanford Literary Lab discusses, designs, and pursues literary research of a digital and quantitative nature. *About*, STANFORD LITERARY LAB, http://litlab.stanford.edu/ (last visited June 18, 2012).

better understand individual texts, the connections between texts, and the overall evolution of literary language. As Matthew Jockers explains, by exploring the literary record writ large, researchers can better understand the context in which individual texts exist and thereby better understand those individual texts.[203] As Franco Moretti has further noted, "a field this large cannot be understood by stitching together separate bits of knowledge about individual cases, because it isn't a sum of individual cases: it's a collective system, that should be grasped as a whole . . . ."[204] Grasping a system as a whole is not possible without the ability to make non-expressive uses of digitized text. For some, the possibility of mining huge digital archives has been a major catalyst for changing the very conception of humanities research. For others, it is a useful tool for testing old theories or suggesting new areas of inquiry.

Researchers in Information Retrieval frequently use text-mining and computer-aided classification to identify and retrieve relevant documents. Using similar techniques, researchers in the Digital Humanities use text mining and computer-aided classification to identify and retrieve relevant texts, often found in unlikely places. This enables researchers in the humanities to expand their traditional study of a few, canonical works to a study of any one of the several million books in the larger archive of literary history, an archive that has hitherto remained hidden because of the limitations of human reading. Thus, non-expressive use leads to additional expressive use and thus expands the audience (and the potential market) for individual works.[205]

Moreover, digitization also allows scholars to reimagine the relationships between texts. For example, the Google Ancient Places project links the text of public domain books such as *Gibbon's Decline and Fall of the Roman Empire* to a map of the ancient world.[206] The interface allows the user to browse the books, including the full text, at the same time as she browses a map. The places are marked on the map and hyperlinked. Again, the map itself is a non-expressive use of the underlying texts, but such use may well still lead to

203. JOCKERS, *supra* note 116.

204. MORETTI, *supra* note 115, at 4.

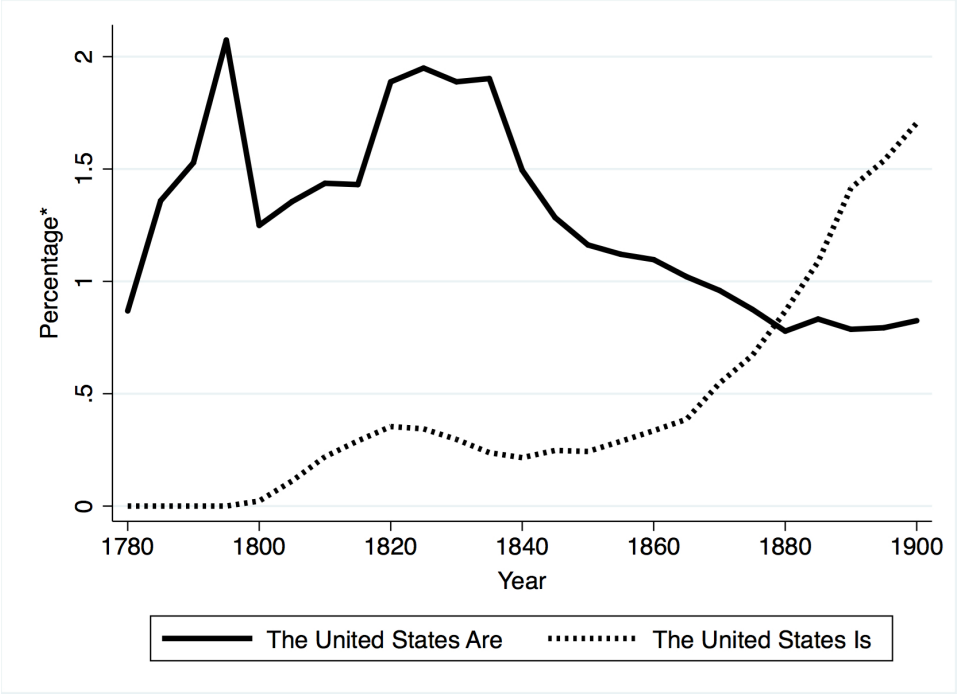
205. For example, Matthew Jockers used text-mining and computer-aided classification to identify an overlooked tradition of whaling fiction predating (and arguably informing) Melville's writing of Moby Dick. *See* JOCKERS, *supra* note 116.

206. *About*, GOOGLE ANCIENT PLACES, http://googleancientplaces.wordpress.com/about/ (last visited July 17, 2012).

additional expressive use and expansion of the audience—and, again the potential market—for individual works.[207]

The Google Ngram tool provides a simple example of such non-expressive use. The comparison of the frequency with which authors refer to the United States as a single entity ("is") versus a collection of individual states ("are") is only possible with a digitized archive of significant size and coverage.[208]

Figure 5: Google Ngram Visualization Comparing Frequency of "The United States is" to "The United States are"



207. In a similar vein, researchers at Stanford University have mapped thousands of letters exchanged during the Enlightenment and have pieced together how these individual networks fit into a complete whole they refer to as the "Republic of Letters." *See Mapping the Republic of Letters*, https://republicofletters.stanford.edu/ (last visited June 20, 2012). One such visualization yields the surprising insight that although Voltaire admired England for its tolerance, freedom and political institutions, surprisingly few letters actually went to England. *See* Patricia Cohen, *Digital Keys for Unlocking the Humanities' Riches*, N.Y. TIMES, Nov. 17, 2010, at C1.

208. *Google Books Ngram Viewer*, GOOGLE.COM, http://books.google.com/ngrams (last visited June 30, 2012). Figure 5 is a reconstruction of data generated using Google Ngram, sampled at 5-year intervals. The y-axis is scaled to 1/100,000 of a percent, such that 1=0.00001%. This particular ngram can be reproduced as follows: http://books.google.com/ngrams/graph?content=The+United+States+is%2C+The+United+States+are&year_start=1780&year_end=1900&corpus=5&smoothing=10.

Note that metadata produced in this visualization was only possible because the entire contents of the relevant books had been digitized. But note also that not a single sentence of the underlying books has been reproduced in the finished product. This kind of non-expressive use may add to our understanding, appreciation, and enjoyment of copyrighted works, but since it does not allow for the underlying works to be reconstructed, it could hardly be said to substitute for the originals.[209]

Google Ngram is just the tip of an emerging iceberg.[210] In a forthcoming book *Macroanalysis: Digital Methods and Literary History*,[211] Professor Jockers draws on a corpus of nineteenth century novels to demonstrate how literary style changes over time. By studying word frequencies, syntactic patterns, and thematic markers in the context of metadata about author nationality, author gender, and historical time period, this kind of work opens up literary study to an entirely new perspective. Thus, in the larger context of the digital archive, Jockers is able to identify both the trendsetters and the outliers. Text mining and computational analysis can lead to surprising results. For example, Jockers demonstrates that Harriet Beecher Stowe's fiction is far more similar to the work of male authors of her generation than to the typically female-authored works of sentimental fiction among which her work is generally categorized.

The macro analysis of text archives has the potential to yield specific insights into literary historical questions, such as the historic place of individual texts, authors, and genres in relation to a larger literary context; literary patterns and lexicons employed over time, across periods, within regions, or within demographic groups; the cultural and societal forces that impact literary style and the evolution of style; the waxing and waning of

209. For additional examples of the use of Ngram, see, for example, Jean-Baptiste Michel, et al., *Quantitative Analysis of Culture Using Millions of Digitized Books*, 331 SCIENCE 176 (2011) (a study of study linguistic and cultural changes in over five million digitized books) *available at* http://www.sciencemag.org/content/331/6014/176.

210. The toolkit available to Digital Humanities researchers is becoming increasingly sophisticated. *See, e.g.*, TAPOR http://www.tapor.ca/ (last visited June 30, 2012) (tools to map word usage over time, including peaks, density, collocations, and types); Andrew Kachites McCallum, *MALLET: A MAchine Learning for LanguagE Toolkit*, http://mallet.cs.umass.edu/ (last visited June 30, 2012) (a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text); *MONK: Metadata Offer New Knowledge*, http://www.monkproject.org/(last visited June 30, 2012) (a digital environment designed to help humanities scholars discover and analyze patterns in the texts); SEASR: THE SOFTWARE ENVIRONMENT FOR THE ADVANCEMENT OF SCHOLARLY RESEARCH, http://www.seasr.org (last visited June 30, 2012).

211. UIUC Press (forthcoming 2013).

literary themes; and the tastes and preferences of the literary establishment and whether those preferences correspond to general tastes and preferences.[212] *Realizing that potential requires analytical tools and capabilities and access to digitized texts*.

And yet, today's digital-minded literary scholar is shackled in time. In the absence of a policy allowing non-expressive use of copyrighted material, literary scholars, historians, and other humanists are all destined to become nineteenth-centuryists: slaves not to history, but to the public domain. To do their work thoroughly and completely—to study literary history, cultural history, and the human record writ large—these scholars simply must have access to the source material of literary, cultural, and human history. This history does not and should not end in 1923.[213]

One of the aims of this Article is to disentangle the library digitization issue for the purposes of data analysis from the broader orphan works debate. There is no orphan works problem for library digitization-search because the copyright owners are not implicated by digitization for the purpose of non-expressive use. The distinction between expressive and non-expressive *works* is already well recognized in copyright law as the gatekeeper to copyright protection. As this Article has shown, the same distinction should generally be made in relation to potential *acts* of infringement. Preserving the functional force of the idea-expression distinction in the digital context requires courts to conclude that copying for purely non-expressive purposes, such as the automated extraction of data, are not infringing. Like transformative uses, such as parody and criticism, non-expressive uses should generally be classified as fair use because, by their very nature, they do not substitute for the author's original expression.

The legal status of actual copying for non-expressive uses was not a burning issue before digital technology. Outside the context of reading machines like search engines, plagiarism software, and the like, courts have quite reasonably presumed that every copy of an expressive work is for an expressive or consumptive purpose. The issue is now, however, squarely before the courts and should be addressed. To apply the words of the Ninth Circuit Court of Appeals in *Sony v. Connectix* in a different context, "[courts] are called upon once again to apply the principles of copyright law to [the use of] computers . . . , to determine what must be protected as expression and what must be made accessible to the public . . . ."[214]

212. *See* JOCKERS, *supra* note 116.

213. *Id.*

214. Sony Computer Entm't, Inc. v. Connectix Corp., 203 F.3d 596, 598 (9th Cir. 2000).

The idea-expression distinction protects the author's legitimate interest in her work while guaranteeing others the breathing space to supplement, reuse, or reinterpret the facts and ideas embodied in the work. A similar distinction should be applied to enable the non-expressive use of copyrighted works in the age of reading machines, even if those machines reproduce the text as a step in the analytical process.

APPENDIX

An unauthorized word cloud illustration of the relative frequency of words in this Article, as depicted below, would not infringe the rights of the copyright owner of this Article.

Figure 6: A Word Cloud Based on this Article[215]

215. Matthew Sag, *Orphan Works as Grist for the Data Mill*, 27 BERKELEY TECH. L.J. 1503 2012). Word cloud produced using Wordle.net. WORDLE, http://www.wordle.net/ (last visited June 30, 2012)