UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

THE AUTHORS GUILD, INC., Associational Plaintiff, BETTY MILES, JOSEPH GOULDEN, and JIM BOUTON, on behalf of themselves and all others similarly situated,	)	
Plaintiffs,	)	05 cv 08136 (DC)
v.	)	
	)	**ECF Case**
GOOGLE INC.,	)	
Defendant.	)	

BRIEF OF DIGITAL HUMANITIES AND LAW SCHOLARS AS *AMICI CURIAE* IN PARTIAL SUPPORT OF DEFENDANTS' MOTION FOR SUMMARY JUDGMENT OR IN THE ALTERNATIVE SUMMARY ADJUDICATION

SAMUELSON LAW, TECHNOLOGY & PUBLIC POLICY CLINIC
Jennifer M. Urban (*Pro Hac Vice*) (CA # 290845)
jurban@law.berkeley.edu
Babak Siavoshy (*Pro Hac Vice*) (CA # 264182)
bsiavoshy@law.berkeley.edu
Jason Schultz (CA # 212600)
jschultz@law.berkeley.edu
University of California, Berkeley, School of Law
396 Simon Hall
Berkeley, California 94720
Telephone: 510-642-6332
Facsimile: 510-643-4625

Matthew Sag
Associate Professor
Loyola University of Chicago School of Law
msag@luc.edu

Counsel for Amici

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Cases

Statutes

Secondary Sources

STATEMENT OF INTEREST OF *AMICI*

Amici are professors and scholars who teach, write, and research in the areas of either digital humanities or the law, and an association that represents Digital Humanities scholars generally.[1] Digital Humanities *Amici* have an interest in this case because of its potential impact on their ability to discover and understand, through automated means, the data in and relationships among textual works. Legal Scholar *Amici* have an interest in this case both because of its impact on socially beneficial research, and because of their interest in the sound development of intellectual property law. Resolution of the legal issue of copying for non-expressive uses has far-reaching implications for the scope of copyright protection, a subject germane to *Amici's* professional interests and one about which they have great expertise. *Amici* speak only to the issue of copying for non-expressive uses and express no opinion on other aspects of this case. A complete list of individual *amici* is attached as Appendix A.

SUMMARY OF ARGUMENT

The significance of this case extends far beyond one discrete product of a single information technology company. Mass digitization, like that employed by Google, is a key enabler of socially valuable computational and statistical research (often called "data mining" or "text mining"). While the practice of data mining has been used for several decades in traditional scientific disciplines such as astrophysics and in social sciences like economics, it has only recently become technologically and economically feasible within the humanities. This has led to a revolution, dubbed "Digital Humanities," ranging across subjects like literature and linguistics to history and philosophy. New scholarly endeavors

[1] *See* Association for Computers and the Humanities, http://www.ach.org/.

enabled by Digital Humanities advancements are still in their infancy, but have enormous potential to contribute to our collective understanding of the cultural, political, and economic relationships among various collections (or "corpuses") of works—including copyrighted works—and with society. The Court's ruling in this case on the legality of mass digitization could dramatically affect the future of work in the Digital Humanities.

In ruling on the parties' motions, the Court should recognize that text mining is a non-expressive use that presents no legally cognizable conflict with the statutory rights or interests of the copyright holders. Where, as here, the output of a database—*i.e.*, the data it produces and displays—is noninfringing, this Court should find that the creation and operation of the database itself is likewise noninfringing. The copying required to convert paper library books into a searchable digital database is properly considered a "non-expressive use" because the works are copied for reasons unrelated to their protectable expressive qualities; none of the works in question are being read by humans as they would be if sitting on the shelves of a library or bookstore.

The type of non-expressive use at issue here is common among copy-reliant technologies: for example, Internet search engines and plagiarism detection software do not read, understand, or enjoy copyrighted works, nor do they deliver these works directly to the public. Such platforms copy the works only incidentally, in order to process them as "grist for the mill"—raw materials that feed various algorithms and indices. *See* Matthew Sag, *Copyright and Copy-Reliant Technology,* 103 NW. U.L. REV. 1607 (2009); Matthew Sag, *Orphan Works as Grist for the Data Mill*, 27 BERKELEY TECH. L. J. ___ (forthcoming 2012).

Further, generating data about a copyrighted work (often called "metadata") does not infringe the original work because, as has been recognized for over a century, copyright law protects only an author's original expression, not facts. That a "fact" might pertain to an expressive works does not change its factual character—or render it an author's exclusive intellectual property under the law. Indeed, making such factual information freely available to all is crucial to the harmony between copyright law and the First Amendment—hence the existence of rules like the "idea/expression" distinction (*see* 17 U.S.C. § 102(b)), the doctrine of *scenes à faire*, and the "merger" principle.

The act of copying works into a database in order to enable the generation of metadata about those works should thus be deemed noninfringing. As numerous courts have found, making intermediate copies that enable socially beneficial noninfringing uses and/or outputs constitutes a protected "fair use" under Section 107 of the Copyright Act. *See, e.g., A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 645 (4th Cir. 2009); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1168 (9th Cir. 2007); *Sony Computer Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596, 609 (9th Cir. 2000); *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1527-28 (9th Cir. 1992). Similarly, the mass digitization of books for text-mining purposes is a form of incidental or "intermediate" copying that enables ultimately non-expressive, non-infringing, and socially beneficial uses without unduly treading on any expressive—*i.e.*, legally cognizable—uses of the works. The Court should find such copying to be fair use.

ARGUMENT

I. The Freedom to Make Non-expressive Use of Copyrighted Works is Vital to the "Progress of Science" in the Digital Humanities

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 massive amounts of data.

"Digital Humanities" scholars fervently believe that text mining and the computational analysis of text are vital to the progress of human knowledge in the current Information Age. The potential of these non-expressive uses of text has already been revealed in the life sciences, where researchers routinely use a variety of text-mining tools to facilitate the search for relevant research across disparate fields and to uncover previously unnoticed "correlations or associations such as protein-protein interactions and gene-disease associations." *See* Sophia Ananiadou et al., *Text Mining and its Potential Applications in Systems Biology,* 24 TRENDS IN BIOTECHNOLOGY 571, 571 (2006) (citing Toshihide Ono et al., *Automated Extraction of Information on Protein–Protein Interactions from the Biological Literature*, 17 BIOINFORMATICS 155 (2001) and Christian Blaschke et al. *Information Extraction in Molecular Biology*, 3 BRIEFINGS IN BIOINFORMATICS 154 (2002)).

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. Just as the power of computing transformed the study of the genome (before which, years were required to sequence a single gene) into the study of

genomics (in which entire genomes can be sequenced within days), revolutionizing our understanding of biology and medicine, computers are also transforming the humanities and bringing about vast new intellectual riches there.

Literary analyses of digitized collections are at the core of Digital Humanities research. Large-scale quantitative projects like those being undertaken at the Stanford Literary Lab are unearthing previously unknowable information about individual works, genres, and even entire eras.[2] Digitization enhances our ability to process, mine, and ultimately better understand individual texts, the connections between texts, and the evolution of literary language. As University of Nebraska Professor 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 the texts themselves. *See* Matthew Jockers, *Macroanalysis: Digital Methods for Literary History* (forthcoming February 2013). Along similar lines, Stanford University Professor Franco Moretti has noted that "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 such, as a whole . . . ." Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* 4 (2005).

Researchers working 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 are able to identify and retrieve relevant texts, often from unlikely places. Humanities researchers can thereby expand their

---

[2] The Stanford Literary Lab discusses, designs, and pursues literary research of a digital and quantitative nature. *See* Stanford Literary Lab, http://litlab.stanford.edu/ (last visited June 29, 2012).

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 humans' reading capacity. In the process, nonexpressive use leads to additional expressive use, expanding the audience (and the potential market) for individual works.[3]

Mass digitization also results in the creation of data that enables scholars to reimagine relationships between texts—for example, by linking texts with maps. Thus, Google's "Ancient Places Project" links the text of public domain books like *Gibbon's Decline and Fall of the Roman Empire* to a map of the ancient world.[4] The interface allows the user to browse the books, including the full text, at the same time as she browses a map. The places mentioned are marked on the map and hyperlinked.[5] Similar maps could be made with reference to works still under copyright—importantly, *without* ever making the text of the book available for free viewing. Extracting such data from texts to create these maps is a quintessential *nonexpressive* use of the underlying texts

---

[3] 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*.

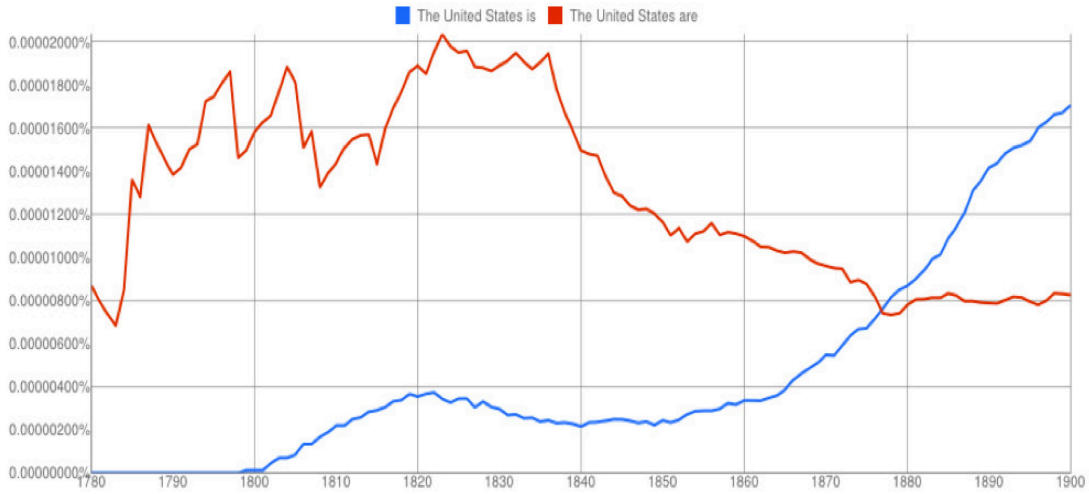
[4] Elton Barker, Eric C. Kansa, Leif Isaksen, *GAP: A Neogeo Approach To Classical Resources*, in EUROPEAN CONFERENCE ON COMPLEX SYSTEMS 2010 (Sep. 2010).

[5] In a similar vein, researchers at Stanford University have mapped thousands of letters exchanged during the Enlightenment and thereby devised a theory of how these individual networks fit into a coherent whole, which the scholars refer to as the "Republic of Letters." Mapping the Republic of Letters, https://republicofletters.stanford.edu/ (last visited July 2, 2012). Such aggregation yields surprising insights: for example, "the common narrative is that the Enlightenment started in England and spread to the rest of Europe," but the relatively low volume of correspondence between London and Paris suggests otherwise. *See* Patricia Cohen, *Digital Keys for Unlocking the Humanities' Riches*, N.Y. TIMES, Nov. 17, 2010, at C1.

that does not implicate any copyright-protected use—let alone infringe the copyrights of—the works in question.

Google's "Ngram" tool provides another example of a nonexpressive use enabled by mass digitization—this time easily visualized. Figure 1, below, is an Ngram-generated chart that compares the frequency with which authors of texts in the Google Book Search database refer to the United States as a single entity ("is") as opposed to a collection of individual states ("are"). As the chart illustrates, it was only in the latter half of the Nineteenth Century that the conception of the United States as a single, indivisible entity was reflected in the way a majority of writers referred to the nation. This is a trend with obvious political and historical significance, of interest to a wide range of scholars and even to the public at large. But this type of comparison is meaningful only to the extent that it uses as raw data a digitized archive of significant size and scope.[6]

Figure 1: Google Ngram Visualization Comparing Frequency of "The United States is" to "The United States are"



[6] Google Ngram is available at http://books.google.com/ngrams. 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.

To be absolutely clear, 1) the data used to produce this visualization can *only* be collected by digitizing the entire contents of the relevant books, and 2) not a *single sentence* of the underlying books has been reproduced in the finished product. In other words, this type of nonexpressive use only adds to our collective knowledge and understanding, without in any way replacing, damaging the value of, or interfering with the market for, the original works.[7]

Google Ngram is just the tip of the iceberg.[8] In *Macroanalysis: Digital Methods and Literary History*, Professor Jockers draws on a corpus of Nineteenth Century novels to demonstrate how literary style changes over time. *See generally* Jockers, *supra*. Examining word frequencies, syntactic patterns, and thematic markers in the metadata-enriched context of author nationality, author gender, and time period, opens up literary study to an entirely new perspective.[9] Trendsetters and outliers are revealed, as when

---

[7] For additional examples of Ngram's uses, *see*, *e.g.*, Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden; *Quantitative Analysis of Culture Using Millions of Digitized Books*. 331 SCIENCE 176 (2011) (a study of linguistic and cultural changes in over five million digitized books).

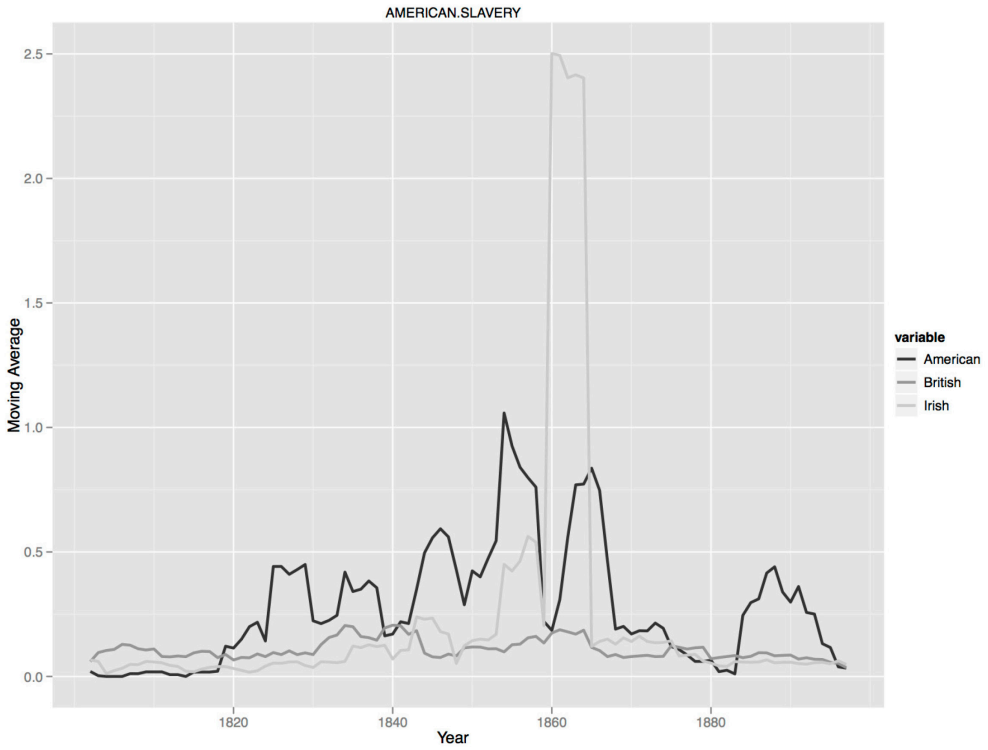
[8] The toolkit available to Digital Humanities researchers is becoming increasingly sophisticated. *See, e.g.*, Text Analysis Portal for Research ("TAPoR"), http://www.tapor.ca/portal/portal (last visited July 2, 2012) (tools to map word usage over time, including peaks, density, collocations, and types); MALLET: MAchine Learning for LanguagE Toolkit, http://mallet.cs.umass.edu/ (last visited July 2, 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 July 2, 2012) (a digital environment designed to help humanities scholars discover and analyze patterns in the texts); Software Environment for the Advancement of Scholarly Research ("SEASR"), http://seasr.org (last visited June 29, 2012).

[9] A recently published study, led by mathematicians at Dartmouth, makes a similar point. *See* James M. Hughes et al., *Quantitative Patterns of Stylistic Influence in the Evolution of Literature*, 109 PROC. OF THE NAT'L ACAD. OF SCI. OF THE U.S. 7682 (2012).

Jockers' text mining and computational analysis demonstrated that Harriet Beecher Stowe's fiction is far more similar to the work of male authors of her generation than to the female-authored works of "sentimental fiction" among which her work has traditionally been categorized. *See* Jockers, *supra*.

Figure 2 provides another fascinating example of Professor Jockers' research. The chart shows the extent to which British, American, and Irish authors focused on the theme of American slavery during the Nineteenth Century, based on a corpus of 3,450 novels from that time period. Although it comes as no surprise that slavery was most often addressed by American authors, the strong Irish reaction to the American Civil War (note the spike in the light gray line beginning in 1860) compared with the decidedly muted response by British authors invites—indeed, demands—further investigation.

Figure 2: American Slavery in American, English, and Irish Literature, 1800-1899.



As Jockers' work reveals, "macroanalysis" of text archives has the potential to provide insight into historical literary questions, such as the 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 literary themes; and the tastes and preferences of the literary establishment—and whether those preferences correspond to general tastes and preferences. However, *realizing this potential requires access to digitized texts*.

If libraries, research universities, non-profit organizations,[10] and commercial entities like Google[11] are prohibited from making nonexpressive use of copyrighted material, literary scholars, historians, and other humanists are destined to become 19th-centuryists; slaves not to history, but to the public domain. History does not end in 1923.[12] But if copyright law prevents Digital Humanities scholars from using more recent materials, that is the effective end date of the work these scholars can do.

In short, the possibility of mining huge digital archives and manipulating the data collected in the process has inspired many scholars to reconceptualize the very nature of

---

[10] *See, e.g.*, HathiTrust Digital Library, http://www.hathitrust.org/about (last visited July 3, 2012); Project Gutenberg, http://www.gutenberg.org/ (last visited July 3, 2012); *About the Internet Archive*, http://archive.org/about/about.php (last visited July 3, 2012).

[11] Google has played a significant role in facilitating research in the Digital Humanities, but the legality of nonexpressive use is an issue that transcends any one digital repository. In 2009, an international coalition of research organizations including the National Science Foundation, the Institute of Museum and Library Services, and the National Endowment for the Humanities sponsored a multi-million dollar competition to promote innovative humanities and social science research using large-scale data analysis. *See* Digging Into Data Challenge, http://www.diggingintodata.org/ (last visited July 3, 2012).

[12] Due to repeated extensions of the copyright term, U.S. copyrights after 1923 do not automatically expire on an annual basis; thus, most modern works are still copyrighted.

humanities research. For others, it has played the more modest—but still valuable—role of providing new tools for testing old theories, or suggesting new areas of inquiry. None of this, however, can be done in the Twentieth-Century context if scholars cannot make nonexpressive uses of underlying copyrighted texts, which (as shown above) will frequently number in the thousands, if not millions. Given copyright law's objective of promoting "the Progress of Science,"[13] it would be perversely counterintuitive if the promise of Digital Humanities were extinguished in the name of copyright protection.

II. COPYRIGHT LAW DOES NOT PROTECT NONEXPRESSIVE ASPECTS OF WORKS

Fortunately, this Court need not contemplate such a scenario, as nonexpressive aspects of copyrighted works—*e.g.*, the facts and ideas contained within the work and concerning it—are not protected by copyright. Such fundamental legal principles as the "idea/expression" distinction (reflected in Section 102(b) of the Copyright Act), the "merger" doctrine, the rule of "*scènes à faire*," and the "fact/expression" distinction all reflect this basic tenet. Metadata—information *about* copyrighted works collected through data mining and used by Digital Humanities scholars in the research described above—either does not implicate copyright protection at all, or is inoculated by the aforementioned doctrines that limit authors' rights to their works' expressive content.

A. The Idea/Expression Distinction

Copyright gives authors the right to set the terms upon which their original expression is made available to the public. But this right is not unlimited. As one of the fundamental—and Constitutional—limitations on those rights, the idea-expression

[13] U.S. Const. Art I., Sec. 8. "Science," as used in the Constitution, referred to knowledge and learning.

distinction strikes a balance between "the interests of authors . . . in the control and exploitation of their writings . . . on the one hand, and society's competing interest in the free flow of ideas, information, and commerce on the other hand." *Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539 (1985) (quoting *Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 429 (1984)); *see also Golan v. Holder*, 132 S. Ct. 873, 890 (2012) (describing the idea-expression distinction as one of copyright's "built-in First Amendment accommodations"). Copyright law protects only *expressive* use: "It is an axiom of copyright law that the protection granted to a copyrightable work extends only to the particular expression of an idea and never to the idea itself." *Reyher v. Children's Television Workshop*, 533 F.2d 87, 90 (2d Cir. 1976).

B. Section 102(b)

Recognizing the importance of access to ideas within expressive works, Congress has placed statutory limits on the rights of copyright holders through Section 102(b) of the Copyright Act, which provides: "In no case does copyright protection for an original work of authorship extend to any idea . . . concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work." 17 U.S.C. § 102(b) (2006). This provision has played a key role in modern copyright cases, ensuring that access to nonexpressive aspects of works is not inhibited. *See, e.g., Peter F. Gaito Architecture, LLC v. Simone Dev. Corp.*, 602 F.3d 57, 67 (2d Cir. 2010) (holding that the principle behind § 102(b) required the court "to determine whether . . . 'similarities are due to protected aesthetic expressions original to the allegedly infringed work, or whether the similarity is to something in the original that is free for the taking'" (quoting *Tufenkian Imp./Exp. Ventures, Inc. v. Einstein Moomjy, Inc.*, 338 F.3d 127, 134-

35 (2d Cir. 2003))). As noted above, text mining extracts and compiles ideas, concepts, and principles in copyrighted works into metadata. This process generates the very types of "discovery" that § 102(b) envisions.

C. Merger and *Scènes à Faire*

The policy of excluding nonexpressive elements from copyright protection is so strong that—even in situations where expressive and nonexpressive elements intertwine—doctrines like that of "merger" and "*scènes à faire*" preclude copyright protection *for expression* "in those instances where there is only one or so few ways of expressing an idea that protection of the expression would effectively accord protection to the idea itself." *Kregos v. Associated Press,* 937 F.2d 700, 705 (2d Cir. 1991); *New York Mercantile Exch., Inc. v. IntercontinentalExchange, Inc*., 497 F.3d 109, 118 (2d Cir. 2007). The "merger" doctrine is built upon the same principle as the idea/expression distinction: the protection of expressive elements of a work cannot, for Constitutional and practical reasons, interfere with the public's "free access to ideas." *New York Mercantile Exch., Inc.*, 497 F.3d. at 116. Relatedly, elements of a work that are *scènes à faire*—that is, "incidents, characters or settings which are as a practical matter indispensable, or at least standard, in the treatment of a given topic"—are not protectable. *Hoehling v. Universal City Studios, Inc*., 618 F.2d 972, 979 (2d Cir. 1980); *MyWebGrocer, LLC v. Hometown Info, Inc*., 375 F.3d 190, 194 (2d Cir. 2004) ("*Scènes à faire* are unprotectable elements [following] from a work's theme rather than from an author's creativity.").

D. Fact/Expression Distinction

Finally, the monopoly rights of authors cannot extend to factual elements that "do not owe their origin to an act of authorship." *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.,*

Inc., 499 U.S. 340, 347 (1991). "The distinction is one between creation and discovery: The first person to find and report a particular fact has not created the fact; he or she has merely discovered its existence." *Id.* The Supreme Court in *Feist* made clear that if an "author clothes facts with an original collocation of words, he or she may be able to claim a copyright in this written expression;" *nevertheless*, "[o]thers may copy the underlying facts from the publication . . . ." *Id*. at 348.

In *National Basketball Association v. Motorola, Inc.*, 105 F.3d 841 (2nd Cir. 1997), for example, a sports reporting service distributing real-time game statistics based on a data feed from reporters was held non-infringing. The Second Circuit reasoned that "[b]ecause [the service reproduced] only factual information culled from the broadcasts and none of the copyrightable expression of the games, appellants did not infringe the copyright of the broadcasts." *Id*. at 847. The Second Circuit has similarly held that one has "the right to avail himself of the facts contained in [another's] book and to use such information, whether correct or incorrect, in his own literary work." *Hoehling*, 618 F.2d at 979 (quotations omitted). In other words, copyright law clearly distinguishes between expressive and nonexpressive content, and deems only *expressive* content protectable.

E. Nonexpressive Metadata Does Not Implicate the Statutory Rights of the Copyright Holder

Metadata about a copyrighted work does not implicate any legally cognizable interest of the copyright holder. Metadata may contain facts about the works themselves, might capture (in different terminology) the ideas contained within the text, or may convey information such as the number of times a given word appears in a particular text, how often a particular author uses a specific literary device, or the essence of what the

work is about.[14] Though it is true that metadata would not exist but for the underlying work, *it does not contain the expression of the work*.

Consider, for example, two facts about *Moby Dick*: first, that the word "whale" appears 1119 times; second, that the word "dinosaur" appears 0 times. While *a whale* is certainly central to the expression contained in *Moby Dick*, this data is not. Rather, metadata of this sort—a simplified version of the metadata surveyed in Section I—is factual and nonexpressive, and incapable of infringing the rights of copyright holders.

The same principle can be illustrated using a recent decision of this court, *Warner Brothers Entertainment Inc. v. RDR Books*, 575 F. Supp. 2d 513 (S.D.N.Y. 2008). Consider the following four statements:

> [1] "Goblin-made armour does not require cleaning, simple girl. Goblins' silver repels mundane dirt, imbibing only that which strengthens it."
>
> [2] "goblin-made armor does not require cleaning, because goblins' silver repels mundane dirt, imbibing only that which strengthens it, such as basilisk venom."
>
> [3] "Statement [1] contains twenty words, and other than 'Goblin', no word in expression [1] is repeated."
>
> [4] "Statement [2] is strikingly similar to Statement [1]."

Statement [1] originates with J.K. Rowling, the author of the *Harry Potter* novels. *See Warner Bros.*, 575 F. Supp. 2d at 527 (quoting J.K. Rowling, *Harry Potter and the Deathly Hallows* 303 (2007)). Statement [2] was held out as originating with a contributor to the *Harry Potter Lexicon* (a reference work for the "*Harry Potter* universe"), which was found to infringe because too much of its contents consisted of

---

[14] See *supra*, Part I for examples of the use of such metadata.

direct quotations or close paraphrases of vivid passages in the Harry Potter books, as the comparison between [1] and [2] illustrates. *Id.* at 527. Statements [3] and [4], by contrast, are classic metadata; they would not exist but for the underlying work, and yet neither passage is substantially similar—or indeed, bears any resemblance at all—to the expressive elements of the underlying work.

Even more importantly, this metadata *does not originate with the author* of the underlying work. As the Supreme Court held in *Feist Publications*, "copying of constituent elements of the work that are *original*" is an essential element of a copyright infringement claim. 499 U.S. at 361 (emphasis added); *see also* 17 U.S.C. § 102(a) (2006) (copyright subsists only in "*original* works of authorship") (emphasis added).

Amici wish to emphasize that metadata is not the same thing as so-called "invented facts." J.K. Rowling's conception and description of goblin armor and thousands of other details in the Harry Potter series could be regarded as "invented facts" because, quite simply, she made them up. As laid out in the case law, if such facts and their associated expressive descriptions are reproduced in sufficient quantity, they may "constitute creative expression protected by copyright because characters and events spring from the imagination of the original authors." *Warner Bros.*, 575 F. Supp. 2d at 536 (quoting *Castle Rock Entm't Inc. v. Carol Publ'g Grp., Inc.*, 150 F.3d 132, 139 (2d Cir. 1998)). Metadata, however, cannot be accurately characterized as "invented facts," but only as facts *about* "invented facts." The distinction is significant: once again, facts are not eligible for copyright protection.

Nor does metadata of the sort described above not infringe the author's right "to prepare derivative works based upon the copyrighted work[.]" 17 U.S.C. § 106(2) (2006).

As this Court held in *Warner Brothers*, an analytical work that provides insight into an copyrighted work but does not "recast, transform, or adapt" that work does not violate the derivative work right. 575 F. Supp. 2d at 539; *see also Ty, Inc. v. Publ'ns Int'l Ltd.*, 292 F.3d 512, 520 (7th Cir. 2002) (holding that collectors' guide to certain copyrighted works did not violate 17 U.S.C. § 106(2) because the guides did not "recast, transform, or adapt the things to which they are guides").

Amici urge the Court to carefully distinguish the facts of the instant case from those in *Castle Rock Entertainment v. Carol Publishing Group*, 150 F.3d 132 (2d Cir. 1998). In *Castle Rock*, the Second Circuit held that a quiz book based on the popular television series "Seinfeld" was, quantitatively and qualitatively, substantially similar to that series, considered as a whole. *Id.* at 138–39. The quiz book in that case, however, was not an analytical work; rather, it essentially recast "Seinfeld"'s copyrightable characters into a new format, as if the defendant had made miniature dolls of those same characters. *See Hasbro Bradley, Inc. v. Sparkle Toys, Inc.*, 780 F.2d 189, 192-93 (2d Cir. 1985) (upholding copyrightability of "Transformer" robotic action figures as sculptural works). The supposed "facts" conveyed in the "Seinfeld" quiz book were not truly *facts* about the television program; they were "in reality fictitious expression created by *Seinfeld*'s authors." *Castle Rock Entm't*, 150 F.3d at 139.

By contrast, the many forms of metadata produced by the library digitization at the heart of this litigation do not merely recast copyrightable expression from underlying works; rather, the metadata encompasses numerous uncopyrightable facts *about* the works, such as author, title, frequency of particular words or phrases, and the like.

F. Nonexpressive Metadata Is Also Noninfringing Because It Does Not Allow the Public to Perceive the Expressive Content of a Work

The significance of public perception runs deep in copyright law. Indeed, controlling authority suggests that the copyright holder's exclusive rights are limited to the right to communicate the expressive aspects of her work to the public. For example, in *New York Times Co. v. Tasini*, 533 U.S. 483 (2001), a case about the scope of the 17 U.S.C. § 201(c) "privilege" of the copyright owner to reproduce and distribute individual contributions "as part of [a] collective work," the Supreme Court held that "[i]n determining whether the Articles [at issue] have been reproduced and distributed as part of a revision of the collective works in issue, we focus on the Articles *as presented to, and perceptible by, the user[s]* of the Databases [containing the Articles]." 533 U.S. at 499 (emphasis added; internal quotation marks and citations omitted). The Court elaborated: "the question is not whether a user can generate a revision of a collective work from a database, but whether the database itself *perceptibly presents the author's contribution* as part of a revision of the collective work." *Id.* at 504 (emphasis added).[15]

This point is especially evident in cases where plaintiffs have argued that, although a defendant's final product does not support an allegation of infringement, the defendant has violated the Copyright Act by making a reproduction of the plaintiff's work that is merely intermediate, transient, and *imperceptible by the public*. In *Davis v. United Artists, Inc.*, for example, this Court rejected out of hand the allegation that the

[15] It is true that the text of Section 106(1) of the Copyright Act, the reproduction right, has no express limitation of this sort, yet courts have focused their analysis in this way. Even where courts conclude that reproduction without public reception constitutes a prima facie case of infringement, the question of public reception remains important. *See Cambridge Univ. Press v. Becker*, No. 08 Civ. 01425, Slip Op. at 99–108, 133, 203, 235, 283 (N.D. Ga. May 11, 2012) (holding that reproduction and posting of works online, without subsequent viewing by users, was *de minimis* use and therefore not infringing).

defendant's unpublished screenplays were substantially similar to plaintiff's novel, refusing to "consider the preliminary scripts" because "the ultimate test of infringement must be the film as produced and broadcast" to the public. 547 F. Supp. 722, 724 n.9 (S.D.N.Y. 1982). *See also Fuld v. Nat'l Broad. Co., Inc.*, 390 F. Supp. 877, 882 n.4 (S.D.N.Y. 1975) ("[T]he ultimate test of infringement must be the television film as produced and broadcast — and not the preliminary scripts . . . ."); *Walker v. Time Life Films, Inc.*, 615 F. Supp. 430, 434 (S.D.N.Y. 1985) ("The Court considers the works as they were presented to the public.")[16]

III. Text Mining Creates Value by Facilitating the Advancement of Our Collective Knowledge; To Protect That Value, Mass Digitization and Similar Intermediate Copying For Data Mining and Other Nonexpressive Purposes Should Be Considered "Fair Use"

As demonstrated above, nonexpressive metadata itself is noninfringing. However, *Amici* recognize that this Court must also consider the legality of the process of making copies to generate that metadata. Fortunately, numerous courts have held that copying to enable purely nonexpressive uses, such as the automated extraction of data, does not infringe the statutory rights of the copyright holder. Like copying employed for other "transformative" purposes, such as parody, criticism, and reverse engineering, intermediate copying for the purpose of extracting nonexpressive metadata is fair use.

---

[16] Courts in other circuits have adopted the same view. *See*, *e.g.*, *Stromback v. New Line Cinema*, 384 F.3d 283, 299 (6th Cir. 2004) ("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).

A. Nonexpressive Copying to Expand Our Knowledge in the Digital Humanities Is An Activity of the Sort that Copyright Law Should Favor, Through Fair Use

First among the statutory factors relevant to a fair use analysis is "the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes." 17 U.S.C. § 107(1). The more "nonexpressive" the use of a copyrighted work, the less it substitutes for the author's original expression. As such, nonexpressive uses are properly considered equivalent (though not identical) to highly transformative uses: their "purpose and character" is such that they do not merely supersede the objects of the original creation. *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 583 (1994). *See also Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1165 (9th Cir. 2007) (holding that search engines are "highly transformative" because "[a]lthough an image may have been created originally to serve an entertainment, aesthetic, or informative function, a search engine transforms the image into a pointer directing a user to a source of information"); *Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818 (9th Cir. 2002) (holding that use of images in search engine was transformative because they served "as a tool to help index and improve access to images on the internet and their related web sites" and their use was "unrelated to any aesthetic purpose"); *Bill Graham Archives v. Dorling Kindersley Ltd.*, 448 F.3d 605, 609 (2d Cir. 2006) (finding critical to fair use analysis that publisher's use of copyrighted images of concert posters in book was "plainly different from the original purpose for which they were created"). As the process of digitization for text mining is intermediate and nonexpressive, and its purpose is to produce nonexpressive metadata, this factor favors fair use.

Moreover, "there is a strong presumption that factor one [in the fair use analysis] favors the defendant if the allegedly infringing work fits the description of uses described in [17 U.S.C.] § 107," which includes "scholarship" and "research." *NXIVM Corp. v. Ross Institute*, 364 F.3d 471, 477 (2d Cir. 2004). The crucial role that mass digitization plays in promoting the progress of research and scholarship in the Digital Humanities weighs heavily in favor of fair use here. *See also* Pierre N. Leval, *Toward A Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990) ("If [a] secondary use adds value to the original – if the quoted matter is used as raw material, transformed in the creation of new information, new aesthetics, new insights and understandings – this is the very type of activity that the fair use doctrine intends to protect for the enrichment of society.")

Similarly, courts have ruled in favor of fair use when copying allowed defendants or third parties to use facts from copyrighted works in news reporting or court proceedings. *See*, *e.g.*, *Bond v. Blum*, 317 F.3d 385, 395 (4th Cir. 2003) (holding that "the narrow purpose of defendants' use of the manuscript . . . for the evidentiary value of its content" weighed "heavily" against a finding of infringement); *Religious Tech. Ctr. v. Lerma*, 908 F. Supp. 1362, 1366 (E.D. Va. 1995) (finding fair use in part because documents were copied for "news gathering, news reporting and responding to litigation," not to "scoop" copyright owner). Significantly, both the *Bond* and *Religious Tech. Ctr.* courts' fair use holdings went further than the text mining at issue here, because the users in those cases had to glean the necessary facts by reading the materials, rather than mining the text with computers. *Bond*, 317 F.3d at 393; *Religious Tech. Ctr.*, 908 F. Supp. at 1364-65. If a human's *reading* of copyrighted expression to extract nonexpressive material is fair use, the result must be the same when a computer performs the extraction.

Finally, the commercial nature of Google Inc. should not weigh against a finding of fair use where, as here, its scanning of academic library holdings enables noninfringing uses that promote one of copyright's central purposes: the expansion of knowledge. *See*, *e.g.*, *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 638 (4th Cir. 2009) (finding that commercial character does not weigh heavily against fair use when there is a transformative or publicly beneficial purpose); *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1166 (9th Cir. 2007) (same); *Sony Computer Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596, 607 (9th Cir. 2000) (same); *Sega Enters. Ltd. v. Accolade, Inc.*, 977 F.2d 1510, 1522-23 (9th Cir. 1992) (same). As the Supreme Court has recognized, "[if] commerciality carried presumptive force against a finding of fairness, the presumption would swallow nearly all of the illustrative uses listed in the text of § 107, including news reporting, comment, criticism, teaching, scholarship, and research, since these activities 'are generally conducted for profit in this country.'" *Campbell*, 510 U.S. at 584 (internal citations omitted). As these cases make plain, because transformative uses (including the nonexpressive uses at issue in this case) do not substitute for the author's original expression, Google's commercial status carries no adverse presumption.

B. The Nature of the Works in Question Is Neutral to the Fair Use Analysis of Mass Digitization for the Advancement of Digital Humanities Research and Scholarship

When the purpose of a secondary use is socially beneficial, the second fair use factor, "the nature of the copyrighted work," is rarely dispositive. *See, e.g.*, *Bill Graham*, 448 F.3d at 612 ("The second factor may be of limited usefulness where the creative work of art is being used for a transformative purpose.") This is especially true in

"intermediate copying" cases like this one, where the material ultimately reaching the user is not the expressive content of the copyrighted work at all, but rather ideas contained within it or facts about it.

Nevertheless, to the extent that the second fair use factor is relevant here, it weighs in favor of fair use. Looking to this factor, "[c]ourts generally hold that 'the scope of the second fair use is greater with respect to factual than non-factual works'. . . . [F]ictional works, on the other hand, . . . require more protection." *Basic Books, Inc. v. Kinko's Graphics Corp.*, 758 F. Supp. 1522, 1533 (S.D.N.Y. 1991) (quoting *New Era Publications Int'l, ApS v. Carol Pub. Group*, 904 F.2d 152, 157 (2d Cir. 1990)). A detailed study of the copyrighted works in the collections from which Google has created its digitized corpus have concluded that the "overwhelming majority – 92 Percent . . . – were non fiction." Brian Lavoie & Lorcan Dempsey, *Beyond 1923: Characteristics of Potentially In Copyright Print Books in Library Collections*, 15 D-Lib Mag., http://www.dlib.org/dlib/november09/lavoie/11lavoie.html.

Furthermore, as one court explained, the second fair use factor weighs in favor of fair use where humans "cannot gain access to the unprotected ideas and functional concepts contained in [the copyrighted work] without . . . making copies." *Sega*, 977 F.2d at 1525. This is effectively the case for Digital Humanities scholars, as there are no plausible ways to conduct analyses of the sort described in Section I other than mass digitization and algorithmic analysis, both of which require making intermediate copies.

C. To the Extent Relevant, Mass Digitization Uses a Reasonable "Amount and Substantiality" of the Works in Question, in Light of the Socially Beneficial Purpose of Facilitating Data Mining for the Advancement of the Digital Humanities

The third fair use factor asks whether the amount and substantiality used are "reasonable in relation to the purpose of the copying." *Campbell*, 510 U.S. at 586–87. Because the metadata created here does not contain any infringing material, the third factor "is of very little weight." *See*, *e.g.*, *Connectix*, 203 F.3d at 606. This is true even where many intermediate copies are made. *Id.* at 601. Moreover, as Section I shows, it is not only reasonable to use mass digitization of an entire set of works to enable the creation of noninfringing metadata about those works, it is a practical necessity, as there is no equivalent human means of doing so. In order for Digital Humanities research and scholarship to be as accurate and complete as possible, every word or image in a copyrighted work must be mined.

Other courts have relied upon similar rationales to support full copying in intermediate and nonexpressive fair use cases. *See*, *e.g.*, *Vanderhye*, 562 F.3d at 642 (finding mass digitization of entire student essays to be fair use when reasonable as a means to check for plagiarism); *Perfect 10*, 508 F.3d at 1167-68 (finding thumbnail reproduction of entire photographs reasonable in light of defendant's use of the images to improve access to information on the internet versus artistic expression); *Kelly*, 336 F.3d 820-21 (same); *Bond*, 317 F.3d at 396 (noting that "[t]he use of the copyrighted material [as evidence in a custody proceeding], even the entire manuscript, does not undermine the protections granted by the [Copyright] Act"). In light of practical necessity and ample precedent in support, the Court should find that the "amount and substantiality" factor favors the making of intermediate copies for nonexpressive use.

D. Allowing Intermediate Copying in Order to Enable Nonexpressive Uses Does Not Harm the Market for the Original Works in a Legally Cognizable Manner, As The Practice Does Not Implicate the Works' Expressive Aspects in Any Way

The fourth statutory fair use factor is "the effect of the use upon the potential market for or value of the copyrighted work." 17 U.S.C. § 107(4). In the case of expressive uses such as parody, and nonexpressive uses such as reverse engineering, courts have consistently held that the protection that copyright affords is limited to certain cognizable markets. *Campbell*, 510 U.S. at 591-92 ("[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."); *Sega*, 977 F.2d at 1523-24. 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. *Campbell*, 510 U.S. at 591; *Fisher v. Dees*, 794 F.2d 432, 438 (9th Cir. 1986) ("This is not a case in which commercial substitution is likely . . . . The two works do not fulfill the same demand."). As illustrated by the examples in Section I, above, nonexpressive uses have no potential substitution effect on any legally cognizable market for copyrighted works, because copyright only protects markets for *expression*, and not markets for discoveries, ideas, facts, principles, or concepts. *See*, *e.g.*, *Vanderhye*, 562 F.3d at 644 ("[N]o market substitute was created by [defendants], whose archived student works do not supplant the plaintiffs' works . . . so much as merely suppress demand for them . . . In our view, then, any harm here is not of the kind protected against by copyright law.") Indeed, in many instances, the use of metadata made by scholars could actually enhance the market for the underlying work, by causing researchers to revisit the original work and reexamine it in more detail.

In short, there is no reason to disallow the digitization of libraries, whether by libraries themselves, or commercial search engine companies, so long as that digitization is for nonexpressive use. Nonexpressive uses such as those practiced in the Digital Humanities hold great promise for *Amici*, other scholars, society at large—and copyright owners, too.

Dated: August 3, 2012

Respectfully Submitted,

/s/ Jennifer M. Urban
Jennifer M. Urban
Samuelson Law, Technology & Public Policy Clinic
396 Simon Hall
Berkeley, CA 94720-7200
T: (510) 642-6332/ F: (510) 643-4625
jurban@law.berkeley.edu

Counsel for Amici Curiae (with Matthew Sag, Jason Schultz, and Babak Siavoshy)

APPENDIX A

Erez Lieberman Aiden
Fellow
Harvard Society of Fellows
Harvard University

Steve Anderson
Professor
School of Cinematic Arts
University of Southern California

The Association for Computers and the Humanities
http://www.ach.org

Dr. Elton Barker
Department of Classical Studies
The Open University, UK

Ann Bartow
Professor of Law
Pace Law School

Matthew Bernius
PhD Student, Cultural Anthropology, Cornell University
Researcher At Large, Open Publishing Lab @ the Rochester Institute of Technology

Jeremy Boggs
PhD Candidate, History
George Mason University

danah boyd
Berkman Center for Internet & Society
Harvard University

Annemarie Bridy
Fellow and Visiting Associate Research Scholar
Center for Information Technology Policy (CITP)
Princeton University

Susan Brown
Professor
School of English and Theatre Studies
University of Guelph
Director, Orlando Project; Project Leader, Canadian Writing Research Collaboratory

Dan L. Burk
Chancellor's Professor of Law
UC Irvine School of Law

Dr. Kate Byrne
School of Informatics
University of Edinburgh

Prof. Dr. Irene Calboli
Director, Intellectual Property and Technology Program
Marquette University Law School

Alexandra Chassanoff
Doctoral Student
School of Information and Library Science, University of North Carolina at Chapel Hill

Alan D. Corré
Emeritus Professor of Hebrew Studies
University of Wisconsin-Milwaukee

J. Stephen Downie, PhD
Associate Dean for Research
Professor
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

Penelope Eckert
Professor of Linguistics
Professor by Courtesy of Anthropology
Stanford University

Edward Finegan
Professor of Linguistics and Law
Director, Center for Excellence in Teaching
University of Southern California

Laura N. Gassaway
Paul B. Eaton Distinguished Professor of Law
University of North Carolina School of Law

Deborah R. Gerhardt
Assistant Professor of Law
UNC School of Law

David Theo Goldberg
Director and Professor

University of California
Humanities Research Institute

Eric Goldman
Director, High Tech Law Institute
Santa Clara University School of Law

Martin Holmes
Programmer/Consultant
University of Victoria

Dr. Leif Isaksen
Lecturer in Digital Humanities
University of Southampton

Matthew Jockers
Assistant Professor of English
Fellow, Center for Digital Humanities Research
University of Nebraska, Lincoln

Faye E. Jones
Professor
The Florida State University College of Law

Dan Jurafsky
Professor of Linguistics
Professor by Courtesy of Computer Science
Stanford University

Eric Kansa
UC Berkeley School of Information
Alexandria Archive Institute

Dennis S. Karjala
Jack E. Brown Professor of Law
Sandra Day O'Connor College of Law
Arizona State University

Lauren F. Klein, PhD
Assistant Professor
School of Literature, Communication, and Culture
Georgia Institute of Technology

Virginia Kuhn, PhD
Associate Director, Institute for Multimedia Literacy
Assistant Professor, School of Cinematic Arts
University of Southern California

Norma Leistiko
Reference Librarian
Hillsboro Public Library

Daryl Lim
Assistant Professor
The John Marshall Law School

Professor Jarom McDonald
Director, Office of Digital Humanities
Brigham Young University

Mark P. McKenna
Professor of Law and Notre Dame Presidential Fellow
Notre Dame Law School

Jean-Baptiste Michel
Postdoctoral Fellow
Harvard University

Franco Moretti
Professor of English
Stanford University

Lateef Mtima
Professor of Law
Howard University

Martin Mueller
Professor Emeritus of English, Classics, and Comparative Literature
Northwestern University

Alexander Nakhimovsky
Computer Science Department
Director, Linguistics Program
Director, Project Afghanistan
Colgate University

Dr. Bethany Nowviskie
University of Virginia Library

Frank A. Pasquale
Schering-Plough Professor in Health Care Regulation and Enforcement
Seton Hall Law School

Susan H. Perdue
Director, Documents Compass Virginia Foundation for the Humanities

Aaron Perzanowski
Assistant Professor
Wayne State University Law School

Malla Pollack
Co-Author, Callmann on Unfair Competition, Trademarks & Monopolies (4th ed. Thomson-Reuter)

Alex H. Poole
Doctoral Student
University of North Carolina at Chapel Hill

Kenneth M. Price
Co-director, Walt Whitman Archive
Hillegass University Professor
Department of English
University of Nebraska-Lincoln

Joseph Raben
Professor emeritus
Queens College / CUNY
Founding editor, Computers and the Humanities
Founding president, Association for Computers and the Humanities

Dr. Stephen Ramsay
Associate Professor of English
University of Nebraska-Lincoln

Stan Ruecker
Associate Professor, IIT Institute of Design, Chicago

Ivan A. Sag
Sadie Dernham Patek Professor in Humanities and Professor of Linguistics
Stanford University

Pamela Samuelson
Richard M. Sherman Distinguished Professor of Law and Professor of Information
UC Berkeley School of Law and School of Information

Ted Sichelman
Professor
University of San Diego School of Law

Jessica Silbey

Professor of Law
Suffolk University Law School

Grant Simpson
Senior Systems Analyst and PhD Candidate, Indiana University

Katherine J. Strandburg
Professor of Law
New York University School of Law

Rebecca Tushnet
Professor of Law
Georgetown Law

Deborah Tussey
Professor
Oklahoma City University School of Law

Ted Underwood
Associate Professor of English
University of Illinois, Urbana-Champaign

Thomas Wasow
Professor of Linguistics
Clarence Irving Lewis Professor of Philosophy
Stanford University

Matthew Wilkens
Professor of English
University of Notre Dame

Glen Worthey[17]
Digital Humanities Librarian
Stanford University

Vika Zafrin
Institutional Repository Librarian
Boston University

[17] Opinions expressed here are my own and not necessarily those of Stanford University, my employer.

CERTIFICATE OF SERVICE

I hereby certify that on August 3, 2012, I electronically filed the foregoing with the Clerk of the Court for the United States District Court, Southern District of New York by using the CM/ECF system. I further certify that all participants in this case are registered of the CM/ECF system and that service will be completed via the CM/ECF system.

Dated: August 3, 2012

By: /s/ Babak Siavoshy
Babak Siavoshy