Michael A. Keller

Ida M. Green University Librarian
and Director of Academic
Information Resources

Cecil H. Green Library
Stanford, California
94305-6004

Michael.Keller@stanford.edu
telephone 650-723-5553
fax 650-725-4902

The Stanford University Libraries

BY FACSIMILE & FEDERAL EXPRESS

8 September 2009

The Honorable Denny Chin
United States District Court
Southern District of New York
500 Pearl Street
New York, New York 10007

Re: *The Authors Guild et al. v. Google, Inc*. Case No. 1:05 cv 8136 (S.D.N.Y.)

Dear Judge Chin:

Request to submit Amicus Letter

Stanford University respectfully requests this Court's permission to submit this letter as an amicus curiae supporting final settlement approval in the above-referenced case .

Description of Stanford University

Founded in 1891, the Leland Stanford Junior University ("Stanford") is a leading academic research institution with approximately 15,000 students (6,800 undergraduates and 8,300 graduate students) and 1,800 faculty members.

Stanford University Libraries has amassed a collection of over 8.5 million printed volumes, in addition to hundreds of thousands of audiovisual and digital resources. There are 14 libraries within the Stanford University Libraries system, in addition to the coordinate libraries maintained by the Hoover Institution, Crown Law Library, J. Hugh Jackson Library at the Graduate School of Business, Lane Medical Library and SLAC National Accelerator Laboratory Library. Stanford also maintains several publishing concerns including Stanford University Press, which publishes about 175 books per year and HighWire Press, an electronic journal hosting service, which produces for about 150 scholarly publishers more than 1,270 peer-reviewed online journals.

Stanford University (along with the University of Michigan, Harvard University, Oxford University Press and the New York Public Library) was one of the original libraries to partner with Google in the Google Library Project part of Google Book Search to digitize, search and index the world's printed books.

Statement of Support of Proposed Settlement

After legal review of the Proposed Settlement both by outside attorneys and Stanford counsel, Stanford University finds the settlement to be "fair, reasonable and adequate." Fed. R. Civ. P. § 23(e)(2). Stanford defers to Google and others to provide legal analysis and support for the settlement itself; the purpose of this letter is to express general support for the Settlement Agreement and highlight one aspect of the Proposed Settlement: the Research Corpus.

Stanford's View of the Role of Digital Information

Before turning specifically to the Research Corpus, we offer our viewpoint regarding the role digitization plays in access to information. Within the appropriate bounds of copyright law, as an academic institution we believe that not only our students, but every student and researcher, benefits from easy, uncomplicated and remote access to Stanford's library collection and all of the great library collections in the Google Library Project. Every aspect of the human and world condition is improved through increased access to information: every child should have access to a robust library; every teacher should have access to online teaching resources; every doctor should have access to the latest research results in her field; and every diplomat should have access to the literary works from other cultures (translated into a language of the diplomat's choice). The electronic age brings about the potential for such extraordinary and effortless access to information.

For many years Stanford University Libraries has understood the power of digitization. In the 1980s Stanford made its print card catalog available online and circulation of the collection increased by almost 50%. Stanford has also engaged in multiple digitization projects to bring about improved access to information through such electronic tools as linking, cross-indexing and associative searching. Digital information is so much more accessible to researchers than print materials, not only because of the power of remote access, but also because of electronic text identification tools to identify relevant information.

Indeed digital information is so accessible and useful that we are concerned about the marginalization of information that is not digitally available. Our experience at Stanford is that many of our students are so firmly entrenched in the digital age and rely so heavily on the Internet as a research tool, that information preserved only in print books runs the risk of being ignored by future generations.

Stanford University has a long history, along with many other universities, of contributing to the commonwealth of knowledge. In that long history and as part of its research and pedagogical programs, the university has amassed deep and rich collections of the records of civilization in its libraries. In addition to the development of new knowledge and educating young people for productive lives and leadership, Stanford sees the Google Book Search project as a means to make much of its collections available under the terms of the proposed Settlement Agreement. Under this vision, American citizens anywhere, especially in small towns with small libraries, might have new opportunities of discovery. The Google Book Search project has the promise to contribute to K-12 education by providing a rich store of information and expression for school teachers and students everywhere. Just the indexing of every word and phrase of the books digitized by the Google Book Search will open the content of thousands and thousands of libraries in the U.S. alone, thus increasing the return on the investments made on library collections. Stanford's participation in the Google Book Search project is another way for it to contribute to the betterment of the lives of all Americans. It is another way for Stanford to return to the society that supports it a multiple of the value of its on-going programs. Ratifying the Settlement Agreement will make even the smallest of American libraries expand its offerings to the rich and varied collections of some of the world's best libraries -- an incalculably beneficial opportunity that will be lost without the settlement.

Proposed Research Corpus

In any era, the current working hypotheses in most disciplines are relatively modern and generally rely only on the research of the past several decades to support theories. If one views information and knowledge as a mountain, at any given time we are at the top of the mountain with access only to the most recent layers of learning. In the form of the Research Corpus, the Proposed Settlement would uncover many layers of buried knowledge through the digital access of past works. Such unprecedented access to centrally gathered information offers extraordinary opportunities to scholars and researchers.

Section 7.2(d) of the Proposed Settlement agreement provides for the development of a Research Corpus, hosted at two institutions, and containing a digital copy of every book scanned as part of the Google Library Project (save for those books removed or withdrawn from the project by the holder of the copyright).[1] This Research Corpus has the potential of becoming a latter-day and repurposed digital Library of Alexandria – the worlds' books brought together and placed into a collective repository for non-consumptive research.

Different from our current understanding of a library, this corpus of works would not be made available for the purpose of reading the works. Instead, this group of works is intended to be made available to researchers for computational analysis. As provided in section 1.90 of the proposed agreement:

> *"Non-Consumptive Research" means 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. Categories of Non-Consumptive Research include:*
> *(a) Image Analysis and Text Extraction – Computational analysis of the Digitized image artifact to either improve the image (e.g., de-skewing) or extracting textual or structural information from the image (e.g., OCR).*
> *(b) Textual Analysis and Information Extraction – Automated techniques designed to extract information to understand or develop relationships among or within Books or, more generally, in the body of literature contained within the Research Corpus. This category includes tasks such as concordance development, collocation extraction, citation extraction, automated classification, entity extraction, and natural language processing.*
>
> *(c) Linguistic Analysis – Research that performs linguistic analysis over the*

[1] See Proposed Settlement § 1.130.

> *Research Corpus to understand language, linguistic use, semantics and syntax as they evolve over time and across different genres or other classifications of Books.*
> *(d) Automated Translation – Research on techniques for translating works from one language to another.*
> *(e) Indexing and Search – Research on different techniques for indexing and search of textual content.*

While several of these categories are self-explanatory, we highlight a few of the ways the Research Corpus could be used for scholarship:

History and literature: Research opportunities based on the corpus, spanning centuries of publications in numerous languages from numerous cultures and political systems using natural language processing include investigations of ethnic identity, class consciousness, political and moral philosophy, legal and regulatory theories and practices, creative processes in literature and the arts, generative techniques, and expressive vocabularies. Because of the huge span of time of publications in the corpus, comparative work in the reception and criticism of fiction and non-fiction works will be made easier. We see the re-opening of research themes and the opening of new themes formerly impossible or very difficult due to the problem of assembling evidence and examples. As in the case of the other examples following, the availability of such a rich collection of texts for research is unparalleled, possible leading to a new Renaissance of learning and understanding. An example of this will be found in the comparison of themes and passages of the cultural exchange among East and West by the publication of classic and modern fiction and commentary in translation. For instance, consider the matter of diffusion and adaptation of cultural emanations through the example of Haiku, more or less under continuous development in Japan since the 17th-century, but its principles absorbed into Western European poetics mainly in the 20th-century, recently becoming the preferred poetic form of the American poet laureate, Robert Hass. Another aspect of cultural exchanges awaiting deeper explication is the transmission of ideas of open markets in international trade in the taxonomies of economists, business persons, and financiers from the founding of the General Agreement on Tariff and Trade to the Doha round of the World Trade Organization’s debates.

Automated Information Searching Tool Development: The research corpus provides opportunities to develop tools to search, index and identify information through automated textual analysis. Such tools could include:

- associative searching tools – providing access to texts by statistically associating terms within individual texts and providing a relevance ranked list of similar texts based on the degrees of difference in the association;
- taxonomic searching tools -- providing access to ideas more or less independent of the exact expressions of ideas across many texts; or
- metadata searching tools
- automated assignment of hyperlinking from express and implied citations to the digital versions of the cited references – thus speeding research and facilitating critical thinking on the evidence for or against particular findings

Individuals lack the capacity to manually review all of the information in such a large data set, but searching tools harness the information and become the keys to unlocking access to information. These tools hold the potential of making undiscovered connections between different data sets. For example, associative searching tools could potentially identify similarities between two medical syndromes that had not previously been associated with each other.

Linguistics: The Google Book Search Project has already been used to understand word origin and development, and how meaning changes over time, and the Research Corpus will expand this new opportunity. In a blog written on June 17, 2009, Ken Feinstein describes his efforts to track down word origins as a researcher for a dictionary in the 1990s. About Google Book Search he states: *Most of my work, though, turns out to have been largely wasted. 15 years later, early uses I spent hours to find can be beaten within minutes using Google's Book Search. A sad (for me) example is "bow hunting". I looked through dozens of books about hunting with a bow to find an early use of that term, not to mention dozens of volumes of old magazines. The oldest citation the [dictionary] has is from 1947; that's the earliest one I could find after hours of work in 1993. Using Google Book Search today, it took me less than a minute to locate a citation from 1923. . .. A little more digging could probably locate even earlier examples. As Google Book Search expands its corpus, the date could go even further back.* In addition, the corpus will lead linguists to new and much needed advances in machine translation. Beyond that, we expect so see work proceed more rapidly on taxonomic analysis and semantic searching across languages, something nearly impossible now, but generating the possibility of comparing contrasting contemporaneous views of events, of literature, and of philosophies from different national and linguistic contexts. For example, we envisage better understanding by students and scholars of mid-Twentieth-century America of the variety of views and expressions as our nation grappled with discrepancies of confronting political systems in Nazi Germany that were inherently racist, then only a few years later validating our own political principles, established in the Enlightenment, but not fully delivered to all Americans regardless of race or national origin until the civil rights movement arose and persevered.

Protection of Works in Research Corpus

The Proposed Settlement has architected the Research Corpus so that copyright interests will be appropriately safeguarded. Prior to creation, a Host Site must enter into an agreement with the Registry incorporating the strict security standards of Article VIII of the settlement.[2] Individuals may not access the Research Corpus until becoming a “Qualified User” including having an association with a library participating in the Google Library Project and agreeing to

[2] 7.2(d)(ii).

use the Research Corpus appropriately.[3] Specifically:

> *Prior to engaging in Non-Consumptive Research, a Qualified User will file with the Host Site:*
> *(a) a Research Agenda,*
> *(b) an agreement between the Qualified User and the Host Site, as agent for the Registry, that prohibits access to and use of the Research Corpus except for permitted Non-Consumptive Research and that makes the Qualified User directly liable to the Registry for any breach of its terms, and*
> *(c) a letter from a Fully Participating Library, a Cooperating Library, the Registry, Google or the Host Site indicating that the submitting entity will accept responsibility for the Qualified User's use of the Research Corpus.*[4]

Finally, in addition to the rigorous security standards, researchers are prohibited from using information extracted from the Research Corpus for direct commercial profit.[5] Collectively the protections placed around the Research Corpus ensure the security of the information while offering multiple opportunities to qualified researchers.

[3] 1.121 "Qualified User" means a Person who
(a) wishes to conduct Non-Consumptive Research,
(b) is (i) affiliated with a Fully Participating Library or aCooperating Library or
(ii) a suitably qualified individual
(1) who has the resources to perform such Non-Consumptive Research,
(2) who has an affiliation described below,
(3) who is pre-registered by a Fully Participating Library or a Cooperating Library (*i.e.*, registered prior to conducting Non-Consumptive Research), and
(4) for whose use of the Research Corpus such Fully Participating Library or Cooperating Library takes responsibility, and
(c) is bound by an agreement described in Section 7.2(d)(xi)(2) (Research Agenda).
A for-profit entity may only be a "Qualified User" if both the Registry and Google give their prior written consent. Except as set forth in the preceding sentence, a Qualified User must have an affiliation with one of the following:
(a) an accredited United States two (2)- or four (4)-year college or university;
(b) a United States not-for-profit research organization, such as a museum, observatory or research lab;
(c) a United States governmental agency (federal, state or local); or
(d) to the extent that an individual does not come within clauses (a) through
(c) above in this Section 1.121 (Qualified User), an individual may become a "Qualified User" by demonstrating to a Fully Participating Library or a Cooperating Library that he or she (directly or through the entities with which he or she is affiliated) has the necessary capability and resources to conduct Non-Consumptive Research, provided that such individual (or the entities with which he or she is affiliated) may be required by the Registry to enter into other terms and conditions with respect to such Non-Consumptive Research and the commercial exploitation of any of the results thereof consistent with the restrictions set forth in this Settlement Agreement.

[4] 7.2(d)(xi).

[5] 7.2(d)(viii)

Conclusion

At its core, this Proposed Settlement represents exponentially improved access to information, which will be of great benefit not only to Stanford but to students and researchers worldwide. Of particular importance, but which has generated less attention than other aspects of the Proposed Settlement, is the Research Corpus. The Research Corpus promises to be a resource that will assist scholarship throughout many disciplines and could harness information in ways not previously imagined. Stanford supports the Settlement Agreement and is proud to be part of the project.

Thank you for your consideration of this submission.

Respectfully submitted,

Michael A. Keller

Michael A. Keller
  University Librarian

Lauren K Schoenthaler

Lauren K. Schoenthaler
  Senior University Counsel

C: Prof. John Etchemendy, Provost, Debra Zumwalt, General Counsel