PRELIMINARY AND INCOMPLETE DRAFT

PLEASE DO NOT CITE OR REDISTRIBUTE WITHOUT PERMISSION

SEARCH FOR TOMORROW: THE GROWING RELIANCE ON TECHNOLOGY TO LOCATE PRIOR ART

Andrew Chin[*]

ABSTRACT

Information technology has changed not only how patent examiners, applicants and the public search for prior art, but which documents they find and cite as prior art references. In this Article, I examine the adequacy of the search tools currently offered by the Patent Office to its examiners and stakeholders. In particular, it appears that the rapidly growing reliance on keyword full-text search has reduced the breadth and diversity of patent prior art and, perhaps ironically, increased reliance on the U.S. Patent Classification System.

For this analysis, a large set of citations that may reasonably be imputed to keyword search is compiled from a comprehensive patent citation database. The resulting synthetic data set substantiates various concerns about the emergence of keyword search as the dominant method for finding prior art. The Article concludes with a survey of recent developments in computer and information science that may improve the performance of the patent system's search for prior art.

I. INTRODUCTION

Visitors to Alexandria, Virginia's King Street Metro station are greeted with a rather jarring array of billboards, in which prior art[1] search firms

[*] Associate Professor, University of North Carolina School of Law. Email: chin(at)unc.edu. Web: http://andrewchin.com. The author thanks Tony Biller for helpful suggestions and Allison Dobson and Matthew Ruedy for research assistance.

[1] Prior art is evidence about the state of technology prior to the date of invention or some other critical date associated with an application for patent, such as a previously issued patent. *See generally* 35 U.S.C. § 102(a)-(b) (providing that no patent shall issue where the claimed invention was patented "before the invention thereof by the applicant for patent" or "more than one year prior to the date of the application for patent in the United States").

enjoy equal billing with fast food chains. This esoteric advertising mix is aimed at the Patent Office employees and the many patent attorneys and agents who work in the vicinity, for whom the search for a pertinent prior art reference may be every bit as pressing as the quest for a quick lunch.

While King Street may be a long way from Main Street in the advertising world, the prominent role of innovation in the high-tech economy has focused considerable public attention on substantive questions of patentability. At the same time, the Patent Office's full-text patent database and World Wide Web search engines have enabled the public to conduct reasonably thorough prior art searches and to draw their own inferences regarding the validity of millions of issued patents and published patent applications. The Patent Office has accommodated these developments recently with procedural changes that offer unprecedented opportunities for patent applicants and the general public to participate in the preexamination search for prior art. With a world of prior art only a click away, the public is poised to engage the patent system and to challenge the comparative advantage of patent examiners as never before.

The popularization of prior art search has coincided with the emergence of full-text keyword querying as the dominant search methodology. The Patent Office recently replaced most of its venerable categorized paper files ("shoes") with dedicated search terminals and Web browsers.[2] The move not only represents an important milestone in the agency's transition to a paperless examination system, but also an institutional expectation that examiners, applicants, and the public will continue to find prior art references primarily through computer-aided searching of patent documents.

Search engine technology is rapidly taking center stage as the common denominator in the search for prior art by an increasingly diverse set of actors. It is therefore well to pause at this juncture to examine the ways in which keyword search might be changing not only *how* prior art is found, but *what* prior art is found. While applicants are under a duty to disclose any prior art known to be material to patentability,[3] and examiners are

[2] *See* Patent Information Users Group, 2005 Annual Conference Report, at 3 (reporting that the Patent Office's new Public Search Facility in Alexandria "has approximately 300 public workstations that provide access to USPTO internal patent and trademark search systems" and that "[t]he paper collection of classified patents was discarded in 2003-2004"); *see also* United States Patent & Trademark Office, *Performance and Accountability Report for Fiscal Year 2004*, at 23 (illustrating the new public search facility).

[3] *See* 37 C.F.R. § 1.56 ("Each individual associated with the filing and prosecution of a patent application has a duty of candor and good faith in dealing with the Office, which includes a duty to disclose to the Office all information known to that individual to be material to patentability. . . .").

expected to conduct a thorough prior art search,[4] both operate under time and other resource constraints that make it difficult to guarantee the adequacy of the cited prior art for analyzing patentability.[5] Whether search technology is to play an effective role in alleviating these constraints will ultimately depend on whether all parties are able to use the technology to conduct a more thorough search of the available prior art.

In this Article, I present empirical evidence of the rapidly growing reliance on keyword search technology and of the resulting impact on the distribution of patents cited as prior art references. These findings suggest that an excessive reliance on keyword search technology will impoverish the breadth and diversity of patent prior art, and indicate that more advanced search tools should be made available to all concerned parties.

This Article also makes a methodological contribution to the empirical literature on patent citations, namely the development and validation of synthetic data sets that approximate the characteristics of citations found using various search methods where actual data on the utilization of search results is unavailable.

The remainder of this Article is organized as follows. Part II reviews the incremental implementation of search technology in the Patent Office and the institutional and public responses to those changes. Part III describes the development and verification of the synthetic data sets for keyword search and other search methodologies that were analyzed. Part IV summarizes the results of the analysis. Part V surveys recent results in computer and information science that may serve as the basis for advances in Patent Office search technology. Part VI concludes.

II. THE PATENT OFFICE'S USE OF SEARCH TECHNOLOGY

A. Early Implementations

The Patent Office first instituted full-text patent search capability in 1984, by installing two dedicated terminals to be shared among all examiners in the office for searching patents issued after 1976.[6] The database, USPAT, was expanded in 1991 to include patents issued between

[4] *See* 37 C.F.R. § 1.104(a)(1) ("On taking up an application for examination or a patent in a reexamination proceeding, the examiner shall make a thorough study thereof and shall make a thorough investigation of the available prior art relating to the subject matter of the claimed invention.").

[5] Patent prior art is also commonly searched in the context of an infringement search; i.e., an inquiry into whether a particular product or process may infringe an issued patent. The scope of this Article, however, is limited to patentability searches, and the term "search" as used herein refers only to patentability search.

[6] Nestor Ramirez, Director, Office of Patent Automation, U.S. Patent & Trademark Office, personal communication, May 15, 2007 (notes of telephone conference, on file with author).

1971 and 1975.[7] The Patent Office connected all of its examiners' desktop computers to the search systems in 1993 and 1994, thereby making the technology more accessible.[8] Even so, according to the Patent Office's automation director Nestor Ramirez, many examiners did not utilize the search capability, preferring to continue the practice of searching through the "shoes."[9] In 1999, however, the Patent Office introduced the Examiner Automated Search Tool (EAST) and the Web-based Examiner Search Tool (WEST) software interfaces for the examiners' desktop computers, triggering what Ramirez describes as a "big transition to the system" in 2000.[10] In 2001, USOCR, a full-text database derived from optical character recognition of scanned paper patents issued between 1920 and 1970 was made accessible through the EAST and WEST systems.[11]

Access to the full-text patent databases has historically been more limited outside the Patent Office. Online tools, including the CASSIS and APS search systems, were installed in certain designated Patent Depository Libraries beginning in the early 1980s.[12] Desktop access, however, only became available to the public in 1997 through the introduction of a Web interface to the PatFT database, which contains the full text of all patents issued on or after January 1, 1976.[13]

B. The Transition to Paperless

The effectiveness of the U.S. Patent Office's keyword search technologies came under scrutiny in June 2002, when the agency requested comments and conducted a public hearing on the decision to switch to an all-electronic public search facility. Dozens of comments were submitted in opposition to the plan, including one from the American Bar Association's Section of Intellectual Property Law.[14] The comments were generally anecdotal, but indicative of systemic problems. Some of the most common

---

[7] Id.
[8] Id.
[9] Id.
[10] Id.
[11] Id.
[12] *See* Patent & Trademark Depository Library Association, *About PTDLA*, available at http://www.ptdla.org/ptdla (visited July 18, 2007).
[13] *See id.*
[14] *See* United States Patent and Trademark Office, Public Comments Resulting From: Notice of Public Hearing and Request for Comments on the Proposed Plan for an Electronic Public Search Facility, June 4, 2002, available at http://www.uspto.gov/web/offices/com/sol/comments/epubsearch/index.html [hereinafter Public Comments on Electronic Search] (comments of Hayden Gregory) (opposing, "at least until an equivalent or better electronic system is demonstrated, the removal of the paper patent files from the PTO facilities, on the grounds that the paper files continue to be an important tool for searching patents").

concerns raised by commentators were:

- Many records in the database appear to be missing, inaccurate, or not readily accessible.[15]
- No text files were available for patents issued prior to 1971.[16]
- Keyword search is an inadequate substitute for class- and subclass-wide search in identifying relevant prior art,[17] and reliance on keyword search will lead to a growing neglect of, and diminishing reliance on, the subject classification system.[18] (One commentator, however, took the contrary position that keyword search was helpful in broadening the scope of a search beyond a particular class and subclass.[19])
- Keyword search may miss references where patent applicants and searchers use different terms to describe the same concept.[20]
- Keyword search does not support visual inspection of patent drawings[21] or searching of chemical formulae.[22]

---

[15] *See id.* (comments of Joseph Clawson, the National Intellectual Property Researchers Association, Robert B. Weir, Randy Rabin, and David Testardi).

[16] *See id.* (comments of Randy Rabin, Michael H. Minns and Mark A. Watkins).

[17] *See id.* (comments of Calvin E. VanSant, Lee Grantham, Charlotte M. Kraebel, and Donal B. Tobin).

[18] *See* Public Comments on Electronic Search, *supra* note 15 (comments of Randy Rabin and Lee Grantham).

[19] Public Hearing on Prior Art, *supra* note 25, at 47-48 (comment of Mary Helen Sears) ("[I]f the examiner who is classifying particular claims in connection with allowing the application happens to make a mistake or two, it makes it very easy to miss U.S. patent references if you're relying on the classification system to search only a particular class and subclass, and today I do believe the computer word searches that are carefully carried out even in U.S. patents can help to alleviate that problem.")

The concern that updates to the U.S. patent classification schedule are failing to keep up with technological developments has recurred in the literature. *See, e.g.*, Leah S. Larkey, *A Patent Search and Classification System*, in PROC. FOURTH ACM CONF. ON DIGITAL LIBRARIES 179, 181 (1999) (describing difficulty of training classifiers and updating schedule).

[20] *See id.* (comments of Allan M. Lowe, Esq., Michael H. Minns and Mark A. Watkins); *cf.* Dale L. Carlson & Robert A. Migliorini, *Patent Reform at the Crossroads: Experience in the Far East with Oppositions Suggests an Alternative Approach for the United States*, 7 N.C. J. L. & TECH. 261, 264 ("[T]here are certain more recently developed technologies, such as computer software and business methods, where identifying the relevant prior art is often difficult with current computerized search tools.").

[21] *See id.*

[22] *See id.* (comments of Charlotte M. Kraebel); but see U.S. Patent & Trademark Office, *Public Hearing on Issues Related to the Identification of Prior Art During the Examination of a Patent Application*, July 14, 1999, at 193, available at http://www.uspto.gov/web/offices/com/hearings/priorart/0714pato.doc (comments of Stephen Kunin) (stating that keyword searching is relatively more useful in "the chemical area where the terms are better defined").

- Examiners may cite prior art on the basis of spurious keyword search results.[23]

Particularly pointed criticism came from the National Intellectual Property Researchers Association (NIPRA), which cited a number of studies on the performance of the Patent Office's search systems. In particular, NIPRA alleged that: (1) more than 100,000 patents issued since 1971 were not text-searchable; (2) "numerous" patents that had been reclassified in the paper files had not been reclassified in the database; (3) identical search queries returned different results; and (4) the number of patents in a particular subclass in the paper files did not match the corresponding number in the database.[24]

The use of keyword search technology was also discussed during the Patent Office's July 1999 public hearing on the identification of prior art at the examination stage.[25] NIPRA's then-president James Cottone presented results from his 1997 article[26] in which he reviewed the records of 421 patentability searches his firm had conducted between 1988 and 1994 to determine how the resulting 787 prior art references had been found.[27] The study found that 358, or 45%, of the references had been found through manual searching in the Patent Office's search room; 294, or 37%, had been found through the Patent Office's online search facilities; 84, or 11%, had been found through manual searches of foreign patents and non-patent publications; and 51, or 6%, had been suggested by a Patent Office examiner.[28]

C. The Current State of the Art

Since the Patent Office's move to Alexandria in 2005, on-site access to the agency's patent prior art collections has been almost exclusively via the EAST and WEST interfaces, through which users access the USPAT and USOCR databases on LiveLink Discovery servers supplied by OpenText corporation.[29] The Patent Office provides extensive training to examiners

---

[23] *See id.* (comments of Lee Grantham, the search department manager at a mid-size patent firm) ("[O]ffice actions are being issued that cite patents that have little to do with the invention but do contain appropriate keywords.").

[24] *See id.* (comments of Robert B. Weir).

[25] U.S. Patent and Trademark Office, Public Hearing on Issues Related to the Identification of Prior Art During the Examination of a Patent Application, [hereinafter Public Hearing on Prior Art].

[26] James F. Cottone, *Online Patent Searching: A Good News Story, But Not the Whole Story*, 79 J. PAT. & TRADEMARK OFF. SOC'Y 233 (1997).

[27] *See id.* at 233-34.

[28] *See id.* at 234-35.

[29] U.S. Patent & Trademark Office, EAST TRAINING FOR PUBLIC USERS (October 2004), at 2 (describing EAST as an interface to BRS databases); Wikipedia, *BRS/Search*, available at http://en.wikipedia.org/wiki/BRS/Search (visited August 5, 2007) (explaining

and members of the public in the proper use of EAST and WEST. In addition to text searches, users are trained to retrieve and browse patent images in the LiveLink Discovery databases.

EAST and WEST support keyword searches ranging from simple single-word queries to highly complex structured queries combining keywords and phrases with class and subclass restrictions and Boolean and proximity operators. Image search queries, however, are limited to individual patent numbers and specific classes and subclasses.

The Patent Office also continues to support off-site searching of the PatFT database via the agency's Web site. The Web interface supports a somewhat narrower range of search queries than is available on EAST and WEST, in that proximity operators are not accepted, and the results from one search cannot be used to build a subsequent search.

D. Future Developments

[more to be added here]

1. The Patent Office's 21st Century Strategic Plan

Examiner work-at-home programs.

Outsourcing of prior art search.[30]

Accelerated examination. The Patent Office in August 2006 introduced an "Accelerated Examination" procedure whereby applicants who satisfy certain additional procedural requirements can expect to have their applications processed within 12 months instead of the more typical 24 to 30 months.[31] The applicant's request for accelerated examination takes the form of a "petition to make special," which previously had been limited to inventions promoting environmental quality, energy development and conservation, and countering terrorism or to applicants of advanced age or failing health.[32] Most germane to this discussion, these procedural requirements include a preexamination prior art search by the applicant and the filing of a statement identifying (1) the field of search by class and subclass and (2) the databases searched and the logical queries used to search those databases.[33] The applicant must search U.S. patents and patent

---

that BRS databases have been re-branded as OpenText's Live Link Directory Servers).

[30] For a critical evaluation of the Patent Office's early proposals, see John A. Jeffery, Comment, *Preserving the Presumption of Patent Validity: An Alternative to Outsourcing the U.S. Patent Examiner's Prior Art Search*, 52 CATH. U. L. REV. 761 (2003).

[31] *See* U.S. Patent & Trademark Office, *Changes to Practice for Petitions in Patent Applications To Make Special and for Accelerated Examination*, 71 Fed. Reg. 36323 (June 26, 2006) (announcing accelerated examination procedures and effective date of August 25, 2006).

[32] *See* 37 C.F.R. § 1.102 (2004).

[33] *See id.* at 36324, at pt. 1, ¶ 8.

applications, as well as foreign patent documents and non-patent literature, unless she can provide a justification for omitting one of these sources.[34] The applicant's search must encompass every feature of the invention as either claimed or disclosed in the patent specification.[35] The applicant must also file an "accelerated examination support document" explaining in detail how each of the references found bears on the patentability of each of the claims.[36]

The advantage of accelerated examination was illustrated by the issuance of a patent for an ink cartridge to Brother Kogyo Kabushiki Kaisha on March 13, 2007, less than six months after the September 29, 2006 filing date.[37] Many applicants may decline to pursue this approach, however, because of the additional burdens and costs of satisfying the procedural requirements[38] and the potential estoppel effects of the representations made in the search statement and support document.[39]

2. Community Activism
   Bountyquest.[40]
   Software Patent Institute.
   WikiPatent, etc.
   Peer-to-Patent.

III. DATA

The primary source data for this study was extracted from the Patent Office's PatFT database, which contains the full text of all patents issued on or after January 1, 1976 and supports keyword full-text search via the Web.[41] The study includes all U.S. utility patents issued on or before May 1, 2007, covering patent numbers 3,930,271 through 7,213,269 inclusive. Excluding withdrawn patent numbers, the full-text patent data set includes 3,266,297 patents.

The limitations on the full-text database impose some further limitations

---

[34] *See id.* at 36324, at pt. 1, ¶ 8(A).

[35] *See id.* at 36324-25, at pt. 1, ¶ 8(B).

[36] *See id.* at 36325, at pt. 1, ¶ 9.

[37] *See* David L. Schaeffer, *USPTO's Accelerated Examination Program: Speed at a Price*, Stroock Client Memorandum (March 26, 2007), at 1, available at http://www.stroock.com/sitecontent.cfm?contentID=58&itemID=501.

[38] *See id.*

[39] *See id.* at 2 ("Such statements become a part of the application record and an adversary might later try to rely on those statements to challenge the patent.").

[40] Better prior art searches could only go part of the way toward addressing the founders' concerns about patent quality, since on-sale and public use bars do not require prior art.

[41] *See* U.S. Patent & Trademark Office, Patent Full-Text and Full-Page Image Databases, available at http://www.uspto.gov/patft/ (visited July 15, 2007).

on the set of patent citations that can be analyzed in this study. While patents of any vintage can be cited as prior art, this study covers only citations to patents within the database itself: i.e., those issued on or after January 1, 1976. Thus, for a citation to be included in this study, both the citing patent and the cited patent must be numbered between 3,930,271 and 7,213,269 inclusive. The base citation data set includes 23,729,900 citations of this form.

A. Imputation of Citations to Search Methods

To characterize the influence of technology on the search for patent prior art, it would be helpful to have data identifying, for each reference cited in the patent, the search method that was used to locate the reference. The patent's prosecution history file provides a good deal of this information, including references cited by the examiner and disclosed by the applicant, patent classes and subclasses searched by the examiner, and logical keyword queries used by the examiner to search the full-text databases. Moreover, this information is more widely available than ever, as the Patent Office's move to a paperless examination system has led to the publication of scanned prosecution history files ("image file wrappers") on the agency's Web site since August 2004.[42] There is nothing in these files, however, to indicate which of the cited prior art references were found through keyword searching or the use of other search technologies. The agency generally does not make such nonpublic information regarding prior art search available even for research purposes.[43]

James Cottone's article[44] illustrates one possible approach to identifying sets of citations that were found through various search methods. Cottone identified a data set of 294 citations that were actually known to have been found through the Patent Office's online search facilities. His study was based on the nonpublic records of searches conducted by his firm, however, and is therefore neither repeatable nor extensible. Moreover, it is unclear whether the 421 patentability searches conducted by his firm were representative of prior art searches in general.

To support more general observations about the impacts of search technology, it would be desirable to generate a much larger data set based on a comprehensive analysis of the available underlying data. Accordingly,

---

[42] *See* U.S. Patent and Trademark Office, Press Release No. 04-13, *Internet Access to Patent Application Files Now Available* (Aug. 2, 2004), available at http://www.uspto.gov/web/offices/com/speeches/04-13.htm; Joseph D. Cohen, *What's Really Happening in Inter Partes Reexamination*, 87 J. PAT. & TRADEMARK OFF. SOC. 207 (2005).

[43] *See* Ramirez, *supra* note 6.

[44] *See* Cottone, *supra* note 26.

I relax the requirement of actual knowledge, and instead attempt to impute patent citations to various search methods based on other known information about the relationships between the citing and cited patents. Each of the resulting "synthetic" data sets consists of those citations in the basic data set that share a particular property in common with the citations that would actually have been found through the method under study. The properties are chosen so as to be characteristic of the method under study and weakly correlated with the characteristic properties of other methods.

For keyword search, our synthetic data set consists of all citations in the base citation data set where both the citing and cited patents contain the same "low-frequency" keyword in both their detailed description and claims sections. I define a keyword as low-frequency if it appears in these fields in 50 or fewer patents in the public PatFT database, as determined by a structured single-keyword query to the Patent Office's Web server. I conducted queries for each of the 354,984 words in the Moby Words II SINGLE.TXT word list, a widely-used public domain text file,[45] and found 29,050 low-frequency words. From this analysis, I was able to produce a list of 61,221 citations imputed to keyword search. For each of these citations, there is a corresponding low-frequency keyword, which I conjecture to have appeared in a logical query during the prior art search for the citing patent whereby the cited patent was found.

We also studied the methods of searching through forward citation tracking (i.e., locating the patents that also cite a cited patent) and backward citation tracking (i.e., locating the patents cited by a cited patent). To produce our synthetic data sets, I identified all citations in the base citation data set where the citing and cited patents both cited a third patent, or where the citing patent cited a third patent that also cited the cited patent. The 7,405,952 citations of the first type were imputed to forward citation tracking, and the 7,624,501 citations of the second type were imputed to backward citation tracking. Note that while backward citation tracking is amenable to manual (paper-based) searching, forward citation tracking is not.

Finally, I studied the method of searching through the entire primary subclass to which the citing patent was ultimately assigned. This method is amenable to manual searching, and corresponds to the time-honored tradition of browsing the shoes in the Patent Office. Our synthetic data set for classification search consists of 2,631,901 citations where the citing and cited patents were both assigned to the same class and subclass, as of the Patent Office's 2006 classification schedule.

---

[45] *See* Wikipedia, *Moby Project*, available at http://en.wikipedia.org/wiki/Moby_Project (visited July 16, 2007).

B. Discussion

1. Possible Biases and Limitations

Nonrandom sampling.

Pre-1976 data. In confining its analysis to patents available in the PatFT database, the present study does not incorporate other data that the Patent Office has made available through its public search facilities. The USPAT database, which contains the full-text of U.S. patents issued since 1971, can be accessed by examiners and the public on Patent Office workstations that run the EAST and WEST software interfaces. While additional data from patents issued between 1971 and 1975 would no doubt yield more informative results, the difficulty of conducting such an extensive study on-site in the Patent Office made it necessary to utilize the more widely available PatFT database.

Examiner- vs. applicant-generated references. Since 2001, the paper versions of U.S. patents have distinguished between prior art references cited by the examiner and those cited by the applicant for patent; however, the PatFT database does not draw this distinction. Our citation data sets are based on data extracted from the PatFT database and therefore do not distinguish between examiner- and applicant-generated references. It is therefore not possible here to determine the extent to which our conclusions relate to reliance on keyword search by examiners rather than applicants, or vice versa. Such a determination would certainly be of considerable interest, particularly in assessing the increasing involvement of applicants and the general public in the search process. Considerable additional resources would, however, be required to perform the necessary data entry tasks, and so this subject is left for future study.[46]

Multiple-word queries. In contrast to the single-word queries used to generate the synthetic data set for keyword search, most search queries are more complex, combining words and phrases with class and subclass limitations, and Boolean and proximity operators. Even so, low-frequency keywords, by their nature, contribute disproportionately to the discriminatory power of a search query when taken in conjunction with other keywords. Recognizing this fact, the Patent Office's training manuals advise users of EAST and WEST to "[s]earch for *unique* words first" and to build more complex queries from there.[47] While low-frequency keywords

[46] For empirical studies of the characteristics of examiner- and applicant-identified citations, *see, e.g.,* Juan Alcacer & Michelle Gittelman, *How Do I Know What You Know? The Role of Inventors and Examiners in the Generation of Patent Citations* (August 2004), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=548003; Bhaven N. Sampat, *Examining Patent Examination: An Analysis of Examiner and Applicant Generated Prior Art*, Working Paper (NBER Summer Institute, 2004), available at http://faculty.haas.berkeley.edu/wakeman/ba297tspring05/Sampat.pdf.

[47] U.S. Patent & Trademark Office, EAST TRAINING FOR PUBLIC USERS (October

need not play a role in every keyword search result, there does not appear to be a loss of generality in restricting the synthetic data set to citations imputed to single-word queries.[48]

Non-patent prior art. While the influence of patent search technology on the search for non-patent prior art was excluded from the present study, it is a subject worthy of further investigation, particularly in fields such as software and business methods.[49]

Changes to the USPTO classification schedule. Our study did not account for changes in the Patent Office's classification schedule, which has been amended from time to time, generally in the direction of further refinement. While the renumbering of classes and subclasses over time does not affect the validity of the synthetic data set for classification search, the refinement of subclasses may have led to the systematic omission of many earlier citations.

2. Validation

We employed various internal and external methods of validating each of the synthetic data sets.

For the keyword search set, I utilized the image file wrappers that have become available on the Patent Office Web site for the most recently issued patents. I compared a random sample of [74] citations from patents issued between January 1, 2006 and May 1, 2007,[50] and their associated conjectural keywords, with the logical search queries listed in the citing patent's Examiner's Search Strategy and Results" reports. These daily reports list each of the logical queries sent to the search engine and the number of hits returned in response in connection with the prior art search for a given patent application. I found the conjectural keyword in the reports for approximately [36 of 74, or 48.6%,] of the citing patents.

Our focus on low-frequency keywords was motivated by the general observations that search engine users tend to browse only the first part of a list of results when the list is lengthy,[51] and that short search engine queries

---

2004), at 180 (emphasis in original).

[48] *See infra* text accompanying note 50.

[49] *See, e.g.*, Allison & Lemley, Growing Complexity, at 102; Julie E. Cohen, *Reverse Engineering and the Rise of Electronic Vigilantism: Intellectual Property Implications of "Lock-Out" Programs*, 68 S. CAL. L. REV. 1091, 1179 (1995) (noting difficulty of finding software prior art).

[50] I focused on the most recently issued citing patents because many of the image file wrappers for patents issued in 2004 and 2005 appeared to be incomplete. Cf. Cohen, *supra* note 42 (noting inaccuracies in and omissions from online image file wrappers).

[51] *See, e.g.*, B.J. Jansen et al., *Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web*, 36 INFO. PROCESSING & MANAGEMENT 207 (2000) (finding that 58% of search engine users view only the first page of results).

tend to be effective only when the keywords are very specific.[52] To verify these observations with respect to this specific application, I also performed a sensitivity analysis on the keyword data set by comparing it with a larger data set that would have been derived from the inclusion of higher-frequency keywords (up to 500 hits in the PatFT database). As shown in Figure 6, I find that the search engine results for higher-frequency keywords contain on average only slightly more information than could be obtained from search engine results for lower-frequency keywords. Also, during the critical periods of search technology implementation in the Patent Office, there is an observed increase in the percentage of citations imputed to keyword search. As shown in Table 1, this trend is exhibited by both synthetic data sets, but the low-frequency keyword set accounts for most of the observed increase by itself, suggesting that the additional citations in the larger set are relatively less strongly associated with keyword search.

IV. RESULTS

A. Analysis

1. Prevalence of Citations Imputed to Citation Tracking

To calculate the trend in the relative prevalence of citation tracking over time, it is necessary to normalize the number of previously issued patents that could either identify or be identified as prior art through citation tracking. Accordingly, I apply a sliding window of 1,000,000 patent numbers to the base and synthetic data sets; i.e., a citation is included in the count if the cited patent was among the 1,000,000 patents issued immediately prior to the citing patent.

Issue Year	Total Citations	Citations Imputed to Citation Tracking			
		Backward		Forward	
		Number	%	Number	%
1990[53]	74,516	15,964	21.42	22,356	30.00
1991	458,747	99,335	21.65	140,858	30.70
1992	480,705	106,631	22.18	152,731	31.77
1993	500,352	113,092	22.60	162,355	32.45
1994	543,618	125,248	23.04	181,307	33.35

[52] *See, e.g.*, Nega Alemayehu, *Analysis of Performance Variation Using Query Expansion*, 54 J. AM. SOC'Y INFO. SCI. & TECH. 379, 380 (2003); K.L. Kwok, *Higher Precision for Two-Word Queries*, *in* PROC. 25TH ANNUAL INT'L ACM SIGIR CONF. ON RESEARCH & DEVELOPMENT IN INFORMATION RETRIEVAL 395, 395 (2002); *but see* Caroline M. Eastman, *30,000 Hits May Be Better Than 300: Precision Anomalies in Internet Searches*, 53 J. AM. SOC'Y INFO. SCI. & TECH. 879, 880 (2002) (describing "anomalies" where the first of a large set of search results is more precise than the smaller set of results from a more focused query).

[53] Partial year.

1995	566,289	135,169	23.87	194,739	34.39
1996	635,450	155,869	24.53	223,244	35.13
1997	655,047	163,012	24.89	238,797	36.45
1998	864,396	207,015	23.95	316,619	36.63
1999	865,653	201,281	23.25	321,753	37.17
2000	772,609	152,526	19.74	251,194	32.51
2001	817,032	156,405	19.14	269,621	33.00
2002	822,401	154,244	18.76	278,689	33.89
2003	847,952	151,010	17.81	295,369	34.83
2004	811,115	138,044	17.02	278,618	34.35
2005	713,267	124,342	17.43	249,671	35.00
2006	866,806	158,851	18.33	313,772	36.20
2007[54]	268,107	51,015	19.03	103,676	38.67

2. Prevalence of Citations Imputed to Keyword Search

Table 2 shows the trend in the relative prevalence of keyword search over time, based on the synthetic data set for low-frequency keywords (2-50 hits) normalized by restriction to sliding window of 1,000,000 patent numbers.

		Citations Imputed to Keyword Search			
Issue	Total	2-50 hits		51-500 hits	
Year	Citations	Number	%	Number	%
1990[55]	74,516	248	0.33	3,171	4.26%
1991	458,747	1,418	0.31	20,122	4.39%
1992	480,705	1,514	0.31	20,868	4.34%
1993	500,352	1,587	0.32	22,201	4.44%
1994	543,618	1,811	0.33	24,230	4.46%
1995	566,289	1,843	0.33	24,465	4.32%
1996	635,450	2,223	0.35	27,468	4.32%
1997	655,047	2,228	0.34	28,926	4.42%
1998	864,396	3,055	0.35	38,555	4.46%
1999	865,653	3,677	0.42	39,917	4.61%
2000	772,609	3,492	0.45	36,291	4.70%
2001	817,032	3,953	0.48	38,916	4.76%
2002	822,401	4,319	0.53	40,072	4.87%
2003	847,952	4,808	0.57	40,950	4.83%
2004	811,115	4,654	0.57	36,412	4.49%
2005	713,267	4,153	0.58	32,155	4.51%
2006	866,806	5,418	0.63	39,968	4.61%
2007[56]	268,107	1,905	0.71	12,418	4.63%

3. Imputed Search Method By Category of Subject Matter

[54] Partial year.
[55] Partial year.
[56] Partial year.

	Citations Imputed to Keyword Search		Citations Imputed to Citation Tracking Forward		Backward	
	Percent of Category	Multiple of Overall	Percent of Category	Multiple of Overall	Percent of Category	Multiple of Overall
Overall	0.258%		31.2%		32.1%	
Chemistry	0.430%	1.667	31.2%	1.000	31.1%	0.967
Communications	0.170%	0.659	24.9%	0.798	28.4%	0.884
Construction	0.237%	0.919	36.2%	1.160	33.4%	1.040
Energy	0.129%	0.500	27.7%	0.889	28.2%	0.877
Engineering	0.203%	0.789	31.1%	0.997	32.5%	1.010
Medicine	0.489%	1.897	40.0%	1.282	42.6%	1.324
Household	0.230%	0.893	33.4%	1.069	31.9%	0.992
Industrial	0.190%	0.735	33.5%	1.073	32.0%	0.996
IT	0.191%	0.739	24.3%	0.778	27.9%	0.867
Material Science	0.280%	1.085	32.6%	1.046	32.3%	1.007
Optics	0.197%	0.764	28.4%	0.910	31.7%	0.985
Packaging	0.185%	0.717	38.1%	1.222	38.7%	1.204
Physics	0.078%	0.304	20.9%	0.671	26.6%	0.827
Tools	0.166%	0.644	35.5%	1.138	34.3%	1.067
Transportation	0.183%	0.708	33.5%	1.073	31.8%	0.990

4. Classification Diversity By Category of Subject Matter

	All Citations		Keyword		Citation Tracking Forward		Backward	
	Same Class	Same Sub	Same Class	Same Sub	Same Class	Same Sub	Same Class	Same Sub
Overall	47.9%	11.1%	61.0%	22.9%	53.4%	15.3%	47.3%	11.0%
Chemistry	46.6%	11.4%	62.9%	22.9%	49.7%	14.4%	44.3%	10.2%
Communications	46.3%	7.5%	57.6%	16.7%	51.3%	10.2%	45.2%	6.8%
Construction	49.0%	12.6%	61.4%	23.9%	56.3%	17.5%	51.6%	13.6%
Energy	50.5%	13.2%	59.7%	26.5%	55.4%	17.2%	49.9%	13.1%
Engineering	34.7%	9.0%	54.6%	19.8%	42.4%	13.8%	33.2%	9.1%
Medicine	49.8%	11.7%	63.2%	23.2%	54.9%	15.4%	49.0%	11.1%
Household	57.4%	14.1%	67.2%	25.5%	64.1%	19.9%	58.9%	15.7%
Industrial	45.1%	11.7%	57.3%	22.3%	51.6%	16.0%	45.4%	11.9%
IT	41.6%	8.3%	53.3%	17.1%	46.9%	11.5%	39.7%	7.9%
Material Science	37.0%	8.8%	53.2%	21.5%	41.3%	11.8%	35.6%	8.3%
Optics	44.8%	9.2%	53.4%	22.1%	48.8%	12.2%	41.6%	8.4%
Packaging	53.3%	12.8%	62.3%	26.4%	59.9%	16.9%	54.6%	13.0%
Physics	61.5%	7.2%	67.2%	22.5%	63.3%	10.3%	56.7%	6.1%
Tools	45.1%	10.3%	52.9%	17.4%	50.4%	13.7%	45.5%	10.2%
Transportation	59.8%	17.8%	71.2%	29.7%	65.0%	22.9%	60.3%	18.4%

5. Cross-Tabulations Between Synthetic Data Sets

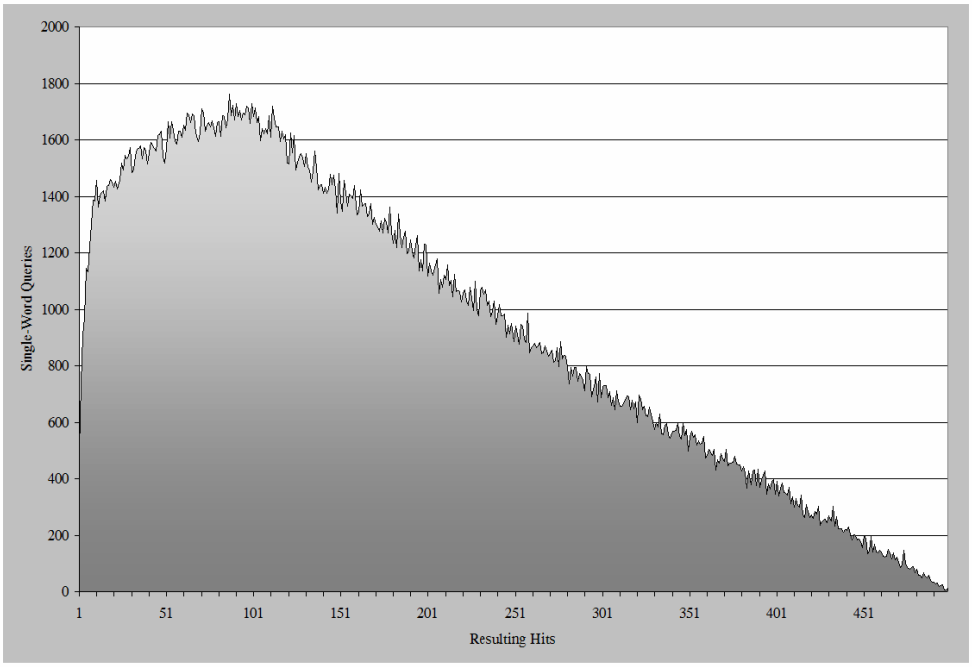
	Keyword	Forward	Backward	Classification	All
Keyword	61,221	32,250	18,140	13,997	61,221
Forward	32,250	7,405,952	2,910,858	1,126,645	7,405,952
Backward	18,140	2,910,858	7,624,501	840,098	7,624,501
Classification	13,997	1,126,645	840,098	2,631,901	2,631,901
All	61,221	7,405,952	7,624,501	2,631,901	23,729,900

6. ESSR Validation

	Matching in Sample		Spurious in Sample		All Synthetic		All Citations	
Total Citations	223		410		7,313		3,397,179	
Same Class	130	58.3%	182	44.4%	4,080	55.8%	1,427,130	42.0%
Same Subclass	28	12.6%	40	9.8%	1,329	18.2%	272,228	8.0%

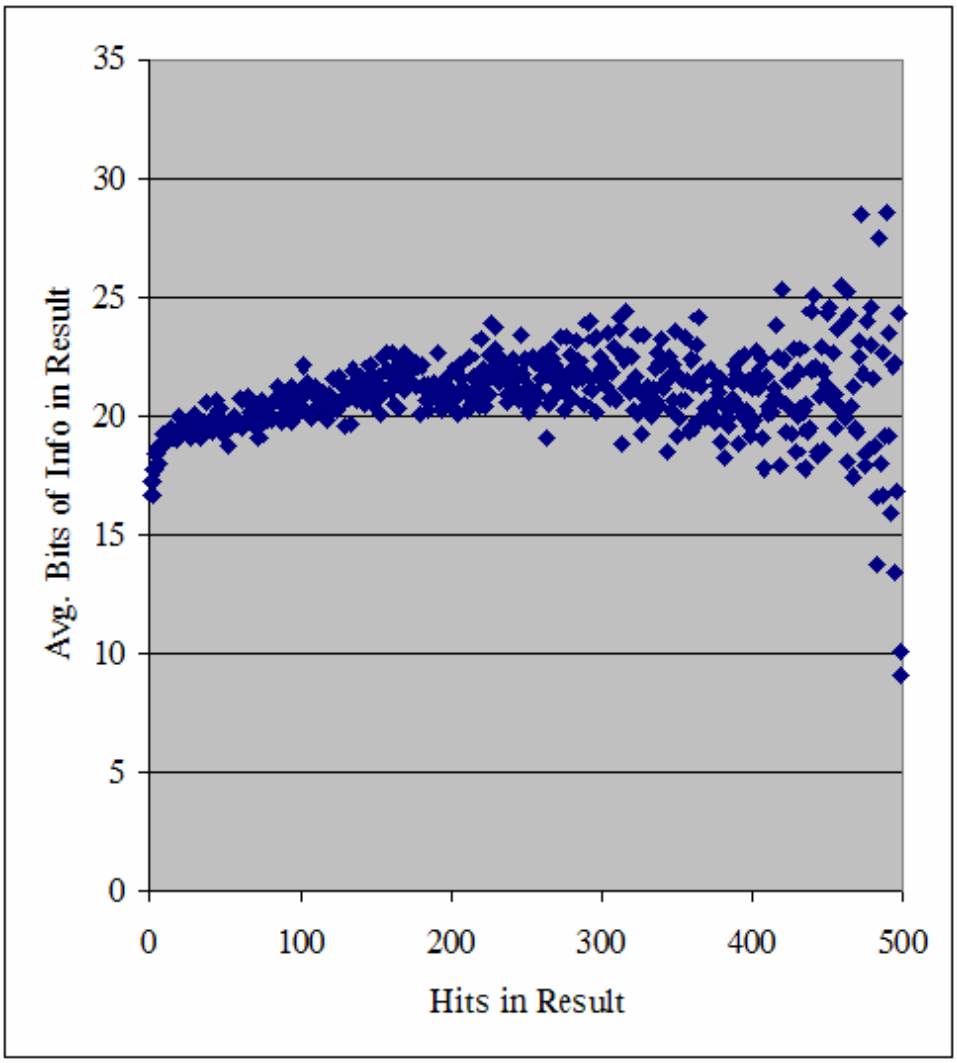
7. Distribution of Hit Counts Among Moby Dictionary Words

Figure 6 shows for each n, $2 \le n \le 500$, the number of words in the Moby SINGLE.TXT dictionary that yield n hits when used as single-keyword queries to the PatFT database.



8. Information Content of Keyword Search Results By Number of Hits

The imputed information content of a keyword search result for patent number P in which k of the n keyword hits to earlier-issued patents were actually used as patent citations is given by $H = \log_2 \frac{\binom{N}{K}}{\binom{n}{k}\binom{N-n}{K-k}}$, where $N = P - 3930270$ is the number of earlier-issued patents represented in the base citation data set and K is the number of citations in the base citation data set in which P is the citing patent. Figure 8 shows the average information content of keyword search results for each value of n, $2 \le n \le 500$.



9. Effect of Keyword Search on Years of Patents Cited

Each row in Table 9 summarizes the respective estimates for the coefficient B in linear regression models of the form

$$p = Ad + Bk + C,$$

where for each patent (observation), p is the fraction of cited patents issued during the indicated five-year interval, d is the issue year of the patent, and k is the number of times the patent appears as a citing patent in the synthetic data set for keyword search, restricted to patents issued after the terminal year of the interval.

Issue Year of Patent Reference	*B* coefficient Estimate	Standard Error	*t* statistic	*p* value
Pre-1956	–0.00144	0.000430	–3.34	0.0008
1956-60	–0.00080	0.000190	–4.22	<.0001
1961-65	–0.00123	0.000234	–5.25	<.0001
1966-70	–0.00186	0.000307	–6.05	<.0001
1971-75	–0.00151	0.000428	–3.53	0.0004
1976-80	0.00940	0.000451	20.87	<.0001
1981-85	0.00602	0.000476	12.64	<.0001
1986-90	0.00432	0.000587	7.36	<.0001
1991-95	0.00351	0.000670	5.24	<.0001
1996-2000	–0.00061	0.000809	–0.75	0.4510

10. Strandburg et al.'s Stratification Parameters

Using nonlinear regressions on a moving window of 500,000 patent numbers, I calculated estimates for coefficients α and β as described in Strandburg et al.[57] for the subgraph consisting of citations in the synthetic data set for keyword search. The nonlinear regression model for a patent's citability p as a function of k, the number of citations previously received, and l, the age of the patent measured in patent numbers, is illustrated in the example below.

$$\left.\begin{array}{c}\cdots\\(6123481,5503043)\\(6123481,5836821)\\(6123481,5916026)\\?\\\cdots\end{array}\right\} \xrightarrow{p} \left\{\begin{array}{c}(4772153,4269445)\\(4830423,4269445)\\(?,4269445)\end{array}\right.$$

$$p = A_k(k)A_l(l) = (k^{\alpha} + \gamma)l^{-\beta}$$

$$k = 2$$

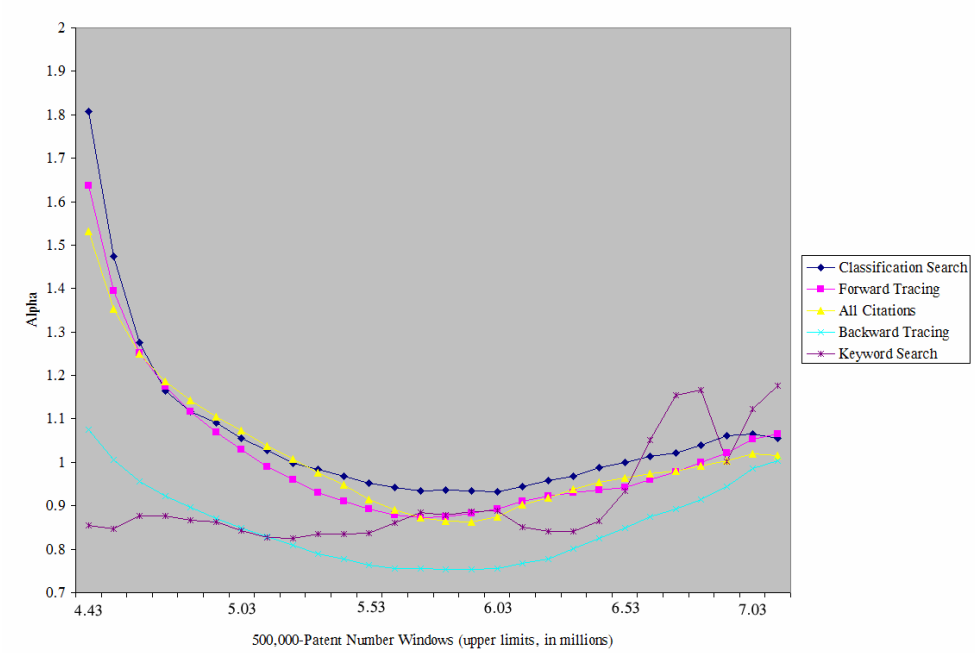
$$l = 6123481 – 4269445 = 1853046$$

Our base citation data set is substantially different from that used by

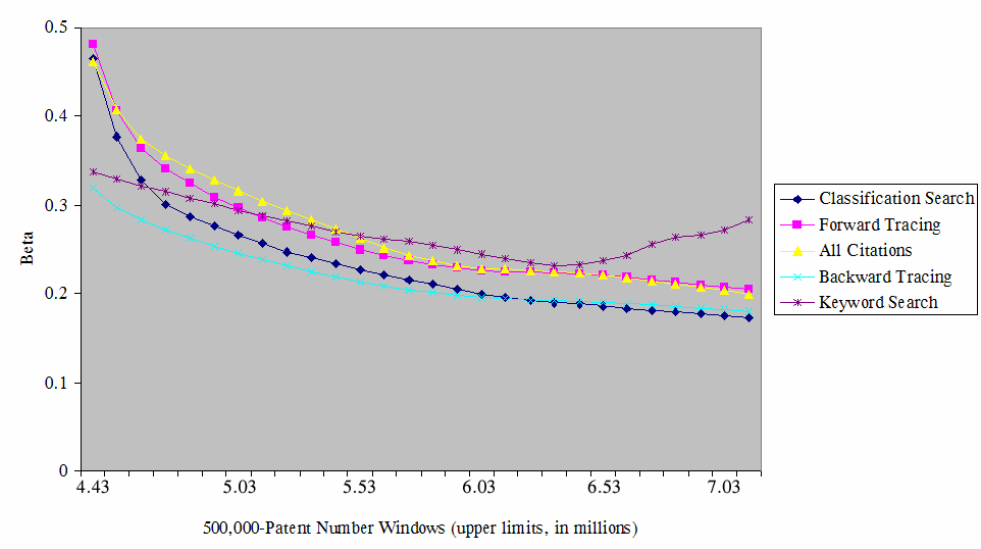
[57] Katherine J. Strandburg et al., *Law and the Science of Networks: An Overview and an Application to the "Patent Explosion,"* 21 BERKELEY TECH. L.J. 1293 (2006).

Strandburg et al., and as a result, the absolute values of α and β differ substantially from those found earlier, even though some significant observations about their relative trends over time continue to hold. With respect to these trends, Strandburg et al. suggested:

> Another possibility is that the change in citability over the years reflects a change in citation practice, rather than a change in inherent patent characteristics. This possibility is especially interesting because the timing of the increased stratification in the late 1980s corresponds to the time at which computerized searching became increasingly prevalent. We are inclined to reject this possibility at present because most of the trends in patent citation practice that we can think of — most notably the increased ease of computerized searching for prior art — seem unlikely to have changed direction from decreasing to increasing in the 1980s. Computerized searching seems likely to have had a "one-way" influence. One way to check for the influence of search technology would be to compare the behavior of the United States patent citation network with other citation networks, such as the European patent citation network or the network of citations in scientific journals.[58]



[58] Id. at 1339.



V. SOME PROMISING SEARCH TECHNOLOGIES

[more to come]

VI. CONCLUSIONS

The Manual of Patent Examining Procedure offers the following general advice to examiners about how to use the Patent Office's search technology:

> Text search can be powerful, especially where the art includes well-established terminology and the search need can be expressed with reasonable accuracy in textual terms. However, it is rare that a text search alone will constitute a thorough search of patent documents. Some combination of text search with other criteria, in particular classification, would be a normal expectation in most technologies.[59]

In providing electronic full-text access to ever-growing collections of patent documents, the Patent Office has manifestly sought to harness the power of keyword search on behalf of its examiners and the public. It is equally clear, however, that the Patent Office neither intended nor desired that keyword search would become the exclusive method for locating patent prior art.

Our data indicate that keyword search has generally led examiners to focus more narrowly on the particular class and subclass to which a patent application has been assigned. Whether the result of selecting unduly domain-specific keywords, or of combining keywords with classifications in their search queries, the rapidly growing reliance on keyword searching

[59] U.S. Patent & Trademark Office, MANUAL OF PATENT EXAMINING PROCEDURE § 904.02, at 900-49 (Aug. 2006).

to find prior art has been accompanied by a growing reliance on the U.S. Patent Classification System.

If, as many commentators suggest, there are mounting deficiencies in the classification system, there is no indication that keyword search is systematically enabling examiners to transcend them. In fact, our data suggest that examiners are taking the MPEP's suggestion to heart, in relying on the existing classification system to correct for the inaccuracies inherent in keyword search. But if both systems are flawed, then at least to some extent, the blind are leading the blind. It is time to develop a better search for tomorrow.[60]

[60] *See also Edited & Excerpted Transcript of the Symposium on Ideas Into Action: Implementing Reform of the Patent System*, 19 BERKELEY TECH. L.J. 1053 (2004) (comments of former USPTO Commissioner Q. Todd Dickinson) ("The examiners . . . need greater access to prior art, and they need better search tools. They have great search tools and they need even better search tools.")