# Comparison of Full-Text Searching to Metadata Searching for Genes in Two Biomedical Literature Cohorts

**Bradley M. Hemminger**
*206A Manning Hall, School of Information and Library Science, University of North Carolina, Chapel Hill, NC 27599–3360. E-mail: bmh@ils.unc.edu*

**Billy Saelim**
*School of Information and Library Science, University of North Carolina, Chapel Hill, NC 27599–3360*

**Patrick F. Sullivan**
*Genetics Department, School of Medicine, University of North Carolina, Chapel Hill, NC 27599–7264*

**Todd J. Vision**
*Biology Department, University of North Carolina, Chapel Hill, NC 27599–3280*

**Researchers have traditionally used bibliographic databases to search out information. Today, the full-text of resources is increasingly available for searching, and more researchers are performing full-text searches. This study compares differences in the number of articles discovered between metadata and full-text searches of the same literature cohort when searching for gene names in two biomedical literature domains. Three reviewers additionally ranked 100 articles in each domain. Significantly more articles were discovered via full-text searching; however, the precision of full-text searching also is significantly lower than that of metadata searching. Certain features of articles correlated with higher relevance ratings. A significant feature measured was the number of matches of the search term in the full-text of the article, with a larger number of matches having a statistically significant higher usefulness (i.e., relevance) rating. By using the number of hits of the search term in the full-text to rank the importance of the article, performance of full-text searching was improved so that both recall and precision were as good as or better than that for metadata searching. This suggests that full-text searching alone may be sufficient, and that metadata searching as a surrogate is not necessary.**

## Introduction

Traditionally, most researchers have searched for scholarly information through bibliographic databases which match search keywords against the metadata that describes the content, with journal articles being the most common form of content (Hersh et al., 2006). Examples of commonly used bibliographic databases include PubMed and the *ISI Web of Knowledge*. The metadata description serves as a surrogate for the complete article itself. With the advent of electronic (i.e., digital) versions of articles being available, there has been an increased interest in searching the complete, or "full-text," article itself. Many publishers are beginning to support full-text searching of their online content (e.g., JStor, Springer, Wiley, ACM Digital Library). The Pew Survey for OCLC in 2003 (Online Computer Library Center, 2005) found that the vast majority of people (89%) turn to search engines to initiate their searches for information while few use library Web pages (2%) or online databases (2%). Even academic research scientists prefer search engines over library Web pages for their information searching for research purposes (Hemminger, 2005, 2007) and are increasingly turning to meta-search interfaces such as Google Scholar to perform full-text searches. Several factors have led to the success of full-text tools such as Google Scholar: having a single simple search interface covering all resources (meta-search), the increasing amount of scholarly material available on Web pages or through resources made available to search engines, instant results and the ability to access the final content via a single click, and the utility of full-text searching versus metadata searching. This article is concerned with the latter issue—understanding in more detail how full-text searching compares with metadata-based searching of scholarly literature.

While it is clear that full-text matches of search strings yield more matches than does just searching for matches within the metadata of articles, it is not evident how many

Received November 28, 2006; revised June 14, 2007; accepted June 24, 2007

 ● Published online 30 August 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20708

more matches or previously undiscovered articles are found on average, or how relevant they are. It is often simply assumed that finding additional articles automatically will be of greater value to the searcher; however, as users have discovered when faced with millions of search engine hits to sort through for Web pages, more is not always better. Some authors have argued that the low precision of search engines (i.e., a small number of relevant results compared to the large number of returned results) limits their usefulness (Beall, 2006). While the lowest ranked search results for Web pages are generally of lower relevance, it is not clear whether this is true in the medical and scientific literature domain, where there are fewer overall hits and generally more relevance to searches against fairly precise terms such as gene names.

So, even if full-text searching becomes available for all scholarly literature, how helpful would it be? The aim of this study is to better quantify the number of articles discovered by full-text searching in addition to those discovered by metadata searching, to better qualify what percentage of these articles are useful, and to evaluate the potential effectiveness of several standard information retrieval article features for ranking the usefulness of the articles (e.g., the number of matches of the search string within the article) (Salton, 1968; Tenopir, 1984).

Throughout this article, reference is made to "full-text" or "full-text only" searching versus "metadata" searching. Metadata searching by scientists in the biomedical literature generally means searching via PubMed et al. (2007), where the metadata fields are the title and abstract. Thus, in this article, searching for the schizophrenia gene "*COMT*" means searching for the string "COMT" in the text of the title and abstract, which is a subset of the full-text. Note that the character string *COMT* could occur in the full-text of the article, but as long as it also appeared in the metadata, it would be considered a "metadata" match since it would be discovered via searching just the metadata. A "full-text only" article, on the other hand, does not have the character string present in the metadata fields but does have it present in the full-text of the article. This definition of metadata differs from some contexts where the metadata may include other fields such as manually assigned index terms or categories that are not part of the full-text.

## Background

The particular domain investigated in this study is the biomedical literature used by researchers studying genetics. The literature in this area is undergoing explosive growth, which makes it particularly challenging for researchers to keep track of all the scholarly information relevant to their work (Müller, Kenny, & Sternberg, 2004; Shatkay & Feldman, 2003). Additionally, research articles of interest may occur in many different journals, often outside the researchers' core area of interest, making them difficult to discover (Swanson, 1987, 1990). To investigate this problem, two genetics research laboratories which collaborate with our laboratory were recruited to participate in a study comparing searching for information about genes in their literature via metadata (the current standard practice) versus full-text searching. The first research laboratory was in biology (Vision, 2006), and researchers there studied the genetics of Arabidopsis, a plant in the mustard family commonly used as a genetics model. The second laboratory was in the Neuroscience Department in the School of Medicine, where researchers studied the genetic causes of human schizophrenia (Sullivan, 2006). Researchers in both labs typically used PubMed to search the MEDLINE (2006) database for particular gene names within a set of relevant journals, sometimes qualified by a particular species or disease process within the species. A typical search for Arabidopsis information was just the gene name itself. An example search string is "ERD10." A typical search for the schizophrenia researchers was "schizophrenia genename" within a set of schizophrenia-related journals. An example search string is "schizophrenia COMT." The experimental tasks in this study use the same literature, the same search tasks, and similar evaluations to those commonly utilized by these researchers in their daily practice.

There is extensive previous work in text searching within the biomedical literature community (e.g., Chiang & Yu, 2003; De Bruijn & Martin, 2002; Tanabe et al., 1999), and Hirschman et al. (2002) provided a good review. A significant body of research also has been developed in the general text-retrieval community. Perhaps most well known is the Text Retrieval Conference (TREC), which works to develop common test collections and facilitates the comparison and evaluation of different information retrieval strategies. In 2003, TREC introduced a genomics track with the goal of creating a large test collection to facilitate researchers developing and improving their genomics search systems (Hersh et al. 2006). The TREC Genomics Tracks have been very successful and provide valuable resources and insights. In the TREC Genomics Track, evaluations are performed using the MEDLINE bibliographic metadata because of its availability, although the authors recognize the growing significance of online full-text materials (Hersh et al., 2006). The work described in this article differs from TREC in that it focuses on full-text, and evaluates differences in the quantity and quality of articles that are retrieved when using full-text as compared to metadata searches. This article does not evaluate different algorithms for information retrieval; rather, it performs the same simple text matching so that searching is standardized across the two source types (i.e., full-text and metadata). Relevance judgments, however, are very similar to the TREC 2006 Genomics Track ad hoc retrieval task (Hersh et al., 2006), in using a panel of expert reviewers and being structured around "generic topic templates" that involve finding articles involving a gene and related topics such a as disease, process, or mutation (Hersh et al., 2006). Another related challenge-based workshop was the 2002 Knowledge Discovery and Data Mining Challenge Cup. The evaluation in this case was to rank the usefulness of articles and make a binary decision whether to curate them. The articles and lists of genes present in the article were provided. These efforts


DOI: 10.1002/asi

(Yeh et al., 2003) addressed higher level decision making based on knowledge extracted from the articles, and thus are different that what is addressed in this article.

The most relevant literature is that studying the utility of full-text versus metadata searching. With the advent of computer processing of text, there came the belief that full-text searching would be a significant improvement (Salton, 1970; Swanson, 1960). Some of the initial work did not always find this to be true (Blair & Marion, 1985), as early information retrieval systems did not scale well with larger document sets. A few have even argued that with the overwhelming amount of content available to meta-searches of full-text documents (e.g., GoogleScholar), full-text-based searches cannot provide accurate enough precision to be useful (Beall, 2006). The standard trade-off between precision and recall suggests that full-text searches will discover more documents (higher recall), but with less precision than metadata searches. This was borne out in a comprehensive study of the biomedical literature (McKinin et al., 1991), which analyzed 100 searches performed against a database of several hundred thousand articles in MEDLINE. They found that roughly twice as many relevant articles were discovered by full-text searches as compared to metadata searches; however, the precision of the full-text retrieved articles was statistically significantly less than that of the metadata articles. One possible limitation of the McKinin et al. (1991) study was that judgments of relevance were based on the citation including abstract, and not on the full-text of the article. More recent studies have shown evidence that having the full-text available allows for the possibility of increasing recall and potentially precision. Müller et al. (2004) reported that the availability of full-text is critical for achieving a satisfactory recall rate for researchers working with biological literature. Donaldson et al. (2003) found that their classifier for extracting biological domain knowledge (PreBIND/Textomy) required the use of full-text articles to be successful. Very recently, Hawking and Zobel (2007) tested whether topic metadata improved users' performance when searching university Web pages. They found that the metadata was of little value in ranking answers whereas quantitative measures such as link counts and URL length did improve baseline performance. Perhaps the strongest argument is the increasing use of full-text search engines using relevance rankings (e.g., Google, Google Scholar) by scientists and the corresponding decrease in use of metadata-based citation resources Hemminger (in press).

Other work has evaluated whether abstracts are representative summaries of the full-text, and whether word-occurrence frequencies can indicate relative importance of articles. Tenopir (1984) found a correlation between word occurrences of the search term and relevance, and suggested that it would be useful to establish word-occurrence thresholds associated with levels of relevance. Search-term occurrences, or hits, also have been more recently used in search engine relevance calculations, notably in the original Google description (Brin & Page, 1998). In other work, several studies have found that abstracts were inconsistent with the full-text or that terms occurred in full-text but not in the abstract (Pitkin et al., 1999; Weinberg, 1981). Contrasting this, Ries et al. (2001) compared the frequency of occurrence of index terms in the abstract versus the full-text, and found the abstract to be representative of the full-text in 96% of the 1,138 medical articles they examined.

This work attempts to better understand the utility of full-text searching versus metadata searching, and whether metadata searching as a surrogate for full-text searching is still necessary.

## Methods

Two sets of analyses were performed. The first analysis, referred to as *Article Discovery*, examined the frequency at which scholarly journal articles are discovered via metadata searches versus full-text-only searches, for a large set of scholarly literature in the two domain areas: (a) the human complex disease schizophrenia and (b) the plant Arabidopsis. The second set of analyses, *Article Review*, involved an observer experiment for each of the two domain areas where expert reviewers scored the value (i.e., relevance) of individual articles and classified the context in which the gene was discussed in the article. This allowed correlations to be made between the value of articles discovered and their method of discovery (full-text vs. metadata search) as well as other features of the articles commonly used in information retrieval (e.g., number of occurrences of the search term in the article). The resources used to conduct the two sets of studies are summarized in Table 1.

## Reviewers

Reviewers were experienced researchers in one of two areas: Arabidopsis plant biology genetics or human schizophrenia genetics. One of the reviewers for each of the two studies was the "senior" reviewer. The senior reviewer helped identify the journals most appropriate to the research area, and provided or reviewed lists of genes that the scientific community was actively studying as being related to the research area. The senior reviewer was a faculty member working in that research area. The additional reviewers were postdoctoral and doctoral students working in the research area.

## Journal Articles

Articles used in the studies were selected based on the significance of the journal and the availability of the journal articles in electronic form. First, the senior reviewer identified primary journals for the group's research area. This list was then culled to keep only the journals available in MEDLINE, to which our university library subscribed and for which we could electronically retrieve both the metadata and full-text versions. Articles were collected from 1994 to 2005 (a 12-year period) because before 1994 many articles are not available electronically, or if they were, they were usually in a scanned


DOI: 10.1002/asi

TABLE 1. Summary of information about the different article sets used in the analyses. The left column describes the named groups, or sets, of articles. The second and third columns describe the source of articles (journals), what type of article (whether articles were found by metadata searching, or full-text-only searching), and the counts of articles in each set. The second column describes articles in the Arabidopsis study, and the third column describes those in the schizophrenia study.

| | Arabidopsis | Schizophrenia |
|---|---|---|
| Article Discovery Set | *Plant Cell*<br>*Plant Physiology*<br>*Genes Development*<br>*Journal of Experimental Biology*<br>*PNAS*<br>(13,991 total articles) | *PNAS*<br>*American Journal of Human Genetics*<br>*American Journal of Psychiatry*<br>*Archives of General Psychiatry*<br>(12,314 total articles) |
| Article Review base set.<br>Three major journals selected<br>in research area, covering 1994–2005. | *Plant Cell*<br>*Plant Physiology*<br>*Genes Development* | *American Journal of Psychiatry*<br>*American Journal of Human Genetics*<br>*PNAS* |
| Gene names | Candidates (5,175)<br>Article Review subset (10) | Candidates (26,597)<br>Article Review subset (15) |
| Article Review study set | Metadata articles (18)<br>Full-text articles (82)<br>Total (100) | Metadata articles (19)<br>Full-text articles (83)<br>Total (102) |
| Article Review training set | Metadata articles (3)<br>Full-text articles (17) | Metadata articles (3)<br>Full-text articles (9) |

format and digital text searching could not always be accurately performed. The collection of all articles from this list of journals is referred to as the *Article Discovery Set* (Table 1). From this list, the senior reviewer then selected the top three journals. This set of articles from all three journals, over the 12-year time period, is referred to as the *Article Review Base Set*, with one set for each of the two studies. For the Arabidopsis genetics article review, the journals selected were *Plant Cell*, *Plant Physiology*, and *Genes Development*. The *American Journal of Psychiatry*, the *American Journal of Human Genetics*, and *PNAS* were the journals selected for the schizophrenia study. The Article Review Study Set is the subset of articles selected from the Article Review Base Set used to perform the review study.

## Matching Articles

Searches were intended to model real-life searches by researchers. The reviewers indicated that they use PubMed for most all searches, and often limited their search to a fixed set of known journals and then searched for articles that contained the gene name of interest. In the case of the schizophrenia researchers, they also commonly included "schizophrenia" as a search term in addition to the gene name. To match their standard practice, the search string for the Arabidopsis set was "gene" while the search string for the schizophrenia set was "schizophrenia gene." Thus, for instance, one of the search-term sets for schizophrenia was "schizophrenia COMT." For the Article Discovery analysis, results were computed for both "schizophrenia gene" and "gene." For Arabidopsis, the genes chosen for the review study were randomly selected. Since only a small number of the possible human genes are currently considered relevant to schizophrenia, they were all included (the list being provided by the senior reviewer). Commonly used aliases for the candidate genes were determined and included. Aliases for the schizophrenia genes were determined from the HUGO (2006) resource. The Aradidopsis gene aliases were extracted from the Arabidopsis Information Resource (TAIR, 2006).

Two representations of each article were required for both the Article Discovery and the Article Review experiments: a full-text version and a metadata version. The metadata representation of each article was downloaded directly from the MEDLINE using their efetch interface (NLM, 2006). The full-text version of each article was retrieved directly from the respective journal.

## Experiments

The Article Discovery experiment addressed whether there were differences in the number of articles discovered from searches of the same starting collection of articles, depending on whether the search was done via the normal bibliographic database using matches against the metadata or via a direct search of the full-text of the article. All articles from all journals listed in the Article Discovery Set (see Table 1) were retrieved, in both full-text and as metadata forms. There were 13,991 articles in the Arabidopsis set


DOI: 10.1002/asi

and 12,314 in the schizophrenia set. For each gene in the Gene Names (Table 1), a search was performed against both full-text and metadata representations of each article. Each gene name was searched against each article, and the results were classified into four categories: (a) The gene name was discovered in metadata fields only, (b) it was discovered in the metadata and full-text fields, (c) it was discovered in the full-text only, or (d) it was not present in the article. The total counts in each of these categories were computed for both the schizophrenia and Arabidopsis article sets. Additionally, the number of individual matches of the gene name in each individual article was counted.

For the Article Review experiment, expert reviewers rated articles as to their usefulness. Metadata articles are defined as articles that could be found by searching for the search string in the articles' metadata. Full-text-only articles are the articles that contain the search string only in the full-text of the article, and not in the metadata. To generate the metadata and full-text-only sets, two separate searches were performed against the Article Review Base Set for each gene name. The metadata representations of the articles were retrieved from MEDLINE and stored locally in a MySQL (2006a) database. The metadata record contained the abstract, title, and full-text. The full-text-only representations of the articles were retrieved electronically from the journals in PDF format (Adobe Systems). They were then converted to plain text via PDFbox (2006), which captures all the text including text in tables and figure captions, but not text embedded in images. The plain-text version of the article also was stored in the MySQL (2006a) database. Searches against the metadata or full-text versions of articles in the database were performed using the MySQL database full-text search functions (MySQL, 2006b), the MySQL stop word list (MySQL, 2006c), and matching only the correct capitalization of the gene name. Note that articles in the metadata category also could include the search string in the full-text as well as in the metadata, and most did.

Based on power estimates for resolving differences in mean quality ratings between the full-text-only and metadata sets using three readers, a target of at least 20 articles in both groups was chosen. Additionally, power estimates projected needing 80 articles to analyze feature subsets of the full-text-only set. As a result, the Article Review Study Set was designed to have 20 metadata articles and 80 full-text-only articles, for a total of 100 articles for each of the two domains. Gene candidates were randomly assigned to be added until the total number of articles reached the desired targets. The final schizophrenia study set contained 102 articles, and the Arabidopsis set contained 100. There were a total of 15 schizophrenia gene names and 10 Arabidopsis gene names. Final counts of the articles in the study sets, broken down by full-text-only and metadata categories, are shown in Tables 2 and 3. Appendix A contains the counts of articles in the test set broken down by the individual gene search terms, for both studies. Articles were presented one at a time to reviewers. The order of presentation of the full Article Review set to the reviewers was randomized, and they were blind as to whether the article was from the metadata-only or full-text-only set. Prior to running the test set, users were trained on the "training" sets (Table 1).

TABLE 2. Describes the counts of article by article type (discovered by metadata search or full-text-only search) for schizophrenia review set.

| Schizophrenia | | | |
|---|---|---|---|
| Journal | Metadata | Full-text only | Total |
| *American Journal of Psychiatry* | 12 | 47 | 59 |
| *American Journal of Human Genetics* | 6 | 25 | 31 |
| *PNAS* | 1 | 11 | 12 |
| Total | 19 | 83 | 102 |

TABLE 3. Describes the counts of article by article type (discovered by metadata search or full-text-only search) for Arabidopsis review set.

| Arabidopsis | | | |
|---|---|---|---|
| Journal | Metadata | Full-text only | Total |
| *Plant Physiology* | 10 | 42 | 52 |
| *Plant Cell* | 4 | 38 | 41 |
| *Genes and Development* | 4 | 2 | 6 |
| Total | 18 | 82 | 100 |

## Article Review Presentation

The NeoRef article review system was used to present the articles. A demonstration version of this system using only publicly available articles is available on the Web (NeoRef, 2006). The article review system used in this study is part of a larger system, which is being developed to improve the ways science researchers search and organize scholarly literature. The complete NeoRef system allows researchers to enter a search string similar to the Google Scholar interface, and it returns matches against all available scholarly literature. As part of the literature-review capabilities, the search results can automatically be organized and brought up as consecutive PDFs for the reviewer to step through. The interface supports the observer making qualitative annotations as well as giving quantitative scores, and storing these results in a bibliographic database (e.g., Endnote). In this experiment, only the review portion of the interface was used, as the search results are automatically generated from the defined search strings before they start the review process. The normal full-text PDF formatted presentation of the journal article appears automatically on the screen, with an added right-column interface which provides the reviewer with the ability to score the currently viewed article, as well as other options (e.g., come back later, skip this article). All instances of the gene keyword are highlighted in the PDF, and the viewer comes up with


DOI: 10.1002/asi

the first matched keyword location displayed on the screen. When the user has completed reviewing the article and scoring it, he or she can click on the "next article" button, and immediately the next article comes up.

## Article Review Task

Reviewers were asked to rate journal articles as to their usefulness. The schizophrenia reviewers were instructed "to play the role of a researcher with general experience in this research field (e.g., knowledgeable about genetics and brain disorders), but new to this particular research area (i.e., genetic causes of schizophrenia)." The Arabidopsis reviewers received the same instruction for "the genetics of Arabidopsis." They were asked to "review the article and judge its relevance to them as someone new to the biology of the gene trying to build an understanding of the state of knowledge in that research area." For each article, they were asked to do two things: (a) to score the usefulness of the article quantitatively (Table 4) and (b) to assign one or more terms from a controlled vocabulary that describe features of this article with respect to their study area and genetics (Table 5). The quantitative scoring system (ranking of 1–5) matched what is used in our existing NeoRef system, which was developed based on input from scientists using the system. It is slightly more fine grained, but comparable to the three-level relevance ranking used by TREC Genomics Track (Hersh et al., 2006), and similar to the five-category scale used by McKinin et al. (1991). Reviewers assign the quantitative score as part of the user interface; the assignment of terms was done by verbalizing them to the experimenter who recorded them. The controlled vocabulary was developed by the experimenter and the senior reviewers during pilot testing on approximately 20 cases from their study-subject area.

## Results

### *Analysis Summary*

Several analyses were performed to better understand the utility of searching for keywords in the full-text of an article. First, a quantitative comparison was made of the number of times articles are discovered via full-text-only matches compared to metadata plus full-text matches in the Article Discovery set. Second, for the Article Review Study Set,

TABLE 4. Possible ranking scores assigned by reviewers to each reviewed article. The left column is the actual rating score assigned, the middle column is the label for that quantitative value, and the right column is the meaning used by reviewers (established during the pilot sessions and shared with reviewers during the training sessions).

| Rating | Rating name | Rating usage |
|---|---|---|
| 1 | Definitely Useful | Right on topic, very helpful, primary initial study, excellent review, etc. |
| 2 | Probably Useful | On topic and potentially important material. |
| 3 | Possibly Useful | Has some material or references that are likely useful, but not certain without further checking. |
| 4 | Probably Not Useful | Unlikely, but may have some use, for instance, references to check out. |
| 5 | Definitely Not Useful | Not on topic; nothing of direct value, not worth keeping. |

TABLE 5. Controlled vocabulary used in assigning terms to the article.

| Code | Definition |
|---|---|
| SPDG | Same process/pathway/phentotype/disease different gene. |
| SGDO | Same gene different organism (e.g., mouse). |
| SGDD | Same gene different disease/phenotype. |
| DGDD | Different gene different disease/phentotype. |
| MUTANT | Characterization of a mutant of the gene of interest; includes molecular biology of variant function, transgenic models. |
| FAMILY | Not the gene of interest, but a closely related gene. |
| SEQUENCE | Sequence analysis including this gene. |
| INTERACTION | Identification of interaction (gene characterized in paper interacts with gene of interest, or vice versa). At DNA, RNA, or protein level. |
| PROCESS | Identification of something in the process (biochemical effect, metabolic effect). |
| STRUCTURE | Identification of something in the structure (biochemical effect, metabolic effect). |
| UP | Upstream, includes anything upstream of it (i.e., not only promoter). |
| DOWN | Downstream, mutant of gene of interest affects gene in paper. |
| REVIEW | Review, summary, overview types of articles. |
| MARKER | Use of gene as physiological/developmental marker; not for characterization of gene(already known) but just for use as marker. |
| FP | False positive from literature search. Matched the gene of interest in name, but it was actually something different from the gene (e.g., matched LACS2 when looking for ACS2, or matched e-mail address which contained the gene name). |
| REFERENCE | The only matches occurred in the references (citations). |
| TABLE | The matches are only in table or figures. |
| MIP | Mentioned in passing. The main discussion of the paper is not about this gene, it is just mentioned in passing. |
| IMG | Imaging Study. The method of observation is an imaging study (as opposed to genetic analysis, or organism's visible phenotype). |


DOI: 10.1002/asi

three comparisons were made: (a) comparing differences between the mean rating values of the full-text-only articles and the metadata articles, (b) studying correlations between the observer's coding of article features and the article's mean rating value, and (c) the correlations between the automatically calculated features of the article and its rating score. Finally, the reviewers provided qualitative feedback on the review-system interaction as a tool for literature searching.

### *Article Discovery*

This analysis was to determine the number of articles in the Article Discovery set that would have been discovered by a full-text-only search versus a metadata search. The results are summarized across all genes in each domain (Table 6) because the large number of genes in each category precludes listing the results by individual gene. For the Arabidopsis genes, only about one quarter more articles were discovered through full-text-only search ($n$ = 5,705) compared to that through metadata search ($n$ = 22,305). For the schizophrenia set, about 10 times as many full-text-only searched articles ($n$ = 1,671) were discovered compared to metadata-discovered ones ($n$ = 161). Overall, a schizophrenia researcher using current searching practice (PubMed search of metadata for "schizophrenia COMT") would find only one tenth of the total articles in PubMed that actually included the search terms somewhere in the article. The number of metadata-only matches differed between the two literature cohorts as well. For schizophrenia, only a very small number of the articles were discovered through metadata-only matches, and thus, tools that perform a full-text search on the articles without using any metadata would discover most all the articles currently discovered via metadata searches, as well as discovering all the other articles previously undiscovered. On the other hand, in Arabidopsis, the number of metadata-only discovered cases is almost half as many as the full-text-only discovered cases, and not discovering these articles would mean a substantial number were not retrieved. For comparison purposes, the same calculations for the Schizophrenia study set were computed for just "gene" without "schizophrenia" as a search term as well, and the results show a decrease in the percentage of full-text-only articles (from 83–59%) and a corresponding equal increase in metadata-only and metadata-discovered articles. Overall, the percentage of articles found only by full-text searches thus depends on both the corpus as well as the search terms.

### *Article Review: Quality Differences*

Given that full-text searching could allow for the discovery of many previously undiscovered articles, how does the quality of the additionally discovered articles compare? To see if there are differences in usefulness to the researcher searching for information, the reviewers' mean ratings of articles in the two categories (metadata and full-text-only discovered articles) were compared in the Article Review study. The results are shown for both the Schizophrenia and Arabidopsis sets in Tables 7 and 8, respectively. A test of mean differences using analysis of variance adjusted for multiple observations within readers was used (SAS GENMOD, SAS Institute Inc. SAS/SAT, Cary, NC) to compare the differences between the mean quality value for an observer's full-text discovered articles versus the mean quality value for their metadata discovered articles. In both cases and for all observers, their mean quality rating values were lower (i.e., more useful) for the metadata-discovered articles. There were statistically significant differences between the mean quality rating for the metadata-discovered articles versus the full-text-discovered articles for the Arabidopsis set at the $p < .05$ level ($M_{\text{diff}} = -1.35$, $df = 1$, $p < .0001$) and the schizophrenia set at the $p < .05$ level ($M_{\text{diff}} = -1.26$, $df = 1$, $p < .0001$).

Thus, the implication is that the value of the articles discovered by metadata is on average more useful than those discovered by full-text. This may be due to certain subsets of the full-text-only discovered articles being of lower value to researchers, for instance, when the match occurs only in the references and not the text of the article. This is investigated later in the analysis studying correlations of mean rankings with article features.

Agreement between pairs of observer's ratings over all the review articles was calculated using a weighted $\kappa$ statistic. There was moderate (Landis & Koch, 1977) pairwise agreement between all observers in the Arabidopsis experiment (0.43, 0.48, 0.45), and between two of the observers in the schizophrenia experiment (0.56); however, the remaining schizophrenia observer, who was the least experienced, had only fair agreement with the other two observers (0.25, 0.22).

TABLE 6. Counts and percentages of how many times each of the schizophrenia genes and the Arabidopsis genes were discovered in their literature cohort. The percentages given are just calculated across the number of articles that had a match since in the vast majority of cases (99.9%), a given gene did not match a given article. Separate totals are given for the genes found in articles, and for the complete set.

| | Schizophrenia + schizophrenia gene | | Schizophrenia gene | | Arabidopsis gene | |
|---|---|---|---|---|---|---|
| Genes found in metadata only | 172 | 8.58% | 3,541 | 20.63% | 2,712 | 8.83% |
| Genes found in full-text only | 1,671 | 83.38% | 10,125 | 58.99% | 5,705 | 18.57% |
| Genes found in metadata and full-text | 161 | 8.03% | 3,498 | 20.38% | 22,305 | 72.60% |
| Totals for found genes | 2,004 | | 17,164 | | 30,722 | |
| Genes not found | 327,513,454 | | 327,498,294 | | 72,372,703 | |
| Overall total | 327,515,458 | | 327,515,458 | | 72,403,425 | |


DOI: 10.1002/asi

TABLE 7. Mean rating values for each of the observers (A, B, C) of the Schizophrenia study, including mean ratings broken down by full-text-only discovered articles versus metadata-discovered articles, as well as the difference between these two. The right column is the average of the mean ratings across all three observers.

| | Schizophrenia | | | |
|---|---|---|---|---|
| Observer | A | B | C | *M* |
| *M* ratings | 3.29 | 2.51 | 3.05 | 2.95 |
| *M* ratings (full-text) | 3.58 | 2.71 | 3.27 | 3.19 |
| *M* ratings (metadata) | 2.05 | 1.63 | 2.11 | 1.93 |
| Difference in *M* rating (full-text − metadata) | 1.53 | 1.08 | 1.16 | 1.26 |

TABLE 8. Mean rating values for each of the observers (D, E, F) of the Arabidopsis study, including mean ratings broken down by full-text-only discovered articles versus metadata-discovered articles, as well as the difference between these two. The right column is the average of the mean ratings across all three observers.

| | Arabidopsis | | | |
|---|---|---|---|---|
| Observer | D | E | F | *M* |
| *M* ratings | 3.09 | 2.83 | 2.85 | 2.92 |
| *M* ratings (full-text) | 3.43 | 3.00 | 3.07 | 3.17 |
| *M* ratings (metadata) | 1.56 | 2.06 | 1.83 | 1.82 |
| Difference in *M* rating (full-text − metadata) | 1.87 | 0.94 | 1.24 | 1.35 |

This suggests that a single experienced reviewer's article rating may serve as a reasonable predictor of usefulness for most readers.

### *Precision and Recall Measures*

To understand these results in relationship to prior work, it is helpful to cast these results in terms of precision and recall measures. As with many large corpora, it is not practical to exactly determine by hand the relevance of many thousands of full-text articles. Calculating relevance in such situations is usually based on a smaller sample "pool," generally the top items returned (TREC uses the top 100.), which generally has been determined to not be statistically significantly different from the entire corpus (Spark & van Rijsbergen, 1975; Voorhees & Harman, 1996). Since relevance judgments for the review experiment are available for all the articles containing a match of the search query, recall and precision are calculated using the "usefulness" rating score as relevance judgments. Ratings of 1, 2, and 3 (Definitely Useful, Probably Useful, and Possibly Useful, respectively) are considered relevant, and ratings of 4 and 5 (Probably not Useful, Definitely not Useful, respectively) are considered not relevant. Ratings are averaged across all reviewers for a mean rating score, and mean ratings ranging from 1 to 3.5 are considered as relevant while mean ratings ranging from 3.5 to 5 are considered not relevant. Using these definitions, the precision (i.e., number of relevant items returned ÷ the number of all relevant items) and the recall (i.e., number of relevant items returned ÷ the number of items returned) were calculated and are shown in Table 9. The recall calculations were corrected because the sample is not representative (see Appendix B for details). The results show the typical trade-off between precision and recall, with the full-text-discovered cases having higher recall but poorer precision as compared with the metadata-discovered cases.

### *Article Review: Feature Analysis*

When the reviewer searches on dopamine receptor and schizophrenia within just the three schizophrenia journals used for the article study, the user would be returned 284 matching articles. When presented with such large numbers of results, the user tends to review just the first ones presented. This has been well documented with search engine results, where users rarely consider results beyond the first page (Jansen et al., 1998). While scholars may be more inclined to work through more of the search results, it would be important to be able to rank the returned articles based on likely utility to the searcher, such that the most useful articles could be displayed first. Features that might correlate well with usefulness of the article could be used to rank or filter the resulting hits so that only the most useful (i.e., relevant) results were displayed or retained. Being able to automatically recognize such features that correlate with relevant articles could improve the precision of full-text searching.

When the observers' mean ratings were compared for the full-text-only and metadata-discovered articles, the metadata articles were rated as more useful than were the full-text ones. To study this further in the context of article features, mean ratings were computed for each of the reviewer-coded and automatically recognized article features. Analyzing the relationship in more detail between reviewer-identified features and the reviewers' mean ratings value will be the focus of another publication. Such information also may lead to improvements in automatic classifiers by documenting the significance of the scientists' manually identified features. A count of how many articles were classified by at least one of the reviewers as being of a particular feature type is shown in the first part of Table 10. Note that articles can be classified by more than one term, so the percentages add up to more than 100%. In addition to the classification terms assigned to articles by reviewers, automatic classification of articles was made into one of four types (i.e., Standard Text,

TABLE 9. Results from the Review study calculated in terms of precision and recall. Original calculations of recall, before correcting for the sample versus population distribution bias, are in parentheses.

| | Schizophrenia | | Arabidopsis | |
|---|---|---|---|---|
| | Recall | Precision | Recall | Precision |
| Metadata discovered | 15.7% (16.6%) | 94.7% | 84.1% (84.1%) | 100% |
| Full-text-only discovered | 100% | 63.7% | 100% | 69% |


DOI: 10.1002/asi

TABLE 10. Counts, percentages, and mean ratings for each of the reviewer and automated article classification codes, for both the Schizophrenia and Arabidopsis review sets.

| | Schizophrenia | | | Arabidopsis | | |
|---|---|---|---|---|---|---|
| Search term | No. of matches | Percentage of articles matched | Mean reviewer rating for article class | No. of matches | Percentage of articles matched | Mean reviewer rating for article class |
| SPDG | 8 | 7.84 | 3.04 | 39 | 39 | 3.13 |
| SGDO | 2 | 1.96 | 1.67 | 20 | 20.00 | 2.27 |
| SGDD | 25 | 24.51 | 3.39 | 0 | 0.00 | 0.00 |
| DGDD | 10 | 9.80 | 3.90 | 0 | 0.00 | 0.00 |
| MUTANT | 11 | 10.78 | 1.55 | 21 | 21.00 | 2.05 |
| FAMILY | 2 | 1.96 | 2.00 | 42 | 42.00 | 2.71 |
| SEQUENCE | 26 | 25.49 | 1.94 | 17 | 17.00 | 1.92 |
| INTERACTION | 4 | 3.92 | 3.33 | 25 | 25.00 | 2.23 |
| PROCESS | 28 | 27.45 | 2.46 | 44 | 44.00 | 2.32 |
| STRUCTURE | 7 | 6.86 | 2.10 | 7 | 7.00 | 1.76 |
| UP | 0 | 0.00 | 0.00 | 26 | 26.00 | 2.56 |
| DOWN | 0 | 0.00 | 0.00 | 2 | 2.00 | 2.33 |
| REVIEW | 18 | 17.65 | 2.59 | 10 | 10.00 | 2.63 |
| MARKER | 1 | 0.98 | 4.00 | 17 | 17.00 | 3.31 |
| FP | 2 | 1.96 | 4.00 | 5 | 5.00 | 3.87 |
| REFERENCE | 36 | 35.29 | 3.22 | 13 | 13.00 | 4.03 |
| TABLE | 3 | 2.94 | 3.44 | 8 | 8.00 | 3.29 |
| MIP | 38 | 37.25 | 3.46 | 36 | 36.00 | 3.55 |
| IMG | 15 | 14.71 | 3.38 | 1 | 1.00 | 1.33 |
| Text | 67 | 0.66 | 2.80 | 90 | 0.90 | 2.79 |
| References only | 33 | 0.32 | 3.31 | 9 | 0.09 | 4.19 |
| Letter | 3 | 0.03 | 3.22 | 1 | 0.01 | 4.00 |
| Errata | 1 | 0.01 | 2.33 | 0 | 0.00 | 0.00 |

Letter, Errata, ReferencesOnly). Standard Text is a normal journal article. Letter refers to letters to the editor, and Errata are published corrections to standard articles. ReferencesOnly is the same as the reviewer coded "REF" classification from Table 5 and corresponded to articles that had matches of the search term only in the references (citations) section, and none in the standard-text section. These automatically recognized classification counts are included at the end of Table 10. The final automatic classification performed was counting the number of matches of the search term within each returned article.

Certain reviewer classifications of articles (e.g., MUTANT, SGDO, SEQUENCE) were rated as more useful by the reviewers (i.e., numerically lower average rating scores). Others such as REF, MARKER, FP, TABLE, MIP, and IMG were rated as less useful than average (i.e., numerically higher rating scores). Of the automatically recognized classifications, Errata, Letter, and ReferencesOnly had lower than average usefulness. Excluding articles of these three types from the Article Review Study Set, and rerunning the ANOVA analysis comparing full-text-only (without these articles) versus metadata-discovered articles shows that the metadata articles were still statistically significantly rated as more useful (identical conditions as earlier, with $p$s still $< .0001$ for both the Schizophrenia and Arabidopsis sets).

The number of hits or matches of the search term within the returned document is a commonly used feature to rank returned articles. To test the value of this feature, the number of hits was correlated with the mean quality ranking for each article (averaged across all observers). The results clearly show a relationship where articles with many matches of the search term tend to be much more highly valued. To determine if there were thresholds in number of matches that correspond to statistically significant differences in reviewer quality ratings, a Tukey grouping analysis was performed. Based on the frequency distribution of the number of hits per article, the frequency range of hits was broken into groupings that minimized the differences between members of a group and maximized differences between members of different groups. Tables 11 and 12 show the Tukey analysis for the two article sets.

For schizophrenia, the threshold was 20 hits per article. Articles with 20 hits or more had a statistically significant lower mean rating value (i.e., higher usefulness) averaged across observers, as compared to articles with less than 20 hits. Articles with 5 to 19 hits had lower mean rating values than did articles with one to four hits, but these two groups were not statistically significantly different (Table 11). Similarly for the Arabidopsis article set, articles with 15 hits or more had statistically significant lower mean rating value (i.e., higher usefulness) averaged across observers than did articles with less than 15 hits. Articles with 5 to 14 hits had a lower mean rating value (i.e., higher usefulness) than did articles with 1 to 4 hits; however, the Tukey analysis did not find a statistically significant difference between the two groups (Table 12).


DOI: 10.1002/asi

TABLE 11. Tukey analysis of the mean rating values by the number of matches (i.e., hits) of the search term in the full-text of the article. The Tukey analysis showed three distinct groupings: A, B, and C.

| Schizophrenia gene | | | |
|---|---|---|---|
| Group | Range | *M* rating value | Different from groups |
| A | 1–4 hits | 3.24 | C |
| B | 5–19 hits | 2.88 | C |
| C | > 20 hits | 1.62 | A, B |

TABLE 12. Tukey analysis of the mean rating values by the number of matches (i.e., hits) of the search term in the full-text of the article. The Tukey analysis showed three distinct groupings: A, B, and C.

| Arabidopsis | | | |
|---|---|---|---|
| Group | Range | *M* rating value | Different from groups |
| A | 1–4 hits | 3.41 | C |
| B | 5–14 hits | 2.94 | C |
| C | >15 hits | 1.69 | A, B |

These results are similar to those of Tenopir (1984), who found a significant threshold at 10 hits and a clear relationship between the number of hits and precision. These results also are very similar to what is found in Web searching, where link counts are now commonly utilized to improve the relevance of search engine results. The question, then, is whether the features discovered in this study, such as the number of hits in full-text, can be used to improve the precision of the full-text results so that full-text searches have the same high precision as do metadata searches, but still find more relevant articles. Repeating the calculations on the schizophrenia and Arabidopsis Article Review sets, but limited to only matches with high hit counts (Schizophrenia $\geqslant$ 20 hits and Arabidopsis $\geqslant$ 15 hits) shows that precision for the full-text is now the same (100% in Aradidopsis) or slightly better than that of the metadata retrieved articles (95 vs. 94.4% in schizophrenia); however, the number of additional cases discovered by full-text searching is now only slightly better, finding 5% more cases in schizophrenia and 28% more in Arabidopsis. This suggests that (a) full-text searching can perform as well as or better than metadata searching in precision and recall and that (b) the best solution might be to provide a dynamic interface allowing the user to trade off between precision and recall by controlling the threshold of the number of hits by which the results are filtered.

### *Article Review Interface Evaluation*

At the end of the experiment, the reviewers were asked about the pros and cons of the Article Review interface, and whether they would use it if available. All participants found it useful and said that they would use it. Five of the 6 thought such an interface would be very valuable to them in their research literature searching, in particular, to review and code articles quickly and to be able to directly dump their structured answers into their bibliographic databases for later querying and searching. Two drawbacks of the system were noted. First, the matching text was not always correctly highlighted (i.e., sometimes, it would be off by one or two words due to a technical problem with API interface to Adobe Reader viewer). Second, the highlighting of the search terms was useful, but it also would be very helpful to have the ability to move directly to the next highlighted term through a single interaction such as a keystroke. Such an interaction was not available directly as part of the Article Review system; however, it could be approximated by using the Adobe Reader text find function, which produces a sidebar list of the places in the text where there are matches, allowing the user to jump directly to those locations by clicking on the list entry. All users made frequent use of this feature to find the searched gene names as well as to search for related terms and citations within the text.

## Discussion

In the two biomedical literature cohorts investigated in this article, the standard methods of searching would find between one quarter more (Arabidopsis) and 10 times more (schizophrenia) articles containing the gene name search term. On the other hand, the relevance of the metadata-discovered articles was higher than that of the full-text-only discovered articles. This is in general agreement with the more recent and larger scale studies of similar literatures. It parallels the recent Hawking and Zobel (2007) results, which found metadata not to be of help while quantitative features could be used to improve search results. Of importance in this study was that specific quantitative features of the articles could be used to improve the precision in the full-text retrieval system, similar to how link counts are used in Web searching. A good candidate was the number of hits within each article, for which there were significant differences in mean user ratings when certain thresholds were crossed. Using this information to filter the full-text-only results to keep the higher rated articles (i.e., number of hits above certain thresholds) allows the full-text-only retrieval to perform better in recall, and as good as or better in precision than articles resulting from the metadata search.

This suggests that rather than accepting metadata searching as a surrogate for full-text searching, it may be time to make the transition to direct full-text searching as the standard. This could be accomplished by using certain features of the full-text article, such as number of hits of the search string or whether the search string is found in the metadata (i.e., our current metadata search) as filters that allow us to increase the precision of our results. The combination of faster computer processing and the increased availability of full-text articles allows us to work with the full-text, and to dynamically filter based on features of the full-text. The result is that the reviewer can have immediate access to every article containing his or her search term, or can


DOI: 10.1002/asi

initially present the most likely to be relevant articles and extend the search through the remaining less likely to be relevant articles as desired. If these results hold for searching in other domains, then it suggests that researchers could benefit from being able to perform full-text searches of all available literature using a meta-search tool. They would find significantly more literature, and by article features, they could rank the articles so that they could review articles most likely to be of interest first.

One limitation of the study is that searching by gene name may not be representative of general biomedical literature searches. Gene names are specific terms for concepts and, as a result, can be searched for more easily in full-text compared to other concepts in the biomedical literature. Concepts that are referred to by many different terms are why controlled vocabularies such as MeSH (2007) are so helpful in the biomedical domain. In addition, gene names may occur more frequently in the full-text compared to the metadata, as compared to other types of terms. Further, the searching conducted in this experiment is simplified in that only exact matching is performed using basic search functions (MySQL full-text search). Available systems, especially in the biomedical area (e.g., MEDLINE) commonly support expansion of search terms, such as to include aliases, and use more advanced search tools (e.g., Lucene, 2007). A final limitation is that the metadata in this study is a strict subset of the full-text, and so these results may differ from domains where the metadata contains additional information, for instance, manually curated indexing. Thus, it would be desirable to extend this work by conducting similar analyses in other domains, with other types of metadata, with different types of terms, and with more advanced searching capabilities.

## Acknowledgments

Literature used for the study was available through the National Library of Medicine's MEDLINE database. This work was funded in part by NIH Grant P20 RR020751–01.

## References

Beall, J. (2006). The death of full-text searching. PNLA Quarterly, 70(2), Winter, 1984–1998.

Blair, D.C., & Marion, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communication of the Association for Computing Machinery, 28(3), 289–299.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. WWW7/Computer Networks Conference, April 1998, 107–117, Brisbane, Australia.

Chiang, J.H., & Yu, H.C. (2003). MeKE: Discovering the functions of gene products from biomedical literature via sentence alignment. Bioinformatics, 19(11), 1417–1422.

de Bruijn, B., & Martin, J. (2002). Getting to the (c)ore of knowledge: Mining biomedical literature. International Journal of Medical Informatics, 67(1–3), 7–18.

Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., et al. (2003, March 27). PreBIND and Textomy—Mining the biomedical literature for protein–protein interactions using a support vector machine. BMC Bioinformatics, 4, 11.

Hawking, D., & Zobel, J. (2007). Does topic metadata help with Web search? Journal of the American Society for Information Science and Technology, 58(5), 613–628.

Hemminger, B.M. (2005, November). Information seeking behavior of scientists. Panel presentation at the American Society of Information Science and Technology Conference, Charlotte, NC.

Hemminger, B.M. (2007). Information seeking behavior of academic scientists. Journal of the American Society for Information Science and Technology (in press).

Hersh, W.R., Bhupatiraju, R.T., Ross, L., Roberts, P., Cohen, A.M., & Karemer, D.F. (2006). Enhancing access to the Bibliome: The TREC 2004 Genomics Track. Journal of Biomedical Discovery and Collaboration, 1(3), DOI: 10.1186/1747–5333–1–3.

Hirschman, L., Park, J.C., Tsujii, J., Wong, L., & Wu, C.H. (2002). Accomplishments and challenges in literature data mining for biology. Bioinformatics, 18(12), 1553–1561.

HUGO. (2006). [database]. Retrieved October 10, 2006, from http://www.genenames.org/index.html.

Jansen, B.J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the Web. ACM SIGIR Forum, 32(1), 5–17.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159–174.

Lucene. (2007). Lucene full-text search engine application. Retrieved June 14, 2007, from http://lucene.apache.org/java/docs/

McKinin, E.J., Sievert, M., Johnson, E.D., & Mitchell, J.A. (1991). The Medline/Full-Text Research Project. Journal of the American Society for Information Science, 42(4), 297–307.

MeSH. (2007). National Library of Medicine's Medical Subject Headings (MeSH). Retrieved June 14, 2007, from http://www.nlm.nih.gov/mesh/

Müller, H.M., Kenny, E.E., & Sternberg, P.W. (2004). Textpresso: An ontology-based information retrieval and extraction system for biological literature. PLoS Biology, 2(11).

MySQL. (2006a). MySQL Database System. Retrieved October 10, 2006, from http://dev.mysql.com

MySQL. (2006b). MySQL full-text search functions. Retrieved October 10, 2006, from http://dev.mysql.com/doc/refman/5.0/en/fulltext-search.html

MySQL. (2006c). MySQL stopword list. Retrieved October 10, 2006, from http://dev.mysql.com/doc/refman/5.0/en/fulltext-stopwords.html

NeoRef. (2006). NeoRef demonstration article review system. Retrieved October 10, 2006, from http://neoref.ils.unc.edu/review/

NLM. (2006). Efetch routine. Retrieved July 8, 2006, from http://eutils.ncbi.nlm.nih.gov/entrez/query/static/efetch_help.html

Online Computer Library Center. (2005). Perceptions of libraries and information resources (2005). PEW Survey for Online Computer Library Center. Retrieved October 10, 2006, from http://www.oclc.org/reports/2005perceptions.htm

PDFbox. (2006). PDFbox application interface. Retrieved October 10, 2006, from http://www.pdfbox.org/

Pitkin, R.M, Branagan, M.A., & Burmeister, L.F. (1999). Accuracy of data in abstracts of published research articles. Journal of the American Medical Association, 281(12), 1110–1111.

PubMed. (2006). National Library of Medicine PubMed Database. Retrieved October 10, 2006, from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed

Ries, J.E., Su, K., Peterson, G., Sievert, M.E.C., Patrick, T.B., Moxley, D.E., & Ries, L.D. (2001). Comparing frequency of content-bearing words in abstracts and texts in articles from four medical journals: An exploratory study. Medinfo, 10(Pt. 1), 381–384.

Salton, G. (1968). Automatic information organization and retrieval. McGraw-Hill.

Salton, G. (1970). Automatic text analysis. Science, 168(3929), 335–343.

Shatkay, H., & Feldman, R. (2003). Mining the biomedical literature in the Genomic Era: An overview. Journal of Computational Biology, 10(6), 821–855.

Spark, K.J., & van Rijsbergen, C. (1975). Report on the need for and provision of an "ideal" information retrieval test collection (British Library

Research and Development Report No. 5266). University of Cambridge, England, Computer Laboratory.
Sullivan. (2006). Pat Sullivan Research Laboratory. Retrieved October 10, 2006, from http://www.med.unc.edu/~pfsulliv/welcome.htm
Swanson, D.R. (1960). Searching natural language text by computer. Science, 132(3434), 1099–1104.
Swanson, D.R. (1987). Two medical literatures that are logically but not bibliographically connected. Journal of the American Society for Information Science, 38, 228–233.
Swanson, D.R. (1990). Medical literature as a potential source of new knowledge. Bulletin of the Medical Library Association, 78(1), 29–37.
TAIR. (2006). TAIR Database Resource. Retrieved October 10, 2006, from http://www.arabidopsis.org/
Tanabe, L., Scherf, U., Smith L.H., Lee, J.K., Hunter, L., & Weinstein, J.N. (1999). MedMiner: An Internet text-mining tool for biomedical information, with application to gene expression profiling. BioTechniques, 27, 1210–1217.
Tenopir, C. (1984). Retrieval performance in a full-text journal article database. Unpublished doctoral dissertation, University of Illinois School of Library and Information Science, Champaign-Urbana.
Vision. (2006). Todd Vision Research Laboratory. Retrieved October 10, 2006, from http://visionlab.bio.unc.edu/
Voorhees, E., & Harman, D. (2000). Overview of the 9th Text Retrieval Conference.
Weinberg, B.H. (1981). Word frequency and automatic indexing (dissertation). New York: Columbia University, School of Library Service. Ann Arbor, MI: University Microfilms, 1983.
Yeh, A.S., Hirschman, L., & Morgan, A.A. (2003). Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup Bioinformatics, 19(Suppl. 1), i331–i339.

## Appendix A

| | Full-text only | Metadata |
|---|---|---|
| **Schizophrenia search terms** | | |
| dopamine receptor | 24 | 1 |
| AKT1 | 1 | |
| DISC1 | 2 | |
| DRD3 | 4 | |
| G30 | 2 | |
| NRG1 | 4 | |
| PRODH | 1 | |
| RGS4 | 4 | |
| ZDHHC8 | 1 | |
| G72 | 2 | |
| HTR2 | 1 | |
| COMT | 16 | |
| DTNBP1 | 14 | 15 |
| Neuregulin | 4 | 3 |
| Catechol-o-methyltransferase | 3 | |
| Total | 83 | 19 |
| **Arabidopsis search terms** | | |
| FIS | 7 | 2 |
| ERD10 | 8 | |
| BES1 | 8 | 2 |
| DCL1 | 6 | 4 |
| CER1 | 10 | 4 |
| RAB11 | 9 | 1 |
| CCR1 | 8 | 2 |
| LHCB6 | 9 | 1 |
| ACS2 | 9 | 1 |
| HFR1 | 8 | 1 |
| Total | 82 | 18 |

## Appendix B

Calculating recall just on the review set sample would be biased because the percentage of metadata versus full-text cases is not the same as the overall population. Since the Article Discovery experiment provides a good estimate of this ratio, recall for the full population can be estimated from the results of the Article Discovery experiment by correcting them using the results from the Article Review experiment to take into account the actual percentage of relevant articles as judged by the observers. From Table 6, the percentage of metadata-discovered cases (including metadata-only cases) of all matched cases in the Article Discovery study is 16.6% for schizophrenia and 81.4% for Arabidopsis. For Arabidopsis, 100% of the metadata matched items in the Review study are relevant, so recall would be estimated to be the same as that found in the Discovery study (81.4%). Eighteen of the 19 metadata cases in the Schizophrenia review study were relevant (94.7%), so estimated recall for metadata cases would be adjusted slightly downward from 16.6 to 15.7%.


DOI: 10.1002/asi