



SEA CHANGE IN E-DISCOVERY SEARCH:

Defensibility and the Emergence
of Transparent Search

A Clearwell White Paper

Table of Contents

Introduction	3
Challenges with E-Discovery Search.....	3
Issues with Today’s Search Technology.....	5
Black Box Search.....	6
The Need for Transparency.....	7
Introducing Transparent Search.....	8
Benefits of Transparent Search.....	8

INTRODUCTION

As the amount of electronically stored information (ESI) has grown, keyword search has become increasingly important in e-discovery as a way to uncover relevant information and cull-down large amounts of ESI prior to review. Without keyword search, e-discovery would be virtually impractical due to its duration and high cost. Nevertheless, while vital to e-discovery, keyword search is not perfect. It can produce results that are both over and under-inclusive, finding non-relevant documents and missing potentially relevant documents.

As the use of keyword search has increased, the way it is being used in e-discovery is receiving more scrutiny from the courts. Recent opinions, such as Judge Grimm's in *Victor Stanley v Creative Pipe, Inc* 2008 WL 2221841 (D. Md. May 29, 2008), highlight some of the limitations of keyword search and outline approaches that attorneys can employ to defensibly use keyword search given these limitations. Addressing the limitations of keyword search while following these approaches in practice, however, is challenging. The technology that is largely used today to perform keyword search in e-discovery was originally designed for other purposes and works as a "black box." This black box design is a major reason why using keyword search makes it difficult to follow the defensible practices outlined in case law. A new, more transparent approach to search is needed to remove the black box and address the shortcomings of today's search technology. This new transparent search technology enables attorneys and litigation support professionals to defensibly utilize search best practices. Transparent search also reduces the over and under-inclusiveness of keyword search, improving search effectiveness and saving cost and time.

CHALLENGES WITH E-DISCOVERY SEARCH

The primary method employed in e-discovery to find relevant information and cull down large amounts of ESI is keyword search. However, keyword search is not a perfect solution. A keyword search can miss potentially relevant documents (i.e., be under-inclusive), or find non-relevant documents (i.e., be over-inclusive). The "Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery" summarizes many of these issues with keywords as applied to e-discovery. Reasons why keyword searches can be under-inclusive include:

- **Unknown keywords.** Not selecting the right words for which to search.
- **Word variations.** Not including common word variations based on conjugation or tense e.g. searching for "hiring," but not including "hire," "hired," and "hires" in your search.
- **Similar keywords.** Missing documents containing words with a similar meaning as the ones included in your search, but which have not been specifically included in the search e.g. searching for "hiring" but not finding documents where an author has used "employ" or "make an offer to" instead of "hiring"
- **Misspellings.** Not including common misspellings of the specified keywords such as "hier" instead of "hire."

The primary reason keyword searches can be over-inclusive is that words have different meanings depending on the context in which they are used. The classic example is the word “strike” which “could be found in documents relating to a labor union tactic, a military action, options trading, or baseball.”¹ This issue is one of the principal reasons why Boolean and proximity search technologies are commonly used. If you found too many documents in your search for “strike,” you might refine your search by searching for “strike AND price” or “strike NOT baseball.”

In a series of recent rulings, including *United States v O’Keefe, Equity Analytics, LLC v Lundin and Victor Stanley v Creative Pipe Inc.*, the bench has shown increasing awareness of these limitations of keyword search. In *Victor Stanley*, Judge Grimm waived attorney-client privilege and work product protection because the defendants failed to demonstrate the search methodology used to prevent the inadvertent production of privileged documents was reasonable. Judge Grimm pointed out that the defendants had the burden to demonstrate their search methodology was reasonable because of the “well-known limitations and risks associated with [keyword searches].” Judge Grimm went on to suggest that, given these known limitations, there are two current approaches parties can follow in order to use keyword search in a defensible manner.

- **Collaboration approach.** The first approach parties could take would be to “confer with their opposing party in an effort to identify a mutually agreeable search and retrieval method.” Grimm points out this approach would “[minimize] cost because if the method is approved, there will be no dispute resolving its sufficiency.”² If this type of collaboration is not possible, then parties can follow a second approach.
- **Best practices approach.** With this approach, Grimm argues that best practices and appropriate search technologies can be used in order to create a reasonable and defensible methodology in the absence of collaboration.

Grimm goes on to specifically cite the Sedona Conference Best Practices document as a source of best practices. “In this regard, compliance with the Sedona Conference Best Practices for use of search and information retrieval will go a long way towards convincing the court that the method chosen was reasonable and reliable.”

The Sedona Conference Best Practices Commentary on Search includes a section on “practical advice” which contains eight “Practice Points.” Practice point two is the most relevant to the actual implementation of a search methodology. It states “Success in using any automated search method or technology will be enhanced by a well-thought out process with substantial human input on the front end.” The Commentary doesn’t proscribe a specific “process” parties should follow, but it does suggest the following key components could be used in an effective search methodology.

- **Testing.** Searches need to be tested for efficacy, i.e. whether the search is producing over or under-inclusive results.
- **Sampling.** The primary way to test the efficacy of a search is through sampling. In *Victor Stanley*, Judge Grimm states that “The only prudent way to test the reliability of the keyword search is to perform some appropriate sampling of the documents determined to be privileged and those determined not to be in order to arrive at a comfort level that the categories are neither over-inclusive nor under-inclusive.”

¹ *The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery*, 8 Sedona Conf. J. (2007) at 201.

² Grimm also notes that “Additionally, cost can be minimized by entering into a court-approved agreement that would comply with Hopson, or if enacted, Proposed Evidence Rule 502.”

The Sedona Conference suggests the following key components could be used in an effective search methodology:

*Testing
Sampling
Iterative feedback*

- **Iterative feedback.** Finally, the process of testing and refining one's search based on the results of testing needs to be iterative so every refinement can be validated.

Together *Victory Stanley* and *The Sedona Conference Commentary* provide significant guidance on the known limitations of keyword search and how producing parties can defensibly conduct searches given these limitations. In practice, following this guidance can be challenging. One of the primary reasons why is that today's search technology simply doesn't make it easy to do.

ISSUES WITH TODAY'S SEARCH TECHNOLOGY

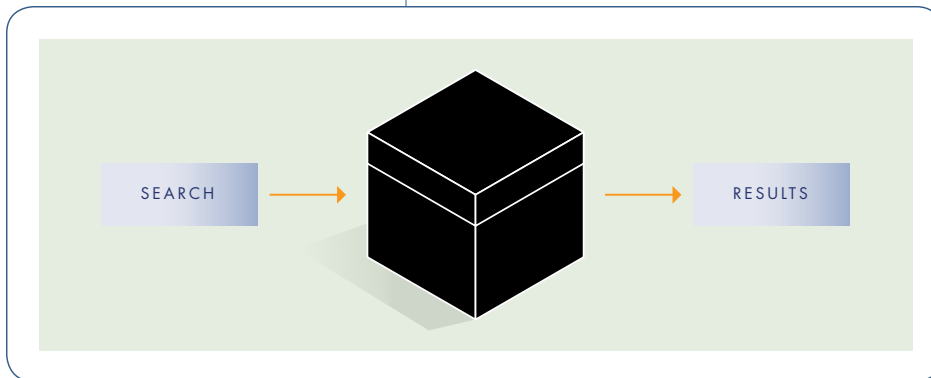
The short-comings of keyword search have been known for some time and many different technologies have been developed to address them in addition to the Boolean and proximity search technologies mentioned earlier. These technologies include wildcard and stemming technology, which was developed in order to address the issue of finding common word variations in specified keywords, concept search, whose objective is to find documents containing words with similar meanings to the keywords in a search, and fuzzy search technologies, which can be used to find misspellings of words.

Despite these advancements, today's search technology still has issues when it comes to e-discovery and specifically when it comes to enabling producing partners to easily perform keyword search in a defensible manner. The three primary issues with current search technology are:

1. **Tradeoffs between search efficacy and efficiency.** While wildcard, stemming, concept, and fuzzy search technologies will find more relevant documents, they will also find more non-relevant documents or false positives. For example, in order to find common variations of "diversity," one might run the following wildcard search, "divers*." This search will improve search efficacy by finding documents containing "diverse" and "diversity." However, it will also find false positive documents containing "diversion," "diversification," and "divers."
2. **High cost to test, sample and refine searches.** Today's search technologies are largely designed around the idea of running one search or query at a time. In e-discovery, however, it is typically necessary to run dozens of queries for a particular case. In this situation, following the best practices of testing, sampling, and refining each search can become very costly and time-consuming. It is also difficult to collaborate with the opposing party. In order for both parties to be satisfied with a set of keyword searches, considerable iteration may be required to balance the opposing objectives of each party. If these iterations are too costly then collaboration simply isn't practical.
3. **Manual documentation of refinement process.** It's not enough to use best practices. Producing parties need to document the practices they have followed so they are able "show their work" to the court. Currently, documenting the search refinement process is mostly manual. As a result, it is either done at a high cost or done inadequately if at all. Weak documentation significantly increases the risk of an adverse ruling if one's chosen search methodology comes into question.

BLACK BOX SEARCH

The issues with the search technology used in e-discovery today arise from the fact that the technology was not originally designed for e-discovery. It was primarily designed for enterprise search. However, what works for enterprise search doesn't necessarily work best for e-discovery search. In enterprise search, search technology has been designed to be a black box. A user enters a single search query and gets the results that match that query. How the search engine interprets the search query the user entered and what it actually searches for is hidden from the user.



Black Box Search

With black box search technology, no information is given about how the results were obtained.

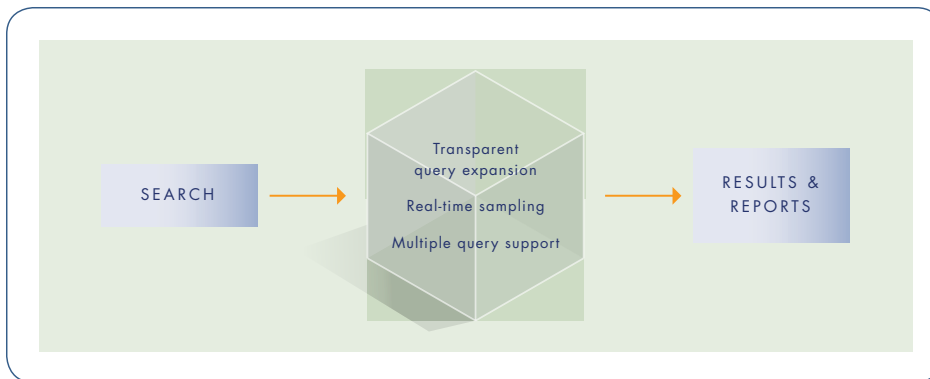
For example, in order to execute wildcard, stemming, concept and fuzzy queries, search engines go through a query expansion process. In this process, a search for “divers*” will be expanded to a search for all the terms that match the wildcard query, such as “diverse,” “diversity,” “diversion,” “diversification,” etc. The search engine will run this expanded set of search queries against an index and produce the results that match all of these queries,

but users will be shielded from the entire query expansion process.

Black box-type search engines have also been designed to run a single search query at a time. When a user enters multiple queries in a search, all of the queries will run as a single search, and the user will have no visibility as to which results are associated with which query. For example, a user that searches for “hiring OR interview” will get the results for the combination of the queries “hiring” and “interview.” They won't know that only five of documents contained “hiring” while 100 documents contained “interview.”

This black box design works very well for enterprise search because enterprise search users typically only want to run one search query at a time and see the most relevant results as quickly as possible. They don't want to know how the search was performed and don't mind if the search is over or under-inclusive as long as the first page or first couple of pages contain the most relevant results. The tradeoff of between search efficacy and efficiency is simply not as important in Enterprise search as it is in e-discovery.

E-discovery search is different, as users are looking for *all* documents, not just the most relevant documents, and it is critical to understand whether searches are over or under-inclusive. Over-inclusive searches result in higher review and production costs. Under-inclusive searches run the risk of missing key information. E-discovery users also typically need to run a lot of searches, and understand which documents are associated with each of these searches.



Transparent Search

With transparent search technology, users have more visibility into how the results are obtained and greater ability to perform sampling and refinement in real time.

THE NEED FOR TRANSPARENCY

Addressing the issues with today's search technology and meeting the unique needs of e-discovery consumers requires a new approach to search. Instead of a black box, search needs to become more transparent. E-discovery users need greater visibility into the search process so they can understand how the results were obtained and reduce the over and under-inclusiveness of keyword search. They also need

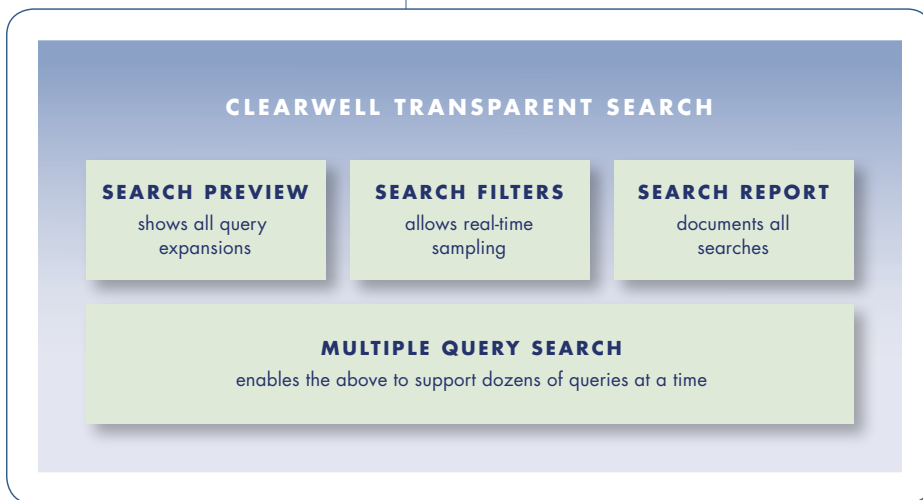
greater transparency so they can easily follow and document defensible best practices including testing, sampling and iterative refinement, and the ability to run multiple queries at once. Specifically, a transparent search solution that meets these needs should have the following key elements:

- **Transparent query expansion.** In transparent search, query expansion is no longer a “black box.” Instead, query expansion should be exposed to the user allowing the user to include or exclude expanded terms. In the previous example, for the search “divers*,” transparent query expansion would present a user with all the terms that matched the search “divers,” “diverse,” “diversity,” “diversion,” “diversification,” etc. A user would then have the ability to exclude false positive expanded terms, such as “divers” and “diversion,” from the search reducing over-inclusiveness and minimize the tradeoff between efficacy and efficiency.
- **Multiple query support.** When a search contains multiple keyword queries, such as “hiring” and “interview,” transparent search should provide visibility into the results for each individual query as well as the combination of all the queries. For example, with the search “hiring OR interview,” users should have visibility into the results for “hiring,” “interview” and “hiring OR interview.”
- **Rapid sampling.** Transparent search should support the ability to rapidly sample the results from all of the individual queries contained within a search. It should also be easy to take a random sample of non-matching documents in order to assess whether one or more searches have identified as many of the relevant documents as possible.
- **Automated documentation.** With transparent search, the results for each individual query that have been run as part of a larger search need to be automatically documented. Which query expansions have been included or excluded as a result of transparent query expansion should also be automatically documented. For example, transparent search would need to document that the user included “divers” and “diversity” and excluded “diversion” and “diversification” when documenting the search “divers*.”

INTRODUCING TRANSPARENT SEARCH

The Clearwell E-Discovery Platform's transparent search features support all of the key elements of transparent search to enable a more defensible e-discovery search process and enhance the ability to cull irrelevant information. The solution includes four capabilities that provide a new level of visibility, interactivity, and auditing of the e-discovery search process: Search Preview, Search Filters, Search Report, and Multiple Query Search.

Search Preview enables transparent query expansion by providing visibility into matching keyword variations for wildcard and stemming searches prior to running a search. This allows users to selectively include relevant variations or exclude false positive variations in their search query, removing irrelevant documents from search results.



Search Filters make rapid sampling possible. With Search Filters, users can filter their results by individual queries or variations, and sample the filtered documents to evaluate the effectiveness of their search. Clearwell also allows users to randomly sample non-retrieved documents to perform quality assurance that relevant documents are not being missed. This also allows users to rapidly identify false positive documents prior to review.

Search Report automates the documentation process. For each search, Clearwell creates a comprehensive report that documents all search criteria and provides

Transparent Search provides a new level of visibility, interactivity, and auditing of the e-discovery search process: Search Preview, Search Filters, Search Report, and Multiple Query Search.

detailed analytics of the results for both the overall search and the individual queries within the search. The report tracks search terms that were included and excluded during search preview, providing a defensible audit trail of search refinement decisions.

Multiple Query Search delivers the ability to run large numbers of queries simultaneously and provides reporting for both the overall search and the individual queries within the search. Large numbers of queries can be tested in minutes instead of days, dramatically decreasing the turnaround time needed to evaluate the effectiveness of keyword searches.

BENEFITS OF TRANSPARENT SEARCH

By addressing the key technology challenges of keyword search, transparent search can provide significant benefits to attorneys and litigation support professionals using search within e-discovery. First, parties that adopt transparent search can improve the defensibility of their e-discovery search practices through collaboration and best practices. By enabling iterative refinement and sampling, transparent search allows users to engage in collaboration or best practices when it was previously impractical to do so. This will enable producing parties to take the necessary steps to comply with the defensibility requirements of case law and reduce the risks associated with e-discovery.

Second, the use of transparent search can substantially reduce downstream production and review costs by removing false positives. For example, it is not uncommon for certain wildcard searches to generate results where 20-40% of the included documents are included due to false positive matches that can be removed by transparent query expansion. In cases of a reasonable size, this can result in thousands of dollars of savings on a single search query.

Finally, transparent search can dramatically reduce the time and cost required to complete the search and culling stage of e-discovery by reducing the time required to refine multiple queries from days to minutes. Currently, it can take hundreds of person hours to run significant number of searches one at a time, document the results of each search and sample, and refine each individual query. With transparent search, running multiple queries and documenting each of the individual results takes minutes. Sampling each of the individual queries takes seconds.

No search technology, including transparent search, is a “silver bullet.” Search will remain an imperfect science with the possibility of over and under-inclusive results. Despite this, search remains the best solution for reducing the ever increasing amounts of information to a reasonable level for review. While attorneys and litigation support professionals can’t completely remove the imperfections of search, they can take action to minimize the impact of these imperfections and meet the requirements of new e-discovery search case law. In doing so, they will be able to reduce the cost and risk of e-discovery and turn their focus to the substance of the case.

FOR MORE INFORMATION

For more information about Clearwell Systems Inc., or the Clearwell E-Discovery Platform, please contact us at:

Clearwell Systems

441 Logue Avenue
Mountain View, CA 94043
650.526.0600 tel
650.526.0699 fax
www.clearwellsystems.com
info@clearwellsystems.com

