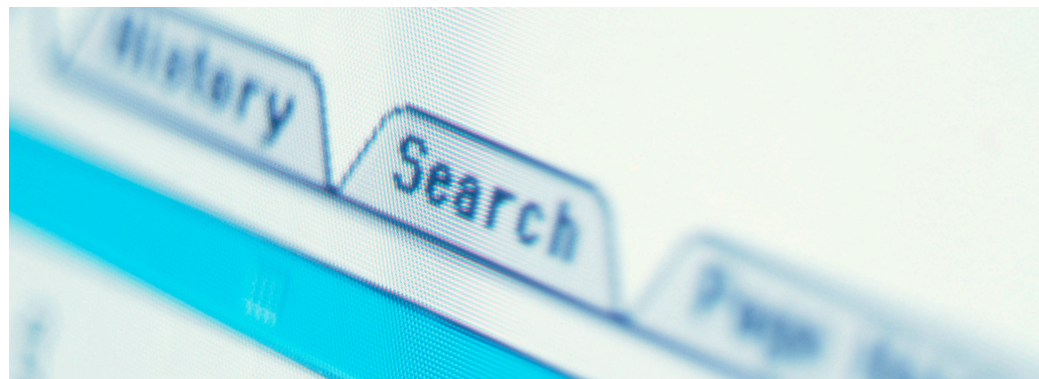


## Conceptual Search Technology: Avoid Sanctions, Prevent Privilege Waiver & Understand Your Data

By, David Chaplin & Regina Jytyla, Esq.

**David Chaplin**, Managing Consultant, heads the Advanced Search Technology division of Kroll Ontrack, Inc. ([www.krollontrack.com](http://www.krollontrack.com)). Mr. Chaplin is an expert in the area of conceptual information access technology.

**Regina A. Jytyla**, Esq. is a Managing Staff Attorney in the Legal Technologies division of Kroll Ontrack ([www.krollontrack.com](http://www.krollontrack.com)). Ms. Jytyla tracks and reports on the evolving law and technology in the areas of litigation readiness and management of ESI, electronic discovery, and computer forensics.



### Introduction

The current weakened economy has placed incredible budgetary pressures on law firms and corporations who must maintain profit margins despite tightening resources. Electronic data proliferation is economically neutral—it grows exponentially in good times or bad, and the costs to manage the discovery of electronically stored information (ESI) in the context of increased litigation and regulatory requirements continue to swell, despite the lack of a corollary budgetary raise. Adding fuel to the fire is a judiciary that is becoming less tolerant of attorneys who fail to adequately search and filter electronic data and as a result, inadvertently produce privilege-protected information, or fail to produce all the responsive data at issue. The seemingly incongruent

pressures of cutting costs while producing higher-quality of work are not actually at odds—both objectives can be achieved through the use of advanced search technologies. Electronic discovery is more complex, time consuming, and costly than ever. In this environment, utilizing advanced conceptual search technology becomes increasingly important in achieving economic efficiency and gaining a greater understanding of the documents under review.

### Courts are More Involved and Holding Parties to Higher Standards

In December 2006, the Federal Rules of Civil Procedure were amended to impose rigorous obligations to locate, preserve and produce electronic data and require early conferences between parties. These

amendments and subsequent case-law, have “raised the bar” and now require that attorneys and clients certify upon information and belief that all responsive, non-privileged data is located and produced. See, Fed.R.Civ.Pro. 26(g).

Today, US Courts require that parties take reasonable steps in preventing the disclosure of privileged electronic information. As such, if a party fails to take reasonable steps and as a result, produces privileged information, privilege is considered to be waived. For example, in the 2008 case of *Victor Stanley, Inc. v. Creative Pipe, Inc.*, Magistrate Judge Paul Grimm found that defendants’ reliance on an insufficient keyword search did not constitute a responsible precaution to prevent disclosure and found that privilege was waived on the inadvertently produced documents at issue. In addition, judges are becoming more involved in determining early search protocol and ensuring that parties are taking reasonable steps to recover and produce responsive data. For example, in the 2008 case of *Equity Analytics, LLC v. Lundin*, Judge Facciola required a party to submit an affidavit from its examiner explaining the limitation of a proposed search and how the search was to be conducted. Similarly, in the October 2008 case *D’Onofrio v. Sfx Sports Group, Inc.*, the court found the defendant’s search protocol to be “highly technical,” “highly restrictive,” and “fundamentally misguided,” and created its own search protocol that included search parameters and locations, time limits, and a requirement that the defendants restore newly discovered ESI.

## Limitations of Keyword Searching

Keyword searching looks for exact matches, or “keywords,” in text. Keyword searching only identifies the exact word that is queried; as a result, it is an “all or nothing” type of search. Keyword searching does not help a user to understand the data or how it relates to other data—rather, it simply looks for matching words. The fundamental deficiency of keyword search in a legal environment is the tendency of the approach to provide either over inclusive or under inclusive search results. The simple fact is; documents are more than just a bag of words and basic keyword searching removes context from the search through its simple process of matching terms.

There are many limitations inherent in keyword search. Keyword search alone does not assist the searcher in recognizing other related terms or help the searcher identify different or new language that should be included in subsequent searches. The keyword search engine does not learn from how words are used in context and through contextual analysis what other terms would be of value. Revealing other terms of interest based upon the analysis of how each and every term is used in a document and across all documents is what differentiates conceptual search from keyword search. With conceptual search the search process becomes a learning event that exposes unknown relationships between terms and documents, thus initiating true discovery in place of simple match.

## Broadening the Scope Through

## Conceptual Search

Conceptual search is defined as the ability to retrieve relevant information without requiring the occurrence of the search terms in the retrieved documents. The search technology that is most often used today is traditional keyword search (discussed above). However, many traditional keyword search engines have mimicked conceptual search through the use of synonym lists and other human-maintained query expansion approaches. True conceptual search retrieves relevant information in a way that does not require the presence of the search terms without the use of query expansion or independently maintained lexicons, taxonomies or synonym lists. This is why conceptual search is distinctly different from keyword search and is the key to why it is able to adapt to changes in language and the use of slang.

Conceptual search allows a user to locate information about a topic by understanding what words mean in a given context. For example, a conceptual search for “cellular” will return documents containing the words “mobile” and “Federal Communications Commission” in a document set involving telecommunications. Moreover, this intelligent search technology will return documents containing the words “genetics” and “molecular” in response to an identical conceptual search for “cellular” if its analysis reveals that the document set contains documents regarding biology. Abbreviations, acronyms, text and email slang, along with industry and corporate specific terminology

are continually progressing. Conceptual search can adapt to changes in the way language is used and the ever growing amount of information.

## Technology Behind Conceptual Search

Conceptual search engines measure subtle patterns and relationships that occur in language. Effective search requires the search engine to address synonymy (different words with same meaning) and polysemy (same word with different meanings). Using the example above, cellular means something different when the context is biology versus wireless communications. Conceptual search understands these differences and, in effect, smoothes out the idiosyncrasies of speech by analyzing words and how they are used in context. The measurement of how terms are used in context provides the conceptual search engine with the ability to learn new terminology without human intervention.

Concept searching technology employs latent semantic analysis, a technique which analyzes relationships between the terms contained in documents by identifying concepts related to the documents. The result is a powerful search engine that allows users to rapidly find information that normally would be time consuming or impossible to find with keyword searching alone. Additionally, the retrieval of related documents identifies documents related to the case and brings these documents together, facilitating greater understanding of the case

and providing valuable insights that can be used in strategic case planning.

## Dominant Approaches and Methods for Conceptual Search

There are two basic approaches to conceptual search: statistical and linguistic. Statistical methods usually learn from text and do not require any pre-built language models. Statistical methods analyze how terms are used within the document collection to be searched and determines the underlying structure of the language based on the documents in the collection. Linguistic methods, including natural language processing (NLP) and syntactic approaches, require models of language that are created and maintained by humans. These models are based on insight into the language and content, or from a training set of related text in order to find universal properties of language and to account for the language's development.

One may further classify conceptual search by scientific method of learning that is applied. Again, there are two basic approaches: supervised and unsupervised. Supervised learning requires feedback to improve and to initially specify what needs to be learned. Explicit examples need to be supplied to the system for the engine to learn. Unsupervised learning is fully autonomous and can arrive at an optimal solution without requiring user feedback or pre-defined training sets.

Finally, conceptual search technology can be query and non-query based. Methods have

been developed that enable conceptual search technologies to automatically cluster or folder documents that are similar in theme. These clusters are labeled and provide the business user with the ability to navigate a large set of information that is organized and appropriately labeled without having to issue a query. The ever increasing influx of information into critical knowledge management systems requires improved methods to automatically organize and make available documents, without requiring the user to know what search to perform. This approach can also be used to enhance or refine existing corporate taxonomies or to provide a "snapshot" of large document collections. Another important by-product of document clustering is topic / folder labels that provide insight into potential search terms for utilization in discovery or knowledge harvesting.

There are also two basic methods in producing conceptual search: automatic and manual. Automatic methods allows a user to present any source of information to the system without considering structure or syntax. The automatic method allows for the engine to learn as a new language is introduced to the document collection without any human intervention. Manual methods require humans to create and maintain a taxonomy, ontology or synonym list in order to create and maintain relationships. The knowledge is fixed and will have to be altered to account for new vocabulary or relationships.

## Conceptual Search is Important When Responding to Business Matters and/or Civil Suits and Investigations

Business professionals, attorneys and litigation support professionals are spending more and more time searching for information to make better and faster decisions. Concept search has been available for several years as a tool to help legal and business professionals review data that has already been collected.

Conceptual search reduces the number of queries, results sets, and redundant hits in the standard process of collecting, reviewing and producing documents in discovery. For example, conceptual search can be utilized to locate near duplicate documents enabling consistency in the review process (it is not uncommon for multiple versions of a document to exist but with slight variations causing the documents to not be exact duplicates). Near duplicate documents will always appear next to each other in the document matrix allowing for organization and consistency in review. Ultimately, conceptual searching techniques allow legal teams to retrieve the maximum number of relevant documents, including information that would not ordinarily be found through keyword searches.

Conceptual search also simplifies the process. The legal team enters a phrase or sentence and the technology organizes corresponding documents into groups of topics and sub-topics available for document review. For example, if a reviewer knows that all documents in a particular folder are related to

stock options and all documents in another folder are related to going out to lunch or birthday celebrations within an office, the reviewer will be able to move through the documents with the level of speed and precision needed to make the most efficient decision about whether the document is important to the matter at hand.

Further, conceptual search provides an intelligent information access layer that sits between the data and the person conducting the search. The analysis of how terms are used in a document and across all documents creates a statistical understanding of the entire corpus of text. This understanding is used to process and associate a query and compare the search terms with the statistical intelligence garnered from the entire set of documents. The value of this technology is important because:

- **Contextual Location of Data:** Relevant information is retrieved based on context, resulting in better and more informed decisions. The understanding of the difference between *cellular* biology and *cellular* phones is learned by the analysis of how words are used in context.
- **Faster Identification of Data:** A more advanced understanding of the information is achieved, facilitating faster location of relevant information via better, more accurate search results which provide quicker decision making ability.
- **No other Technology Needed:** The engine does not require a query language, providing

a faster path to productivity with no training required. The intelligence created by the engine smoothes out the idiosyncrasies of speech without any human maintained intelligence in the form of synonym lists or taxonomies.

- **Automated Application of the Technology:** An intelligent layer is created that understands your information and continues to learn, providing the ability to automate decisions without human intervention.
- **Learns More as Data Volumes Increase:** An intelligent layer that “learns” sits between the business professional and the critical business information, providing accurate and relevant search results as language and terminology change and shift.

Simply stated, conceptual search is the key technology that can facilitate better and faster business and legal decisions in a knowledge economy. Conceptual search provides a mechanism to deliver the right information to the right person at the right time. The ability to automatically contextually understand the information being search significantly improves accuracy and efficiency in any search task. In addition, concept search may be used for search term formulation, fact assessment, strategic analysis, and witness identification.

## Conceptual Search Can Improve Search Term Formulation

For years, lawyers have tried to develop the perfect set of search

terms, the unobtainable objective of which is to find all relevant data while excluding all irrelevant data. Taking an overly narrow approach to search terms results in the team missing relevant data, but taking an overly broad approach will leave the legal team with much more data than it needs to review.

Lawyers spend hours and hours making, refining, and fighting about search term lists. Typically, lawyers for the producing party want a small, narrow list, but lawyers for the requesting party want a large, broad list. But no human being is capable of developing a search terms list that factors into account the taxonomy and lexicon of the data, nor can any human anticipate all of the abbreviations, misspellings, or “code” language intended to deceive that are prevalent in the data. Concept search can help fine-tune search terms through identifying data that has the highest probability of yielding high ‘hit’ rates.

Concept search can be used to group the data before search terms are developed. The grouped data may be reviewed by the producing party and used to develop search terms to propose to the requesting party. The requesting party, on the other hand, could apply concept search technology to a set of production data that had been identified solely by the use of search terms. Analyzing the grouped data, the requesting party could then provide the producing party with additional search terms to apply against the main data collection. Approaching term-based data productions iteratively has always been the most-accurate

approach; however incorporating concept search technology makes the process even better.

### **Conceptual Search May Improve Fact Assessment and Case Strategy**

Lawyers frequently know very little about the facts of a case in the beginning days of investigation. The investigation may start with nothing more than an anonymous call to an ethics hotline or an allegation of potential wrongdoing by a single employee. Through the use of concept search, the legal team can analyze the data of an accused wrongdoer and quickly profile the subject matter of the data. As the legal team rapidly culls out the irrelevant information, the potential facts become clearer and other witnesses emerge as possible subjects of the investigation. In incremental steps, the legal team can then collect data of others and run concept search technology against that data for sorting and grouping. This technology will help the team for a picture of the facts more quickly and more cost effectively than the typical method of having a team of lawyers plod through every email or try to formulate guesses at search terms to zero in on the issues.

While conceptual search requires a query to add value, the intelligence gathered from the documents can also be utilized to group information by topic without a query. Automatic topic grouping (clustering) is able to organize documents into similarly themed groups. Topic labels are assigned to help in identifying what information topics and themes are

contained in the group. Large sets of unknown documents can present many problems in formulating a search strategy. Topic grouping utilizes the concepts present in the documents to reveal what topics and themes are contained in the document set. The ability to do this without a query provides immediate assistance in assessing a case because the technology allows the document to reveal what they are about in the context of the entire corpus of text. Similarly the exact same intelligence can be utilized to reveal near duplicate documents that are contained in the corpus of text. For example, emails about the company picnic will appear in one cluster and fantasy football in another. The creation of a labeled hierarchical tree of document clusters enables a quick “snapshot” of the document set while also providing the ability to navigate tree revealing clusters of likely relevant or non relevant data.

When a new investigation or lawsuit begins, lawyers must start the process of trying to answer the who, what, where, when, why, and how questions. Sometimes lawyers have a reasonably good understanding of the people, places, and things early in the case, but other times they do not. Rarely, however, will the lawyer possess that knowledge to the degree that allows full early case assessment and a full understanding of who the potential witnesses are and what happened in the case.

Concept search can dramatically improve the speed at which the lawyers develop their case theories, increase the accuracy of the

analysis, and decrease the expense of the process. Concept search can help attorneys identify people involved in the dispute, sift through mountains of data, and provide an objective, machine-generated group of data with similar context and improve the accuracy of the typical “search term” approach to data analysis.

### **Conceptual Search May Enhance Witness Identification and Location**

Starting with a limited amount of information about the case, concept search will help a legal team identify one or two witnesses who may have knowledge of relevant facts. Through the use of concept search technology, names of other potential witnesses may be dropped into search groupings without requiring the use of search terms, without knowing in advance the names of these individuals, without having to account for misspellings or abbreviations, and without having to look at the “to” or “from” lines in email headers. Armed with this information at the beginning of a case, attorneys should more quickly focus on the most important witnesses, even those who are not part of the organization, such as customers, suppliers, competitors, and potential wrongdoers.

In addition, having earlier witness identification information will help the legal team ensure that they have preserved data for the right group of custodians. Rather than having to start the data identification process by interviewing each person or by preserving “everything,” early use

of concept search can help the legal team hone in on who is a potentially important witness. The concept search results can then be fine-tuned with custodian interview and analysis to ensure the preservation plan is complete.

### **Conclusion**

The proven benefits of conceptual search cannot be denied. Relying upon a 2007 report by George Paul & Jason Baron titled *Information Inflation: Can the Legal System Adapt?* Magistrate Judge Facciola stated in the 2007 case of *Disability Rights Council of Greater Washington v. Washington Metro Transit Authority*:

I bring to the parties’ attention recent scholarship that argues that concept searching, as opposed to keyword searching, is more efficient and more likely to produce the most comprehensive results.

The great advantage in information retrieval today is the ability to utilize the right tool for the job at hand. When locating an email from a specific person, a simple keyword search may be the best approach. However, conceptual search provides the additional analysis required when locating related emails or revealing unknown relationships between custodians. Properly combining keyword and conceptual search approaches into a blended search strategy provides the practitioner the ability to view the corpus of documents through a different lens. The result can be compared to the different ways light is refracted through a diamond depending on how the gem is moved or rotated.

The same diamond appears totally different to the eye yet it is, in fact, the same diamond. As the volume of electronic data that is created increases exponentially and judicial expectation to “get it right” becomes stronger, the need to incorporate advanced search technology into a standard keyword search protocol becomes clearer. In short, conceptual search can result in cost-savings, minimize production errors, and provide timely insight into voluminous data sets.

Kroll Ontrack  
9023 Columbine Road  
Eden Prairie, MN 55347  
800 347 6105  
**[www.krollontrack.com](http://www.krollontrack.com)**

Copyright © 2009 Kroll Ontrack Inc.  
All Rights Reserved.

All other brands and product names are trademarks or registered trademarks of their respective owners.