



*Efficient Defensible Discovery*

# Case-Centric Search and Sampling Protocols

The Key to Defensible Search and E-Discovery Cost Containment

February 2009 | A Servient™ Whitepaper

## Table of Contents

<i>Introduction</i> .....	1
<i>Importance of Search and Sampling in Electronic Discovery</i> .....	1
<i>Understanding Search Technology</i> .....	2
<i>Understanding Sampling Protocols</i> .....	6
<i>The Need for a “Case Centric” Search Protocol</i> .....	7
<i>Introducing Centric Search™</i> .....	10

## ***Introduction***

Electronic discovery presents the information retrieval profession with one of the most challenging set of requirements. Lawyers want to achieve near perfection in finding the relevant documents from within very large data sets, while at the same time must avoid driving up costs by returning large percentages of irrelevant documents into the review set.

The true solution to the search challenge is to invoke an iterative approach of search, sampling and refinement. An understanding of search technology and the importance that process plays in achieving an efficient and defensible discovery protocol is critical for today’s litigator. This whitepaper presents a “Case Centric” iterative approach to the design and implementation of a defensible search protocol.

## ***Importance of Search and Sampling in Electronic Discovery***

For many years commentators have warned of the implications on litigation of the increasing volume of electronic information. The growing data volume issue has been so widely discussed that no doubt many practitioners equate the issue with the proverbial “the sky is falling” warning.

In reality, however, the volume of data that must be addressed by lawyers is growing at an alarming pace. Just a few years ago, we saw volumes in the 1.5 GBs per custodian range. In 2009, it is not uncommon to be confronted with custodians with 10-20 GBs of stored information. To put that in perspective, a single witness with 10 GBs of active data possesses 75,000-100,000 documents (yes documents, not pages).

This data volume has quickly overwhelmed the litigation support technology utilized in the “coding and scanning” days of the industry. The cost for the inefficient “TIFFing” of data now destroys litigation budgets. The volume of documents quickly reaches into the millions of records for a small group of witnesses and is beyond that capacity of older litigation support platforms. And, even if the processing budget was of no concern and the technical challenges could be overcome, the cost to review the volume of data is simply unmanageable.

There is little debate that some form of search filtering is necessary to narrow the data set that must be reviewed and thereby contain the skyrocketing cost of litigation. Indeed, the creation and implementation of a defensible search filtering protocol has emerged as the most important aspect of electronic

discovery. An effective search protocol is the number one way for legal teams to control the cost of electronic discovery.

The Courts have placed an increasing focus on the defensibility of the search process. A number of cases in 2008 addressed the importance of a defensible search protocol including *Victor Stanley, Inc. v. Creative Pipe, Inc.*, 250 F.R.D.251 (D. Md. 2008); *United States v. O’Keefe*, 537 F.Supp.2d 14 (D.D.C. 2008); and *Equity Analytics v. Lundin*, 248 F.R.D. 331 (D.D.C. 2008). As one comes to understand that search is used as a determinative process to make calls on the potential responsiveness of documents (i.e. if the search does not return a document, the document is neither reviewed by the legal team nor produced in the litigation), the defensibility of the search technology and protocol quickly comes to the forefront of the analysis.

The design and implementation of a defensible search and sampling protocol is a required core competency for today’s litigation.

### ***Understanding Search Technology***

A prerequisite to the design and implementation of a defensible search protocol is an understanding of the features and limitations of search technology.

It is important to be reminded from the outset that perfection has never been achievable in document discovery. Even manual processes used in the days of paper discovery did not achieve perfection. Judge Grimm repeatedly stressed the term “reasonable and reliable” to define the required standard in *Victor Stanley*. In order to construct a “reasonable and reliable” search protocol, one must consider the limitations of the selected technology so that reasonable accommodations can be made to increase the reliability of the search process.

An understanding of search technology starts with consideration of the nature of the data to be searched. As technology offerings differ in the way that they retrieve different types of data, knowing how the data is processed and handled by the search technology is critical. Almost all search applications require some form of data preparation prior to utilizing the search technology. The interaction between pre-processing data to be consumed by the search application and functionality that is available at run-time is a concept that technologists have wrestled with for decades.

For example, many users consider an email message to be a pure textual document – i.e. the email header block is part of the text of the message. When one digs a little deeper, they will discover that there is additional information that may not be included in the user’s assumption of the text of an email, such

as the internet header. And, some search applications will consider the To, From etc., as fielded information. In such a circumstance the user may be required to search within fields in a database to find the semi-structured information; whereas, a search within the full-text may only cover the body of the email.

This potential variance in the way that different applications consume and access the varied data involved in electronic discovery underscores the need of the legal team to understand the nature of the selected application. “Black Box” proprietary systems run by providers who simply license technology raises serious doubts from the outset with the defensibility of the process. The Sedona Conference echoed this sentiment when it noted the risk involved when “e-discovery and litigation support vendors that use the technologies may be several degrees of separation from the original developers.”<sup>1</sup>

The effectiveness of search technology is generally measured by recall and precision which are expressed as:

$$\text{Recall} = \frac{\text{Total responsive documents returned by search}}{\text{Total responsive document in total data set}}$$

$$\text{Precision} = \frac{\text{Total responsive documents returned by search}}{\text{Total documents retrieved by search}}$$

In electronic discovery, the recall rate measures the effectiveness of finding the responsive documents from within the litigation hold data set. And, the precision rate measures the extent of the “false positives” (non-responsive documents) that are returned by the search. Even if the electronic discovery searcher utilizes a search protocol that achieves a high recall rate, the results may render the entire process useless if the precision is so compromised that the review burden created by the inclusion of “false positives” becomes overwhelming.

In other words, electronic discovery presents one of the most challenging set of requirements for search technology. A lawyer defines the requirement as the need to find all responsive documents to comply with the discovery rules while at the same time returning as few non-responsive documents as possible to limit the review burden. A technologist interprets this as requiring achievement of a perfect recall with a high degree of precision; however, generally speaking,

---

<sup>1</sup> *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery.*

search technology usually either sacrifices search precision to achieve higher recall rates or accepts lower recall rates to improve precision.

A single run of simple keywords alone is generally ineffective at identifying responsive documents from within a litigation hold data set. However, there are many search technologies that aid in the process.

Query refinement may involve the use of features that address synonymy (different words with similar meaning). For example, the use of the term “contract” as a keyword will not return documents that contain the term “agreement.”

Common search features that address such limitations include:

- Thesaurus - Either through user selection or through automated query expansion, a thesaurus allows for the inclusion of known synonyms.
- Taxonomy - Either through user selection or through automated query expansion, taxonomy allow for the inclusion of lists of known relationships between hierarchal concepts. (Similar concept to Ontology)
- Stemming - Either through user selection or through automated query expansion, stemming allow for the inclusion of words that share the same root as identified by a stemming algorithm, such as Porter stemming.
- Wildcarding - The user enters a portion of a word with a single or multi-character wildcard operator to return variants of the partial word.

While these technologies may increase the recall rate of a given query, the query expansion will almost certainly reduce the precision. In other words, while the query may return more responsive documents, the query may significantly increase the number of “false positives” and break the review budget.

The legal team may also improve the precision of a search by implementing Boolean search logic. By combining terms with conjunctions (i.e. and, or), phrases and proximity operators, the query may be refined to return a narrower data set. Increasing the complexity of the search with the use of multiple Boolean operators and nested queries will further narrow the result set.

The narrowing strategy, however, presents a risk of reducing the recall rate of the search. While the number of non-responsive documents returned by the search may be reduced and the review burden minimized through use of complex queries, narrower queries may result in missing responsive documents.

Additional search approaches such as classification and clustering also can be considered to aid in the attempt to identify responsive documents. Text classification attempts to assign documents to different pre-selected topical classes based on a statistical analysis of the characteristics of the documents. While there are many variants of text classification, Bayesian classification and Vector space classification are the most common approaches. In essence, text classification evaluates a test set of data with known classification mappings, and attempts to map the remaining data within the existing classifications.

In contrast to text classification, clustering technology involves the automated grouping of documents using a vector space model; clustering allows for unsupervised grouping of documents that can generate insight into the common thematic topics within the data set.

An additional search approach that should not be overlooked is relevance feedback, as it holds a lot of promise in e-discovery. Relevance feedback involves iterative query refinement to improve the search. Basically, the user is presented with the results of a search, gives feedback as to the relevance of results, and the technology considers the feedback to improve the results. Demonstrating that the technology theories are not as new as the legal profession may have come to believe, relevance feedback is seen in the literature in 1971 in the Rocchio algorithm.

As the legal team incorporates search technologies into the case strategy, it is important to be mindful of Practice Point 7 stated by the Sedona Conference: “Parties should expect that their choice of search methodology will need to be explained, either formally or informally, in subsequent legal contexts (including in depositions, evidentiary proceedings, and trials).”<sup>2</sup> A major challenge with proprietary search programs that incorporate various statistical search algorithms is that, if such technology is used as a decisive filter to control the assignment to review (and thus controls the selection for potential production), the basis for the algorithm and data as to the reliability of the algorithm is essential. Yet, this information is not readily available to legal teams.

There is a tremendous amount of research and innovation underway in the search technology disciplines. The near future holds a lot of promise of

---

<sup>2</sup> *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery.*

emerging technologies that will aid in the e-discovery challenge. Legal teams should continue to evaluate and explore new technologies. But, legal teams are also in the unenviable position of requiring a solution today to meet the current case needs; in so doing, the team must be mindful of the requirement to establish a defensible process.

Some legal commentators (and many in the sales profession) have made the mistaken leap of faith that a statistical based clustering approach alone reliably identifies responsive documents. There is a lack of scientific support for such an assertion in the academic research; we have found no verifiable support thus far in our research and development. Indeed, many statistical algorithms tend to increase the size of the result set because of query expansion thus increasing the risk of adding high volumes of “false positives” if used as a decisive filter.

It is incorrect and indefensible to assume that technology alone can be used as a determinative tool to control what documents are pushed to review. The academic research, actual experience and just plain common sense counsel that an iterative search and sampling process conducted by a knowledge searcher using appropriate technology will yield the most effective and efficient search results.

### ***Understanding Sampling Protocols***

Statistical sampling of data sets has always been a recognized part of the information retrieval discipline. Sampling is widely used in data mining to construct models to make predictions about the entire data set. The importance of sampling in the e-discovery context can be seen in *In re Seroquel Prods. Liab. Litig.*, 244 F.R.D. 650 (M.D.Fla.2007), where the court instructed that “[c]ommon sense dictates that sampling and other quality assurance techniques must be employed to meet requirements of completeness.”

There are a number of different types of sampling that may be used, including:

- Random Sampling - Every document has an equal chance of being included in the sample set.
- Systematic Sampling - Every N<sup>th</sup> document is selected to be included in the sample. (This has limited utility in e-discovery sampling)
- Stratified Sampling - Documents of similar nature are grouped together and random sampling is then conducted within the subgroups.

Sampling is helpful to the searcher during the process of query refinement. In such a setting, the searcher can informally review samples of records returned by different subparts of a query to gain a better understanding of the context of the data and the effectiveness of the query segment. The searcher should also create samples of the overall result set before assigning the complete set for review to confirm the query’s effectiveness in terms of recall and precision. A pre-review responsiveness sample can avoid the costly review of data sets that contain large volumes of “false positives.”

A defensible search process also should include some form of quality assurance sampling to test the reliability of the search protocol. A sampling strategy should be designed to sample the data that has not been assigned to review to test the confidence level of the search protocol.

The effectiveness of sampling has a correlative relationship with the size and quality of the sample. The primary challenge with implementing sampling into the e-discovery workflow is the budget for reviewing the required samples. An inadequate sample, however, can create a false sense of confidence in the effectiveness of the search.

### ***The Need for a “Case-Centric” Search Protocol***

An Iterative process that involves continual query refinement and repeated sampling will generate the most reliable results and achieve the most cost containment. Luckily, attorneys are well versed in the iterative process because the process of legal analysis and case development itself is iterative. The best method to design a search protocol is to follow a legal analysis approach. We refer to this as the “Case Centric” search approach.

When an attorney is first called into a matter, the client usually conveys a summary of the dispute. While the lawyer understands that his knowledge is only partial at this stage, and that more information will be derived as the investigative phase continues, the lawyer nevertheless starts the legal analysis of the matter.

In the analysis, the lawyer begins to categorize known information such as:

- Identity and roles of the key witnesses
- Important events such as meetings etc.
- Important date periods
- Known key documents such as contracts and records as well as pointed correspondence that is typically sent as a prelude to litigation

The design of an appropriate search protocol mimics the steps in legal analysis. The legal team should identify the key witnesses and explore the nature of the data collected from the witnesses to discover terms and aliases that can be used to identify the key witnesses within the data set. The legal team should also consider segmenting topical queries by important dates and events so the query terms are bounded to data that has a higher probability of being relevant. Finally, the team should carefully review known relevant documents and pleadings to help in the identification of appropriate terms and concepts. The upfront work necessary to start the search design not only mimics legal analysis, it is in practice an aid to the legal evaluation of a case.

Too often, e-discovery search is merely an exercise in listing simple keywords to be used as a gross level filter. It is much more effective for the legal team to begin the process by searching for specific, important and relevant information without worrying about locating all responsive documents. This of course requires that the team have access to the complete data set and not just a portion of the data received back from a processing shop.

Such a process also requires that the team conducting the search be intimately involved in the case analysis and development. The role of litigation support should no longer be a “support role”, they must be centrally involved in the case in order to design an effective search protocol.

In designing the early strategy, the team should think in terms of multiple queries, not just a single large query. Design queries to attempt to locate specific documents related to different aspects of the legal analysis such as specific concepts in the litigation (i.e. the negotiation of the letter of intent at issue etc.). In constructing the queries, the team should informally sample documents that are returned by elements of the query to gain an understanding of the context of the terms located in actual documents. Of course, if the team has a collaborative opponent (which is often theoretical we know), the team can integrate the opponent’s suggestions to queries allowing for the intelligent negotiation of the search protocol.

The team should next push random samples of the various query results to review. The sampling process will allow the team to evaluate the effectiveness of the queries and avoid pushing high volumes of irrelevant documents through for expensive review.

The early review of documents returned by highly targeted and tested queries provides the legal team with early visibility into the important documents in the litigation. Most importantly, the early review will provide the team valuable insight into the nature of the documents that will allow the team to continue to

iterate through additional broader queries to reach a final document set that governs the review. The iterative approach of query refinement, sampling, understanding the nature of the returned data and further refinement, will produce the most effective and defensible search protocol.

As a final step of the process, the legal team should design and implement a sampling strategy to test the data not committed to review to evaluate the reliability of the identification of relevant documents. This strategy must be tailored to the nature of the case and the available budget. Few litigation budgets can support substantial costs for lawyers to review documents that the team expects to be irrelevant; but this is exactly what the sampling of the remaining data set entails.

To balance the budgetary constraints with the need to conduct quality assurance on the search protocol, the legal team should consider meaningful groupings of the data and then sampling of the groups – a variant of which is known as stratified sampling. The team should attempt to identify groupings that would produce an increased probability of containing relevant documents. For example, the team may first group the data by relevant date ranges, groups of important witnesses and search terms that were too broad for use as a filter. Sampling subgroups of the remaining documents will often allow for more manageable samples while achieving a threshold confidence measure for the process.

The benefits of an iterative “Case Centric” approach are self evident to the experienced litigator. The process is driven by the issues involved in the case, it provides early visibility into the data supporting the legal analysis and it allows for cost containment.

There are a number of requirements to implement the iterative “Case Centric” process, including:

1. Access to the complete litigation hold data set
2. Scalable search technology that handles multiple, complex searches
3. A search platform that gives visibility to the data to assist in query design
4. A search platform that provides meaningful feedback on the effectiveness of queries
5. The ability to informally sample and preview the content of documents while developing search queries

6. The flexibility to design random samples of search results and perform sampling of groups of documents not submitted for review
7. Tracking and management of the sets of data produced by each query that have been committed for review
8. Seamless integration to a true review platform that permits efficient review of documents
9. Full reporting capabilities to document the history of the process

Of course, an iterative approach also requires the commitment of the legal team to integrate greater analysis into the process, as well as appropriate litigation management. The legal team can not push electronic discovery off until later stages in the case – waiting until the eve of depositions will simply not allow sufficient time to conduct an iterative search process.

### ***Introducing Centric Search™***

Servient’s *Centric Search™* application provides the technology platform necessary to design and implement a defensible, iterative search protocol. *Centric Search™* provides the legal team with access to, and control of, the entire litigation hold data set. Because of the iterative nature of the process, the legal team must have direct access to search, sample and preview documents from the entire litigation hold. A workflow that involves sending data out to processing shops to cull and process for inclusion of a subset of the data into an off-the-shelf review system does not support the iterative protocol.

*Centric Search™* can handle huge volumes of data and efficiently execute complex, multiple queries. The searcher must be able to interact with data to develop the optimum queries; the user can’t wait minutes or hours for results. *Centric Search™* can execute complex queries on millions of records in a matter of seconds.

*Centric Search™* provides various automated views into the nature of the data set. Providing visibility to information such as data statistics, unique email addresses, domains, likely related names and aliases etc. provides the user with enhanced knowledge to create the optimum search strategy.

It is also important for the user to have access to meaningful statistics to evaluate the effectiveness of each sub-component of the query. *Centric Search™* utilizes advanced statistical algorithms to identify query fragments that are likely to produce high numbers of “false positives.” *Centric Search™* gives

the user meaningful feedback to focus the searcher on the portions of the query that should be refined.

*Centric Search™* also provides for the immediate preview of the content of documents themselves throughout the workflow. An iterative search process requires that the user be able to quickly evaluate the context of search hits within documents to make meaningful refinements. *Centric Search™* allows the user to take informal random samples throughout the search process to gain a more reliable overview of the nature of the documents.

Support for statistical sampling of the search results, as well as quality assurance sampling of the documents that have not been committed for review, is built right into the *Centric Search™* platform. Tight integration with a full-featured review platform allows for the efficient review of the samples. Lawyers are expensive knowledge workers – while reviewing the samples, the system should capture responsive, topic and privilege calls so that re-review of the documents later in the litigation is not required.

*Centric Search™* tracks and documents the entire search process. This allows for automated management of the iterative process. As lawyers learn more about their case and the documents, they can return to the overall data set to continue running additional queries on the segment of documents that have not been committed for review. Full reporting throughout the workflow allows for documentation of the defensibility of the search protocol.

In the end, *Centric Search™* enables the legal team to implement a creative approach to attack the electronic discovery problem. An iterative, “Case Centric” protocol powered by *Centric Search™* provides early case visibility, increased clarity of analysis, improved defensibility and ultimate cost containment.

**For More Information Contact:**

Servient, Inc.

Toll-Free (866) 590-4893

[sales@servient.com](mailto:sales@servient.com)

[www.servient.com](http://www.servient.com)

© 2009 Servient, Inc. All Rights Reserved