eDiscoveryJournal
Unique Perspective. Independent Insight. Pragmatic Advice.

The eDiscoveryJournal Report:

Guide to Enterprise Search for eDiscovery

By:

Greg Buckles & Barry Murphy

About eDiscoveryJournal

eDiscoveryJournal offers unbiased information, pragmatic advice, and a unique perspective on hot eDiscovery news, trends, and best practices. Whether it's gaining insight on the news of the day, researching the right technology or service solutions, or finding expertise to answer your specific questions, eDiscoveryJournal is the e-zine you want to visit every day.

About the Authors

Greg Buckles is an independent eDiscovery consultant specializing in enterprise technology and work flow solutions with over 20 years in discovery and consulting. His career spans law enforcement, legal service provider, corporate legal, law firm and legal software development. This deep and diverse background combines with exposure to the discovery challenges of Fortune 500 clients to provide a unique industry perspective.

Barry Murphy is the founding Principal of Murphy Insights and a thought leader in all things retention – eDiscovery, records management, and content archiving. Previously, Barry was Director of Product Marketing at Mimosa Systems, a leading content archiving and eDiscovery software. He joined Mimosa after a highly successful stint as Principal Analyst for eDiscovery, records management, and content archiving at Forrester Research

Disclaimer:

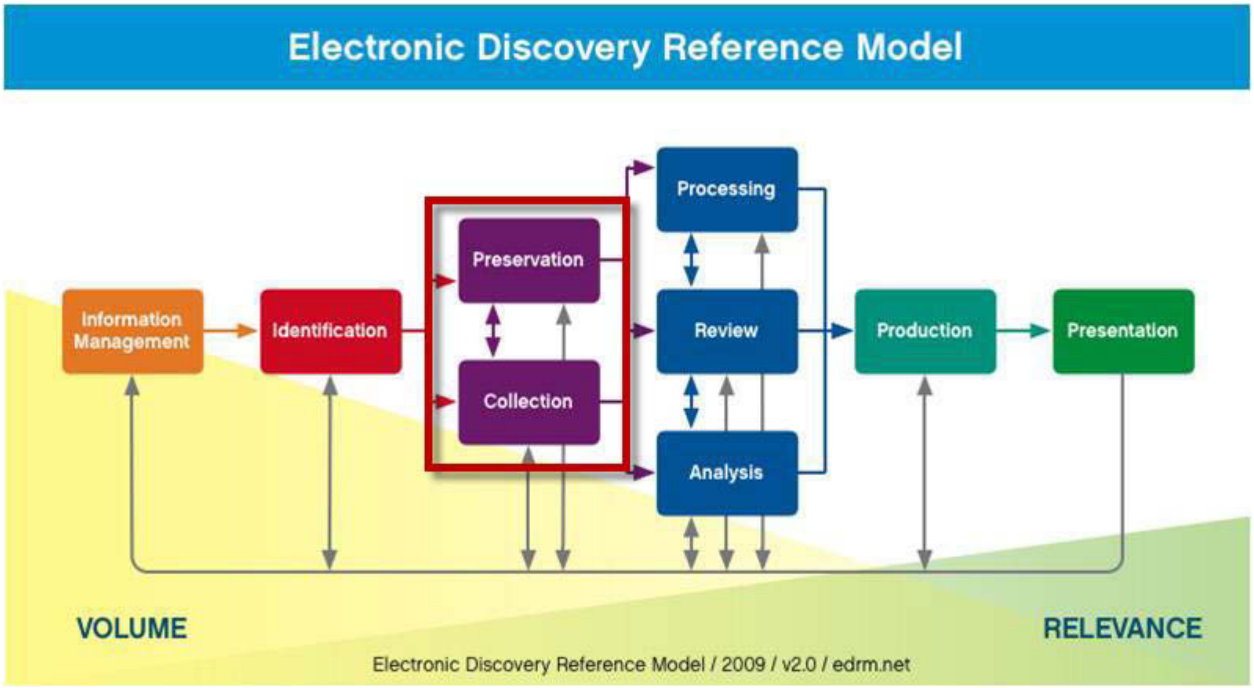
eDiscoveryJournal is not a law firm. All expressed opinions and content are provided for general educational purposes only and are not specific legal advice, even if the author is a practicing attorney. Neither eDiscoveryJournal nor the information contained herein should be used as a substitute for competent legal advice from a licensed professional attorney in your state.

eDiscoveryJournal believes reasonable efforts have been made to ensure the accuracy of all eDiscoveryJournal original content. Content may include inaccuracies or typographical errors and may be changed or updated without notice. All eDiscoveryJournal original content is provided "AS IS" and while we endeavor to keep the information up to date and correct, we make no representations or warranties of any kind, express or implied, about the fitness for a particular purpose, completeness, accuracy, reliability, suitability, or availability with respect to the information, products, services, or related graphics for any specific purpose. Any reliance you place on such information is therefore strictly at your own risk.

In no event will eDiscoveryJournal or any of its contributors be liable for any direct, indirect, punitive, incidental, special, or consequential damages or damages for loss of profits, revenue, data, down time, or use, arising out of or in any way connected with the use of the document or performance of any services, whether based on contract, tort, negligence, strict liability or otherwise.

Introduction

As corporations cautiously develop litigation response and information governance infrastructures, there is a desire to convert outsourced, third-party-managed, reactive processes to processes that are centrally managed and at least somewhat standardized. The transition is not a simple one; it involves contracting, developing or acquiring the people, processes and technologies to fulfill eDiscovery requirements. Meanwhile, corporations grapple with new consumption models such as software-as-a-service (SaaS), cloud infrastructure, managed services and more. Corporations' eDiscovery obligations lie on the left hand of the Electronic Discovery Reference Model lifecycle with prioritization on Information Management, Preservation and Collection.



Electronic Discovery Reference Model / 2009 / v2.0 / edrm.net

Many corporations consider completely in-sourcing the entire eDiscovery lifecycle, but must justify the return on investment. This may be challenging unless the company has an unusual litigation burden or is under heavy regulatory requirements. In the last few years, early adopters have implemented enterprise content management (ECM), archives, collection/preservation appliances, forensic imaging and other single purpose tools to address eDiscovery. Search is the common functional component across these solutions, but of course each solution has its own "search" capabilities. Some enterprise software providers have consolidated software into multipurpose platforms to maximize shared search, storage, indexes, databases and more, but it leaves enterprises wondering what the right approach to search and eDiscovery is. While it might seem like enterprise-wide search is the answer to the eDiscovery challenge, the reality is that it's a bit more complicated than that.

Contents

Selective vs. Enterprise Wide Search

eDiscovery search providers have formed at least two basic camps to address the challenges of enterprise scale, connectivity, heterogeneity, identity management, accessibility and more. There is a philosophical struggle between matter-driven indexing of targeted sources and an enterprise wide index covering all electronically stored information (ESI).

In theory, a single, omniscient corporate index offers simplicity and efficiency. But, in reality, there is a cost to managing such an index. Are technology providers capable of delivering solutions that can meet the demands of true enterprise search? And are corporations ready to execute such capabilities? This whitepaper explores these questions while also providing an overview of the available search technologies along with potential advantages and known eDiscovery hurdles. We will keep our focus on the dominant eDiscovery usage cases that rely on negotiated Boolean searches, though it will be worthwhile to monitor over time how more advanced analytic search method are treated by courts.

Search Architectures

Before diving into comparisons, it’s important to understand the types of search options available as they relate to eDiscovery search.

- **Inventory Index – Context Search.** Operating and storage systems keep the name, location and other context metadata regarding a piece of ESI. Traditional hard drives use some format of File Allocation Table (FAT), while content management systems use an actual database to track this information. This information can be indexed very quickly if the search engine can connect to the data source. Many systems will inventory or survey ESI sources first to enable the eDiscovery team to assess potential relevance prior to fully indexing the actual ESI.

- **Full Text Index - Content Search**. Mention search and most folks will think of Google. You build a full text search index by opening files with a viewer that breaks the rendered text into unique terms and records the relative position of those unique terms within a compiled table. Almost all full text indexes support searching for phrases and Boolean connectors such as AND, OR and NOT. Some have extended functionality that supports proximity searches (within X terms), stemming (locate, located, locating, etc), pattern matching (SSN’s, phone numbers, etc), wild cards and relevance weighting (though this is less common). While these extended search functions provide value and allow users to conduct more granular and complex searches, three is a trade-off – a larger, more difficult to manage index. A simple index can be compressed down to 5-10% of the original ESI data set size, especially if there are a lot of ‘Noise Words’ that it gets to skip. A basic index like this can tell you if a document contains a term, but not where that term is located in relationship to other terms. Thus, while the ability to search for phrases is lost, the index size remains small and search speed may be much faster. The average index size of most traditional eDiscovery applications that support the expected advanced features ranges from 20% to 50% of the ESI data set. Adding in concepts, taxonomies, facets and all the other bells and whistles that allow for faster review of the data, the index can actually be larger than the original ESI. That presents a challenge for storing and managing the index.

- **Selective Search – On Demand Indexing**. Traditionally, potentially relevant sets of files, email and other ESI had to be collected before being processed and indexed. The collection was usually performed by a service provider with special software and technicians. Although custodians might use their installed desktop search to locate potentially relevant ESI, most litigation support personnel knew that these light-weight search systems were designed to retrieve the top hits and were not suitable for eDiscovery searches. Many web search companies have tried to create appliances for eDiscovery, but none so far have found any level of market acceptance. Email archiving platforms were the first enterprise applications to create audited, transaction-based search to meet eDiscovery requirements back in 1999-2000. Email was and is still the highest priority target of discovery requests, but archiving systems were limited to what was in the archive. In the 2001-2003 timeframe, the industry saw the first appliances adapted for eDiscovery search of unstructured ESI in-place on network file shares. Currently, corporations have many options for searching ESI ‘in the wild’ that span the vast majority of structured and unstructured sources.

- **Indexing**. Selective or on-demand indexing does not mean having to index all possible sources every time you have a new matter. It is expected that priority sources like executive and corporate secretary shares will be proactively indexed and incrementally indexed as needed. Rather, the sources indexed are determined by their importance and potential responsiveness. If you know that there is a central contract folder that may contain files responsive to most of your contract dispute litigation, then you would proactively index that folder and schedule monthly updates. The primary difference between a selective indexing and enterprise-wide index architectures is the design for reactive burst capacity (for selective indexing) versus continuous updates (for enterprise-wide indexing). The selective indexing solution’s graphical user interface will support rapid designation of sources and load balancing of high volume indexing. One is built for sprints while the other is designed to run continuously. The actual indexing mechanisms and the indexes that result are similar.

- **Crawl**. Crawl search is worth a short mention here. Several systems derived from forensic imaging applications will read the binary ones and zeros written on the physical hard drive to try to match the raw text against search terms. Generally, they are not opening or unpacking file containers to ‘view’ the file. Instead they use pattern matching technology such as GREP, Generalized Regular Expression Parser, to look for sequential characters that match the search criteria. As you might imagine, this crawl can be very processor and memory intensive. It can also take a substantial amount of time – in some cases it is comparable to the actual indexing time. The difference is that once an index is built, you can search it many times, while a crawl search must reprocess all of the target ESI every search. There are definite scenarios where crawl search can supplement or replace indexing, primarily when you know that you will only have to search a very large volume just one time. Because crawl engines must have custom filters to interpret any file that is not a raw ASCII text format, you should always test them heavily for critical file types that are encoded or compressed such as the new Office 2007 formats.

- **Enterprise Wide Search Index**. We define an enterprise wide search index as any system that proactively, continuously indexes all or most of the active data sources within the corporate enterprise from a single search interface. There are very few players in this market, but they have a very compelling message. Some work through federating the search criteria across multiple native indexes (Sharepoint, desktop, etc), while others create a homogeneous set of indexes from multiple sources. Remember that the goal is to provide a single live search across the enterprise.

Preservation Methods

In the context of eDiscovery, among the primary use-cases for search technology is to find and preserve potentially responsive documents. Even in this use-case, there are various approaches that corporations can take based on their specific requirements and situation:

- **Collect and Preserve**. Collect and preserve is one approach to executing litigation holds on data. In this scenario, a copy of the potentially relevant ESI is collected and moved to a central repository for preservation along with the context information needed to authenticate it (see the eDiscoveryJournal Report: How to Fit Defensible Collection into Information Governance Strategies for more information on the three components of ESI – content, metadata, and context). A typical preservation collection proactively collects based on very broad custodial, chronological and content criteria, generally before there has been any agreed upon discovery criteria. In one sense, this can be seen as overly broad, but it is based on the best knowledge at the time and so reduces the risk of spoliation or loss.

- **Preserve in Place**. Preserving ESI in place can be done with a variety of methods, each with their own trade-offs between security, effectiveness and risk. Traditional legal hold notifications relied on the actual custodians to preserve their own ESI, generally with the expectation that they would refrain from actively deleting potentially relevant email and files. The majority of private and public businesses still rely custodian driven preservation in part or in whole. The problem is that most users do not have the rights or technical know-how to actually fulfill this obligation in the modern dynamic enterprise environment. It also assumes that custodians are cooperative, which is not always the case. For in-place preservation, one approach is to modify the security access rights to ESI where it lives so that potentially responsive data cannot be deleted. Another method is to install a program that watches files and actively prevents them from being deleted based on a set of relevance criteria. There are also preservation systems that leverage existing network and laptop backup systems to keep daily incremental versions. Overall, There are certain IT efficiencies with in place preservation (e.g., IT no longer has to collect data in order to preserve it), however the risk of hardware failure and user mishaps increase the risk of spoliation. We explore the challenges of all of these systems where they relate to the different search architectures.

Benefits and Burdens of Selective vs. Enterprise Search

Knowing the search and preservation approaches on the menu, it is then possible to determine what is best for your organizations. We have now defined the different options and basic components available to meet corporate eDiscovery search and preservation requirements. Thus we can compare and contrast the potential benefits, burden and eDiscovery impact of selective and enterprise wide indexes. There is little doubt that some incarnation of enterprise wide indexing will dominate the infrastructures of public corporations eventually. The real question is which method will meet your current requirements and expectations at an appropriate Total Cost of Ownership(TCO).

Infrastructure Impact

Corporations understand that the TCO for software goes far beyond the license cost and infrastructure means more than hardware. Until fairly recently Legal departments either paid a premium to out-source this cost to service providers or they relied on employees to act as 'distributed, self-directing search agents' with the implicit risk of non-compliance. Now they are looking to search options to supplement or replace both for identification, preservation and collection discovery response. There are a number of factors to consider when measuring the impact on the corporate infrastructure:

Servers & Storage

Selective: Many of these platforms are available as hardware appliances that are meant to be quickly connected to the corporate network with optimized hardware and pre-installed software. Some partner with name brand storage providers to deliver the pre-installed software on standardized hardware that is compatible with the corporate server and storage standards. These systems tend to be optimized for burst performance to support a reactive, case-based workflow. Most of these systems utilize a collect and preserve collection model, which can consume additional storage. However, most also offer single instance storage (SIS) functionality to minimize duplications between matters on legal hold. This method is relatively easy to communicate to courts and can be considered a lower risk option.

Enterprise Wide: These systems must monitor and continuously update central indexes from thousands of different data sources. The architecture is more akin to adding another full communication platform considering the continuous bandwidth, processing and storage required to dynamically build a simple or advanced index of every new or modified piece of ESI. The advantages of a proactive enterprise wide index are self-evident, but the potential TCO can resemble an iceberg in the retrospective when you have indexed all your secondary ESI sources. Some providers have minimized this impact by leveraging the existing disaster recovery infrastructure instead of trying to index the live sources. This does introduce a potential one day lag into the system, but that may be acceptable in many matters.

Network Burden

Selective: On demand indexing or crawl search across a business unit or even a large file store has the potential to consume available bandwidth. Many applications have done a good job with dynamic throttling to manage this impact. Some systems (selective and enterprise wide) use local servlets or agents to index on the local desktops and then transmit back just the index updates to the central index. It should be apparent that any large scale indexing will have network impact. The required speed and volume that you have to index will dictate the potential network burden. Legal departments that invest in Identification processes and technology can limit the potential scope of preservation and collection. For example, having a well documented data map and sampling procedure can limit duplicative sources and peripheral custodians early in the matter.

Enterprise Wide: The initial indexing of ESI sources represents a serious, long term network burden as different classes of sources are put online. It’s impossible to simply flip a switch and build the enterprise index overnight. This is a long term investment for the corporation that requires careful planning for implementation. It will require engaging with IT, compliance and other stakeholders to create appropriate usage policies, change management systems and will affect all future technology purchases. Global corporations can face serious challenges with diverse networks, remote employees, limited bandwidth sites and other factors. You should not allow the promise of universal search to tempt you into impractical implementations. Your enterprise wide coverage may have known exceptions that must be filled with reactive processes.

Index Storage

Selective: We have already discussed the potential trade-off between index size and advanced search functionality. One advantage of the selective indexing is that the legal team may be able to adapt the level and size of index to the requirements of different matters. This does raise potential problems if you have already proactively indexed your email system or other priority sources, but selective indexing does give you more flexibility. Indexing only a targeted portion of your enterprise will result in less index storage than the index everything approach, and in some cases enables additional analysis techniques such as topic classification and concept search which are impractical to apply to the entirety of data across the enterprise.

Enterprise Wide: Consider that you may need to allocate 30-50% of your aggregate network and local storage for your central index. Some systems will try to reduce this index storage by federating search to specific data sources or using local indexes on mobile sources. This can present issues when the federated indexes have different functionality, as previously mentioned. The problem arises because typical business user requirements for desktop search are much simpler than potential discovery searches and most users do not want to allocate 20-50% of their desktop storage for an index. So whether the indexes are federated or centralized, the enterprise wide solution will have a greater overhead on relatively expensive storage.

IT Administration

The relative administration burden of applications varies tremendously, whether selective or enterprise wide. Therefore, we will focus on potentially different administrative roles for the two architectures.

Selective: Some of these appliances and applications are designed to be managed by litigation support personnel rather than a traditional IT department administrator. If you elect to permanently index priority data sources and your system continuously updates those indexes, then it may be more appropriate for IT to directly administrate it so that they can manage the network and system impact. Many corporations see selective indexing systems as a bridge technology to allow the legal department to carry out discovery requests and investigations without additional IT headcount.

Enterprise Wide: These systems generally require dedicated IT administration because of their scale, complexity and their integration into every primary system in the enterprise. Dynamic indexes (indexes being continually updated) have a reputation for instability and corruption. They are very sensitive to slow storage, update lag loops, virus scanning, back up systems and other environmental factors. A well provisioned environment that is actively monitored and administrated is the key to reliable search results and overall index health.

User Impact

It’s possible to leverage indexes of the desktop and network to support end user search, information analytics, compliance and other use-cases beyond eDiscovery search. However, it also carries the potential of availability and performance impact on users because of the local and networking burdens that we have discussed.

Selective: Most selective indexing systems are not designed to support end user search. They are usually controlled by the legal department and focus on priority and matter specific sources. They can be leveraged for compliance investigations and even information analytics on the priority systems, but the typical end user will not be using such tools. The initial indexing or crawl of desktops and laptops must be throttled or scheduled to minimize performance impact, especially if it is accompanied by a preservation collection. The demand on the local hard drive when copying large numbers of files can inadvertently bog down email, internet and other functions. We have seen frustrated users reboot laptops multiple times during remote collections or even just take the systems offline. It’s important to test and manage user impact or coordinate scheduled overnight collections with users. The preservation in-place methodologies seem to work well on central communication and archive systems. However, preservation of unstructured ESI on networks and mobile sources is fraught with issues and hidden risk. That is not to say that it cannot be done, just that it is very hard to do correctly.

Enterprise Wide: Several of enterprise wide search platforms also enable users to search their desktop files. Since most users would not leverage the higher search and organizational functionality of an advanced index, few companies are willing to dedicate 20-50% of the local storage for index storage. The index must be stored locally or the user would not be able to search unless connected to the network. There are some potential benefits to end users when enterprise indexing is integrated into retention system rules and filters, but that represents another substantial investment by the corporation. Other than the initial indexing impact on end users, ongoing incremental updates generally do not have a substantial impact on local performance. Some of the preservation in place strategies that rely on changing security rights or servlet file control have been reported to cause issues with users, active retention systems and back-up systems, but the main concern should be potential loss of ESI from corrupted drives, hardware loss or other factors in mobile sources.

ESI Sources

When considering indexing methods, it is important to examine how they apply to different types of ESI sources. Structured and semi-structured ESI sources have certain impact on organizations that is a bit different than purely unstructured. Structured ESI sources include financial systems and any other fundamentally database driven enterprise application. Although the raw field content can generally be indexed in place, the field and table relationships may render search and results essentially meaningless. Most corporations still run reports and exports through the structured applications to respond to discovery requests. Then there are semi-structured ESI sources like email and enterprise content management (ECM) systems that actually hold individual files, email and other discrete documents that can be indexed externally. Email within mailboxes could be considered a structured source, while corporate Exchange Journal could be considered an unstructured source. The challenge for hybrid sources is collecting these items with their context(s) intact.

Selective: Because structured ESI sources tend to be priority systems, corporate IT may be very hesitant to allow API calls or direct indexing that could impact data integrity or system performance. We are seeing selective indexing systems that can index the hybrid structured sources (Exchange, Sharepoint and ECM) successfully.

Enterprise Wide: This is one source where there is little difference between the two indexing options, other than the targeted sources. Some corporations have literally thousands of Sharepoint sites and so again we come down to how much do you want to index and at what cost.

Unstructured ESI sources have their own impact and burdens to consider. Unstructured ESI sources include network file shares, raw back-up tapes and any other repositories of loose files that are not managed by a central database. It is not as easy as it once was to draw a bright line between structured and unstructured sources. The main challenge with discovery search on unstructured content is resolution of custodial relationships and identity management across seriously heterogeneous sources. Every division may structure their folder shares differently and loose files may have generic Author metadata by default. Even the reliability of chronological fields can vary tremendously.

Selective: The challenge here is how to defensibly identify the right sources to index without being overly broad or unreasonably narrow. Many systems can inventory and sample index with great efficiency, but this requires personnel with the right skills and good process documentation to defend the relevance scope. The key to success here is investment in a standardized, documented work flow with the right people and technology features.

Enterprise Wide: A single search across all of your network file shares and hybrid structured sources is the key value proposition for corporations considering enterprise wide indexing. This proactive indexing moves the search in front of the Identification workflow, meaning that it can support the Identification process rather than be driven by the outcome of interviews, the data map and reactive inventories. This does not negate the need to invest in skilled personnel to craft investigative searches and a documented workflow, but it does give the corporation more options for scoping searches and other early data assessments.

Mobile Sources

Laptops, smart phones, USB storage and employee home computers represent the cutting edge of discovery requests. They also represent the real wild west for data controls, security and retention management. Intermittent connectivity, the risk of physical loss and the dynamic nature of comingled communications, media and documents pose serious challenges to any indexing platform. No search provider wants to be stuck supporting hundred of models and Smartphone OS versions. Luckily, the typically limited storage and relatively short lifespan of most of these sources keeps them from being relevant to some portion of matters.

Selective: Most of the available systems only tackle laptops as mobile sources. Many of them have done a good job of managing the potential user impact during a reactive indexing and heavy collection. There were early issues with some applications managing sources that would drop off the network suddenly, but we have not heard any recent horror stories from the field.

Enterprise Wide: Most of these systems have invested in integrations with mobile sources and their gateway servers. The gateway servers tend to keep recent/live messages cached and most can be used for ongoing preservation requirements. Overall, the enterprise wide systems have more maturity around mobile sources, but at a high potential index storage cost. As laptops continue to become the standard for an increasingly mobile corporate workforce, the sheer volume of individual user offline storage has skyrocketed. A corporation with 10,000 employees, half of whom have laptops with 500 GB drives, represents 2.5 TB that must be indexed onto high speed enterprise class storage, which is very different from your external 1 TB hard drive. A simple term index would range from 250-500 GB while and advanced index could consume 1-2+ TB.

Unified Index vs. Federated Search

We previous touched briefly on federated search, which is using another system’s search index instead of building your own for that ESI source. This is usually encountered in hybrid sources such as Exchange, Sharepoint or ECM systems. Both selective and enterprise systems may utilize federated search to minimize index size, overlap and add sources as easily as possible. The problem is that every search system has different fields, syntax, capabilities and limitations. A federated, heterogeneous search makes it very challenging to understand exactly what the true criteria and results are. Most systems rely on mapping fields and syntax so that the user submits criteria into the central graphical interface and the search system translates that criteria for each of the federated search systems. The problem is that this is effectively a black box function that happens behind the scenes and the results are not clearly segregated. Reliance on federated search places a large burden on the discovery team to understand all of the different systems, how criteria is converted and to potentially declare known exclusions or limitations. So it is important to understand how your selective or enterprise wide search engine tackles each different kind of ESI source before relying on it for discovery.

Index and Collection Lag

Traditional eDiscovery search applications and processing packages are usually offline while new collections are being indexed. Litigation support and legal personnel are used to being confident of the exact scope of collections being searched. If you add new ESI to your matter, then you need to update the index before your search, right? But when discussing tools for searching ESI where it lives on live corporate servers and desktops, we introduce a relatively new wrinkle into search – index lag. Enterprise and desktop search engines run in the background and watch for new, deleted or changed files within the scope of folders that they are watching. The problem is that index updates are never instantaneous, which means that enterprise wide searches are never 100% complete.

Let’s focus in on the potential gaps resulting from the lag between changes on the live system and the index. Anyone who has installed a desktop search application has seen their desktop bog down while the system crunches through all the default file locations. IT admins rolling out an enterprise search solution will almost always throttle the system back to minimize the user impact. Many desktop search engines are set to ‘Zero Impact’ by default. This means that as long as you are active on the machine, it is not indexing so that it will not slow you down. This is based on the business usage assumption that most of what you would be searching for is at least several days old and over a typical day there will be several times when you walk away from your machine. Those inactive periods should allow your index to stay reasonably current.

This assumption can lead to problems in specific discovery scenarios. Consider an example where you need to investigative searches on an executive. He catches wind of the investigation and deletes the critical files from his system after you have gotten hits on them. They are still in the index, but no longer on the desktop. The major problems arise when large volumes are moved, added or deleted on an active system. This temporarily ramps up the difference between index and reality. The index lag on server shares can be significant when they are set for daily updates or when the index is part of enterprise content management, archiving or backup systems.

A reasonable portion of legal matters are only concerned with a specific historical time period. That actually raises an interesting scenario based on the automatic expiry/deletion of files that are beyond their retention period. Most archives and content management systems run daily checks and delete these items immediately. However, it can take time for the index to be updated. If you think that this is not an issue, consider this typical discovery search scenario. An attorney emails a list of custodians and search terms to their litsupport tech. The tech creates a matter and runs the test search. The number of hits is reported and the attorney goes off to confer with outside counsel and even potentially with opposing counsel. It could be days or weeks before they come back with the green light to retrieve the files. Unless the system has automatically preserved those hits, it is probable that some items will be deleted every day if the company has any kind of systematic retention enforcement technology running across all the data sources. Now consider the effort required to manually verify that this is what has happened to thousands of files that show up as hits, but can no longer be retrieved or viewed. This is a different issue from the ‘index lag’, as rerunning the search will eliminate the false hits, but you know that they were there when you first checked and your audit log will show them. Worse, you may have reported those numbers to the other side during negotiations and they may well check the total of your produced, non-relevant and privileged counts. This is more of a preservation issue than an index problem, but it is worth considering.

The last index lag point is a bit technical. When you consider the huge scale of enterprise search, it makes sense that developers will try to minimize the footprint of the most expensive system components like the servers and database storage. This means that many systems keep only pointers in their database and most everything else in the full text index. It makes for lower overhead, but means that the index has to be updated every time a retention category, metadata property, categorization tag, virtual folder or other indexed attribute is altered. Users could decide to clean out a collection of family photos or at least mark them ‘Personal’, but they would still show up in search results as business for a period of time. This really heats up in a review scenario where multiple reviewers are tagging, flagging and commenting on a set of items. There is nothing that drives a reviewer crazy like marking an item privileged and then not finding it when you pull up all your privilege items for the second pass.

Selective: Most reactive indexing systems have workflow built for rapid search and preservation collections. The same index lag issues apply to selective and enterprise wide systems, but discovery teams tend to compress the search and collection timeline when they have had to select sources and provide more specific relevance criteria. In some ways, it is easier to declare a set of target sources that was indexed and searched on a specific date than it is to sign an affidavit that all systems were searched successfully with completely up to date indexes across the enterprise. Selective index systems tend to run incremental indexing and searches on a scheduled basis rather than the continuous update method favored by enterprise wide engines. This means that you must complete an incremental update if you want to include new items created since the last index update. It also means that you may miss items that were created and then destroyed in between indexing runs unless there is some kind of collect-and-preserve system, servlet or other mechanisms keeping versions.

Enterprise Wide: Because these systems offer ‘instant’ search, there is a tendency for non-technical users to assume that these searches are complete and up to date. In any system of sufficient size, it is almost certain that indexes will be being rebuilt, offline, updating or otherwise unavailable when you run an all index search. Unfortunately, many systems bury potential errors or system messages in administrative error logs rather than exposing them to the user. Counsel does not like to hear that nothing is perfect when you exceed a certain size. This is a very subtle issue that requires a fairly sophisticated technician to differentiate a real error from a typical system message. It is even more challenging to troubleshoot ghost hits (search results on items already deleted) or test for false negatives (items that are not in the index).

Geographic and Connectivity Challenges

The global nature of corporations can pose serious challenges to selective and enterprise wide indexing. We will at least highlight some of these for consideration, although they impact both systems.

- Remote locations and employees with poor bandwidth connections can make network based indexing impractical or impossible. Early remote collection and forensic imaging systems found this out the hard way and most now offer stand alone systems that can be shipped and run without a specialist.
- Time changes between the discovery team and the sources can actually affect search results. If a New York litigation support analyst runs a single day search on the files on the Tokyo trading floor, as much as 50% of the files might be missed. Very few search applications resolve chronological search criteria down to the hour level.
- Data privacy laws in the European Union and other countries may make remote search of employee ESI illegal without appropriate mechanisms to ensure that employee's can screen out personal communications and files.
- Many systems have issues with foreign language character sets and some can only load one language dictionary at a time.

What to Consider For Your Search Requirements

We have defined and explored selective and enterprise wide indexing from a variety of perspectives so far. Before you choose any technology, it is worth the time to realistically define the scope and parameters of your discovery requests past, present and potential. Advanced search analytics can be leveraged in every stage of your eDiscovery lifecycle, but we have not seen many requesting parties agree to search criteria based on technology that they cannot render into their familiar Boolean syntax. You could call technology without the appropriate staff and defined process either 'shelfware', i.e. software that never gets used, or the proverbial 'easy button', i.e. black box software that the user operates without any understanding. Look at your current or recent matters and ask yourself what kind of impact different search features would have made. As an example, deduplication may reduce the volume of collections and review, but if you have to produce to the SEC's custodian based protocol, you had better be able to repopulate search results. Do you always break things down by custodians or are you handling a large number of IP matters based on specific lists of terms?

It's important to beware of what we call information overload. Traditional discovery brought counsel most documents with a known context. They generally knew what and why they were looking at based on interviews. Most counsel dreaded having to perform discovery on warehouses full of paper records in boxes or filing cabinets in mothballed facilities. They lost the ‘human context’. An enterprise wide index can resemble that mythical warehouse in the Raiders of the Lost Ark. Instant search can become an instant headache unless you are able to put those half million results back into context of custodian, source, chronology and more. The fundamental issue is information overload. It is easier to understand search context when you start with a core set of priority sources and then add indexes based on potential relevance. If you find yourself limiting your enterprise wide searches to specific sources to eliminate non-relevant hits, then you should ask yourself if you needed everything indexed in the first place. The graphical interfaces of search systems are rapidly evolving to support faceted navigation and dynamic filters. These can enable you to quickly narrow your overly broad results. Of course, these features have a familiar price; increased index size and processing.

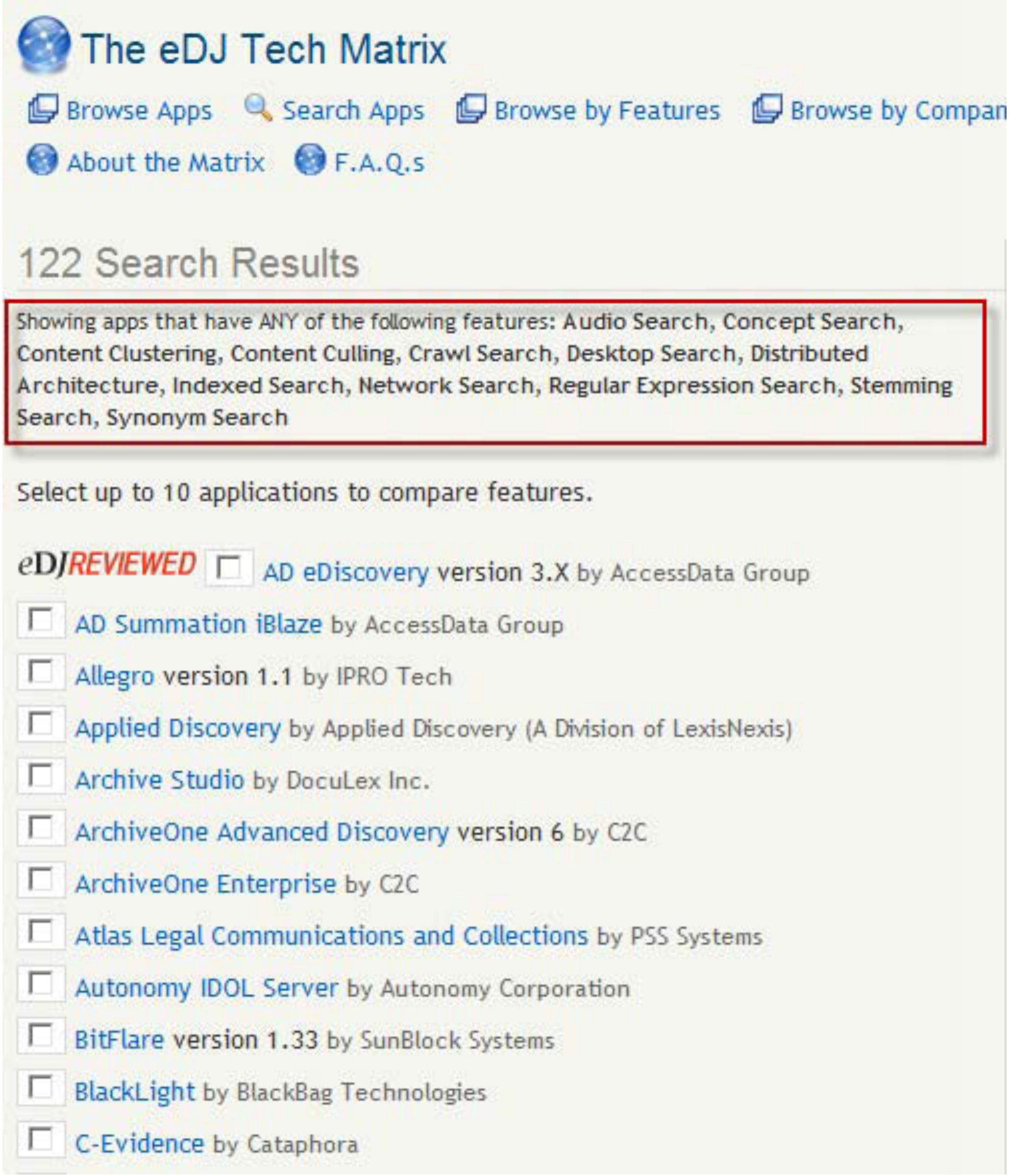
Conclusion

Search is the common thread that runs through almost every stage of the eDiscovery lifecycle. Corporations seek to transform their reactive, out-sourced eDiscovery fire drills into a standardized, managed business process based on search and collection. When exploring the suitability of enterprise wide or selective search platforms, start by clearly defining your resources and requirements. Do not expect new technology to reinvent your discovery lifecycle overnight.

We have discussed many of the potential benefits and costs that each indexing and collection architecture may bring to your infrastructure. Is there a corporate mandate to invest in the value proposition of information governance? Is the primary driver discovery cost control? Is the corporation prepared for the servers, storage, network traffic and administration required to realize the promise of enterprise wide search? There are corporations with the compliance and discovery profile that may justify such a serious investment. One ‘bet the company’ case can be the impetus to consider both options, though you should factor in realistic indexing times in months. Be honest about your ability to your needs, resources and goals. Walk before you try to run.

Appendix A: Researching Applications

There are literally hundreds of products on the market that offer some enterprise search functionality. Not all are actually suitable for either selective or enterprise wide search, but it can be very challenging to figure out who should be considered. The eDJ Tech Matrix is one of our free research tools that can help in this process. In the image below, I browsed through all the available features and retrieved all application listings that contained any of them. As you can see, there are 200 applications (and growing every day) with one or more features, which is far too many to digest and there is a lot of feature overlap with processing and review applications.



122 Search Results

Showing apps that have ANY of the following features: Audio Search, Concept Search, Content Clustering, Content Culling, Crawl Search, Desktop Search, Distributed Architecture, Indexed Search, Network Search, Regular Expression Search, Stemming Search, Synonym Search

Select up to 10 applications to compare features.

eDJREVIEWED □ AD eDiscovery version 3.X by AccessData Group

□ AD Summation iBlaze by AccessData Group

□ Allegro version 1.1 by IPRO Tech

□ Applied Discovery by Applied Discovery (A Division of LexisNexis)

□ Archive Studio by DocuLex Inc.

□ ArchiveOne Advanced Discovery version 6 by C2C

□ ArchiveOne Enterprise by C2C

□ Atlas Legal Communications and Collections by PSS Systems

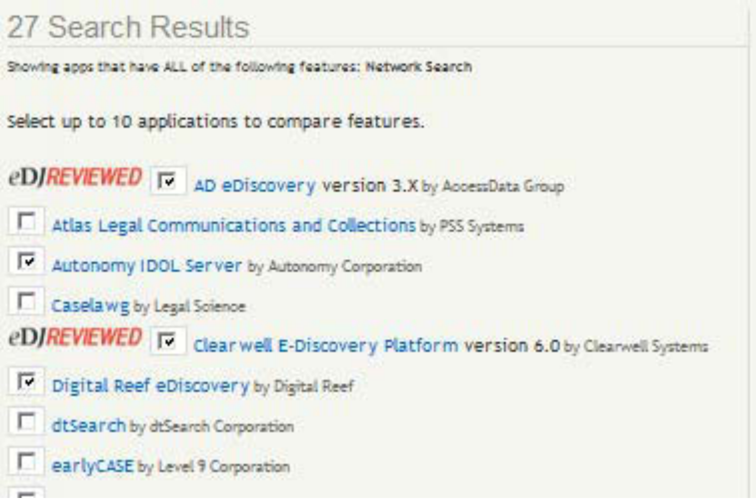
□ Autonomy IDOL Server by Autonomy Corporation

□ BitFlare version 1.33 by SunBlock Systems

□ BlackLight by BlackBag Technologies

□ C-Evidence by Cataphora

The fundamental feature that defines enterprise search is the ability to search across large network file shares, called “Network Search” in the eDJ Tech Matrix. If I ‘Browse by Features’ and select Network Search feature, we get 27 (as of today) applications.



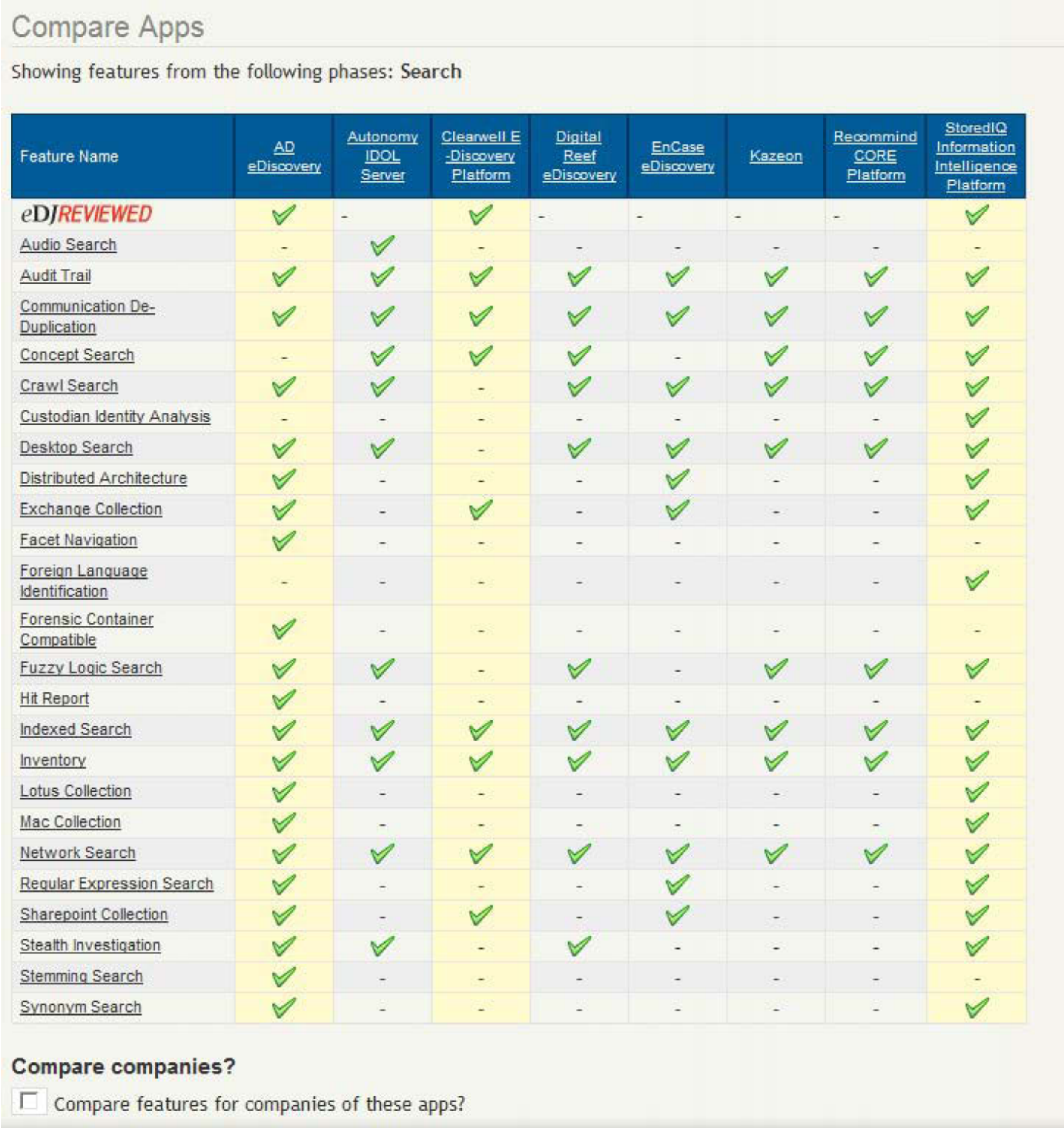
27 Search Results

Showing apps that have ALL of the following features: Network Search

Select up to 10 applications to compare features.

eDJREVIEWED ☑ AD eDiscovery version 3.X by AccessData Group
☐ Atlas Legal Communications and Collections by PSS Systems
☑ Autonomy IDOL Server by Autonomy Corporation
☐ Caselawg by Legal Science
eDJREVIEWED ☑ Clearwell E-Discovery Platform version 6.0 by Clearwell Systems
☑ Digital Reef eDiscovery by Digital Reef
☐ dtSearch by dtSearch Corporation
☐ earlyCASE by Level 9 Corporation

We can next select up to 10 applications at a time to generate a comparison matrix and restrict the comparison features to those associated with the Search category. The image below is an example of what your results could look like. Remember that the eDJ Tech Matrix and the market are incredibly dynamic, with new applications, features, versions and more being released every day. The basic listings in the eDJ Tech Matrix are free for providers or users to create and maintain. We work hard to keep the listing up to date, but only the listings marked eDJ Reviewed have been actively checked with a full product demonstration. The eDJ Tech Matrix gives you a no-spin place to start your research, but it is meant primarily to help untangle the confusion new, overlapping products, terminology and buzz words.



Compare Apps

Showing features from the following phases: **Search**

Feature Name	AD eDiscovery	Autonomy IDOL Server	Clearwell E-Discovery Platform	Digital Reef eDiscovery	EnCase eDiscovery	Kazeon	Recommind CORE Platform	StoredIQ Information Intelligence Platform
eDJREVIEWED	✓	-	✓	-	-	-	-	✓
Audio Search	-	✓	-	-	-	-	-	-
Audit Trail	✓	✓	✓	✓	✓	✓	✓	✓
Communication De-Duplication	✓	✓	✓	✓	✓	✓	✓	✓
Concept Search	-	✓	✓	✓	-	✓	✓	✓
Crawl Search	✓	✓	-	✓	✓	✓	✓	✓
Custodian Identity Analysis	-	-	-	-	-	-	-	✓
Desktop Search	✓	✓	-	✓	✓	✓	✓	✓
Distributed Architecture	✓	-	-	-	✓	-	-	✓
Exchange Collection	✓	-	✓	-	✓	-	-	✓
Facet Navigation	✓	-	-	-	-	-	-	-
Foreign Language Identification	-	-	-	-	-	-	-	✓
Forensic Container Compatible	✓	-	-	-	-	-	-	-
Fuzzy Logic Search	✓	✓	-	✓	-	✓	✓	✓
Hit Report	✓	-	-	-	-	-	-	-
Indexed Search	✓	✓	✓	✓	✓	✓	✓	✓
Inventory	✓	✓	✓	✓	✓	✓	✓	✓
Lotus Collection	✓	-	-	-	-	-	-	✓
Mac Collection	✓	-	-	-	-	-	-	✓
Network Search	✓	✓	✓	✓	✓	✓	✓	✓
Regular Expression Search	✓	-	-	-	✓	-	-	✓
Sharepoint Collection	✓	-	✓	-	✓	-	-	✓
Stealth Investigation	✓	✓	-	✓	-	-	-	✓
Stemming Search	✓	-	-	-	-	-	-	-
Synonym Search	✓	-	-	-	-	-	-	✓

Compare companies?

□ Compare features for companies of these apps?

Appendix B: Search Features Requirements

Below is a list of the primary search features that will help you understand some of the differentiators between products. There are many more features/functions that may be important in understanding what file types (unpacking Zip files) and post search actions that the application can support. Remember to check the active eDJ Tech Matrix for updates and know that you need to actually see how each application has implemented a specific functionality. There are many different ways to perform Crawl Search across desktops or network shares. These features will help you build your initial requirements list for your round of Request for Information responses.

ESI Sources:

- Cloud Collection
  - Ability to collect and preserve ESI from the intra or internet.
- Desktop Search
  - Search across live network to remote user desktop and laptop devices
- Exchange Collection
  - Ability to connect to and collect ESI/email from a live Microsoft Exchange server or an EDB file.
- Forensic Container Compatible
  - Ability to search, extract, process and produce to and from forensic image containers.
- Lotus Collection
  - Ability to access and collect ESI from Lotus Domino and NSF files
- Mac Collection
  - Ability to acquire ESI from Apple Mac computers and servers within the native OS
- Network Search
  - Search across live enterprise network data sources
- Sharepoint Collection
  - Ability to collect files and content from Microsoft Sharepoint.

ESI Types:

- Audio Search
  - Ability to search by conversion to text or direct phonetic search of audio/video format ESI
- Foreign Language Identification
  - Identify and flag ESI that contains foreign languages and characters
- Foreign Language Support
  - Application is capable of processing or searching foreign languages and uncommon character sets such as double-byte characters.

Search Features:

- Concept Search
  - Extraction of concepts based on Latent Semantic Indexing or similar complex frequency analysis.
- Crawl Search
  - Use of GREP or other search that actively checks files without creation of index. This is a point in time search.
- ESI Sampling
  - System for random sampling that can be used to establish ESI characteristics or in quality control or assurance.
- Facet Navigation
  - Ability to navigate, filter or build searches based on dynamically populated ESI property facets such as chronology, custodians, sources and file types
- Fuzzy Logic Search
  - Partial word recognition to compensate for OCR and spelling issues
- Hit Report
  - Report of the number of hits and other properties associated to individual search terms on a single search.
- Indexed Search
  - Creation of an index for search
- Inventory
  - Generates a file inventory or catalogue of the targeted data sources
- Regular Expression Search
  - A search syntax that functions with indexed and crawl search.
- Stealth Investigation
  - Ability to search, browse and collect from custodians's data sources without alerting target custodian or impacting their system adversely.
- Stemming Search
  - Ability to expand search terms based on common stemming rules
- Synonym Search
  - Ability to view synonyms and expand search criteria based on user choice. This is also called White Box criteria expansion