

2009

BizTechReports

Editorial Director:
Lane F. Cooper

EDISCOVERY: SEARCH NOT FOUND

A revealing look at discoverable and potentially crucial text that is too often neglected from the eDiscovery process

eDiscovery: Search Not Found

A revealing look at discoverable and potentially crucial text that is too often neglected from the eDiscovery process

Abstract:

This white paper explores the impact that “embedded images” of text can have on the enterprise and corporate legal communities. Embedded images of text are images of words in documents that are subject to discovery action. Often documents with these images are improperly assessed or missed entirely when traditional eDiscovery tools and procedures are employed. The report provides insights and recommendations on how the affected community of interest (legal counsel, risk managers, compliance officers and technology executives) can address the challenge of embedded images of text in the constantly evolving eDiscovery landscape.

Background:

The role of eDiscovery – the process of applying information technology to search and retrieve electronic documents in response to legal or regulatory actions – is growing in importance as both the volume of information kept by corporation explodes, and the complexity of data types expands. Between structured database files on one hand – and unstructured e-mails, word documents, spreadsheets, presentation decks and multi-media web pages on the other – corporate attorneys are in a perennial race to keep up with technologies capable of identifying and producing discoverable materials in a compliant matter. As a result, researchers at the Radicati Group predict that corporate spending on services and solutions to address this complexity will exceed \$2 billion by 2012.¹

Most executives consider these investments money well spent. Failure to perform discovery actions in a compliant manner can yield disastrous results for organizations and the legal community that represents them. According to analysts at Gartner:

“The trends [driving]...eDiscovery...mean that enterprises and their lawyers have to meet higher expectations from the bench and society...If lawyers do not become technically competent, they are not doing their jobs properly. If [a] judge were to so admonish a lawyer from the bench, it might carry the weight of his office, and lawyers would ignore it only at their peril.”²

In other words, failure to stay current with technology is not a professionally effective position to take when it comes to eDiscovery.

In this report we focus on a new variable in the eDiscovery environment that must be addressed by legal counsel, risk managers, compliance officers and technology executives. Documents today are more complex, often featuring layers of content that may include “images” of words that are relevant to a discovery action.

In the months and years to come, organizations and the legal community run a significant risk of falling out of compliance with discovery actions if issues associated with embedded images of text are not managed appropriately.

¹ "eDiscovery Solutions - Market Quadrant 2008"

http://pr-usa.net/index.php?option=com_content&task=view&id=148088&Itemid=32

² "LegalTech 2009 Shows Higher Bar for E-Discovery"

http://www.gartner.com/DisplayDocument?id=882121&ref=g_sitelink

Embedded Images of Text in Today's Complex Document Environment

There may have been a time when simple, straightforward text documents accounted for the lion's share of content in the filing systems of organizations. But that certainly is not the case anymore.

Document applications today – including spreadsheets, word processors, presentation programs, portable document formats, etc. – increasingly contain a variety of content types within a single file. The reason for this elevated level of complexity of documents is explained by the fact that it simplifies the ability to present sophisticated concepts to readers.

Mature document standards allow practically any knowledge worker to integrate information from different content applications into a common document. They can easily manipulate text, images, original art, and web page “captures” to describe new products, explain complicated procedures, or outline contractual obligations.

“These are precisely the types of documents that must be reviewed when legal differences of opinion are to be resolved in court, or when compliance with regulations must be determined by appropriate authorities.” – Andy Wilson, Founder and CEO of Washington DC-based Logik, a provider of enhanced eDiscovery services to the corporate and legal community.

The problem from an eDiscovery perspective is that when information is taken out of its native application and inserted into a non-native environment, it often loses the characteristics that allow the content to be searched. For example, content is very likely not “searchable” if a worker:

- Cuts and pastes the screenshot of a website into a presentation slide; or
- Captures a bar chart created in a spreadsheet and inserts it into a word processing application.

“In these cases, the screenshot and the bar chart are considered embedded images of text. The presence of text in image form represents a major area of concern because most current eDiscovery programs and procedures do not convert images to text and then search them.” – Sheng Yang, Chief Technology Officer, Logik.

...Search Gaps Can Let Hot Keywords Slip Through Cracks

Two major operational issues arise with embedded images of text. The first revolves around the presumed importance of the embedded images.

If a knowledge worker makes the effort to insert information from another application into a document, it can logically be concluded that the non-native content is highly relevant to the document. However, if the embedded image contains a “hot keyword” – a term that was identified as a high-priority in the discovery action – then it is possible that the document, and the highly-relevant image embedded within it, would be missed.

The second issue stems from the fact that compound documents are becoming increasingly common – perhaps even routine. If this trend continues, and more corporate documents feature

the presence of embedded images of text, then it is likely that existing eDiscovery procedures will allow a growing percentage of relevant content to slip through the cracks.

Both of these scenarios expose organizations and legal counsel to significant risks. For instance, embedded images could be used to cloak fraudulent activities by hiding content in a way that prevents key documents from emerging in a discovery action. Spammers, for instance, use this exact technique to defeat anti-spam measures that use keywords to filter out undesired e-mail correspondence. By embedding text in a non-searchable .GIF image file, the message avoids detection. A similar tactic could be used by rogue workers in an enterprise who want evade discovery procedures.

More than likely, however, litigants will simply be victims of unintended omission – or inclusions – as key documents simply go undetected in the wake of traditional eDiscovery procedures. For instance, there have been cases when eDiscovery processes inadvertently produce and disseminate privileged documents; courts have then found that privilege was waived by the production of these documents.³

...Impact on Affected Parties

The effect of incomplete eDiscovery on all members of the community of interest is significant:

- **Law Firms, Corporate Counsel and Compliance Officers** – The objective of eDiscovery is to automate and streamline attorney involvement in determining documents that should be flagged for review. While significant progress has been made using “keyword” and “concept” searches to identify documents that should be reviewed by attorneys, growing awareness of embedded images – which are not captured by traditional eDiscovery applications – means that more documents must be manually reviewed by expensive legal resources. This can be a daunting, time consuming and expensive proposition. Worse still, missing key documents for disclosure because existing procedures did not capture embedded images can be bad news in the course of a multi-million dollar case. This is especially true if judges or opposing parties determine – and are able to prove – that discovery deliverables and practices were inadequate.
- **C-Level Executives** – Missing key documents can be embarrassing and detrimental to any organization. But it can have a corrosive impact on the reputation and credibility of top executives managing the organizations. This is particularly true when allegations of obfuscation make it to the media.
- **Technology Executives** – Before a discovery action is initiated, data resources typically fall under the stewardship of CIOs and other senior technology managers. The challenge of embedded images adds yet another dimension to the data and content management strategies that must be implemented by the IT staff. The issue of how to address hidden text will most dramatically affect this group as they evaluate new ways of indexing content for later production. If documents are not properly processed with embedded images of text in mind, then documents will be missed.

³ Washington Adopts Test for Determining Waiver by Inadvertent Disclosure, Finds Attorney-Client Privilege Waived. <http://www.ediscoverylaw.com/2008/12/articles/case-summaries/washington-adopts-test-for-determining-waiver-by-inadvertent-disclosure-finds-attorneyclient-privilege-waived/>

- **Judges and the Courts** – In overseeing legal proceedings, judges are interested in preserving the integrity of their cases. Consequently, the courts are being increasingly proactive in demanding that all parties involved utilize defensible processes and tools to demonstrate good-faith efforts to address the full range of complexities involved with eDiscovery. As courts become more aware of the pervasive presence of hidden text, they will require counsel to find the technologies and implement procedures to properly address the challenge.

...Elevating the role of OCR and Smart Indexing in eDiscovery

The growing realization that a significant amount of critical and/or relevant text is currently hidden as images, naturally introduces a higher role for optical character recognition (OCR) capabilities in the eDiscovery process.

“That is why organizations must begin re-evaluating the entire life-cycle of current eDiscovery practices with OCR requirements in mind.” – Xavier Mouligneau, Senior Software Engineer, Logik.

Not only must leaders determine how images should be addressed for conversion during the eDiscovery process itself (in response to specific requests for information) but they should also explore what can be done prior to requests to prepare a streamlined process for:

- Scanning images within documents and files cost-effectively;
- Identifying relevant images in an automated manner;
- Converting images to text in an efficient manner; and
- Indexing the documents appropriately (with flags that alert reviewers of relevant embedded images)

...Developing a Financial Justification Model for OCR-enhanced eDiscovery

There is no getting around the fact that adding an OCR-enhanced capability to current eDiscovery procedures will contribute to the time it takes to go through an eDiscovery process, as well as the cost associated with deploying the technologies and procedures to identify, capture and convert images of potentially relevant text.

Nevertheless, it is still possible to cost-justify the investment of both time and money in a fairly straightforward manner.

- **Addressing Hidden Risks** – The absence of OCR capabilities in current or legacy eDiscovery processes may reduce the cost of the solution, but it leaves an important risk element unknown and therefore unmitigated. It can make sense to spend \$1 million on software and services to save the enterprise \$4 million in risk.
- **OCR-based Automation Optimizes Legal Resources for Asset Review** – Just like traditional eDiscovery solutions, OCR-enhanced services helps to cap the number of documents that lawyers would otherwise be forced to initially review during the triage portion of the discovery process. This can have a direct impact on the bottom line. Specific cost savings can be determined by engaging in a statistical sampling of the target base of documents and then developing a cost justification model. This should be

done by associating the time and effort needed to manually find and review document elements that cannot be found by using basic “keyword” or “concept” search solutions.

- Hence, a lawyer that reviews 10,000 documents for hidden content at a billable rate of \$400/hour may be expected to review 100 documents per hour. This results in an overall cost of \$40,000.
- Using OCR-enhanced eDiscovery, hidden content can be easily tagged and then searched. As a result, the lawyer could find just the documents needed (say 50 percent of the total) within a few seconds and then review just those documents at \$400/hour.

...Conclusion

There are currently few, if any, organizations that are not already exposed to the risks of missing important documents because files contain images of “hot keywords.” Embedded images of text are an increasingly common by product of how workers do business today. Moreover, as technology gets easier to use, and employees mix and match document types into a single document, eDiscovery processes must address multiple layers of embedded images of text.

For instance, a common spreadsheet (XLS) can itself contain an embedded presentation file (PPT) and the presentation file can contain multiple embedded images that need to go through an OCR. Consequently, eDiscovery software not only needs to detect embedded images, but it also needs to dive as deep as possible into the file and extract all embedded content and metadata.

This OCR function should be seen as a critical requirement because embedded images of text that are subject to rules of disclosure – such as the Sarbanes-Oxley Act, as well as the recent amendments to the Federal Rules of Civil Procedure (FRCP) – will therefore have a dramatic impact on the evolution of eDiscovery projects, products and procedures.

The good news is that the technologies that address the search gaps created by embedded images of text are rapidly evolving. They include at their core the ability to identify documents that feature embedded images of text, scan them and execute an OCR protocol to create machine readable text. The ensuing output can then be indexed and flagged for further review by attorneys involved in the discovery process.

Over the next 12 to 18 months, we can therefore expect:

- Plaintiff attorneys to view enhanced eDiscovery as an opportunity to ensure that a comprehensive and accurate review of all relevant documents is conducted;
- Defending legal counsel to use enhanced eDiscovery as a way to shield corporate clients from any allegations of lax compliance;
- The courts to demand participants undertake all reasonable measures to protect the fairness and integrity of legal proceedings by ensuring that a complete record or relevant documents is created.

Until now, the risks associated with embedded images of text have been hidden because the focus of eDiscovery has revolved around the capture of readily identifiable text elements and symbols. However, as awareness of these search gaps rise, disregarding this area of risk becomes increasingly untenable.

###

**Enhanced eDiscovery Vendor Spotlight:
Logik**

The mission of Washington, DC-based Logik is to shrink and simplify the eDiscovery process. Logik achieves this by applying technology, expertise and proven best practices to remove as much irrelevant data as possible. Logik delivers a high quality and clean product that enables clients to assemble a complete and cogent body of evidence for their clients' cases. Often, this means that the Logik team must sift through terabytes of data, using cutting edge search, aggregation and processing technologies developed by Logik.

Gridlogik, for instance, is a technology developed by Logik that identifies duplicate documents, extracts rich text, identifies the languages in a document, searches for keywords, extracts hidden documents, parses metadata, converts documents to static images and then prepares it for various levels of analysis.

The processed data is typically loaded into in-house or web-hosted document review software applications for review by attorneys and content experts. The data is then handed over to the party requesting it, which can be a government agency, a law firm or another corporation.

Logik works with fortune 500 clients as well as many of the top 100 law firms in the country to solve their most complex eDiscovery problems.

For more information visit: www.logik.com