

OCR SOLUTIONS

Although the majority of today's electronic discovery matters focus on email and electronic communications, paper productions continue to be a significant component of Discovery projects. As a result, a key challenge is the search and review of large datasets including OCR processed documents.

OCR: IMPACT ON SEARCH TERMS & SPEED

For both Boolean or keyword search, OCR creates a series of challenges due to the inherently poor quality and inaccuracy of the extracted text. In addition, OCR can dramatically impact the search speeds and scale issues of many current systems due to the extremely large number of terms created during the OCR process. For example, a repository that contains only ESI data is likely to have approximately 750,000 unique, searchable terms. The same repository with OCR images may contain more than 9,000,000 unique terms: words like DISH are recognized as DISH, D1SH, DLSH, and so on. With so many additional terms, SQL-based platforms are limited in scale, search speed, and system responsiveness. Moreover, standard Boolean queries performed on OCR often require a significant amount wildcard searching, which frequently returns results that are overly inclusive and much broader than desired.

INFERENCE FOR OCR: UNMATCHED ANALYTIC SEARCH CAPABILITIES

Inference is one of the few electronic discovery and review applications that can perform baseline Boolean keyword search as well as advanced conceptual search. By combining traditional review methods with concept analytics, Inference can dramatically increase search result accuracy on large OCR data collections. Using Inference's analytics suite, users can focus their attention on the issues of the matter by using concepts derivative of the data to target specific documents and key areas, including OCR datasets, from across the entire data population. Additionally, Inference can accept any available metadata fields to further enhance search capabilities that are extracted electronically or manually coded (i.e., bibliographic coding fields).

SPEED & SCALE VIA DISTRIBUTED ARCHITECTURE

Inference's infrastructure and architecture provide unmatched scalability and system speed to search across millions of terms generated by the OCR process. Through Inference's distributed, load balanced environment, Inference maintains query responsiveness for even the largest datasets. Several of Inference's largest projects have involved repositories containing a majority of OCR data and were migrated from standard SQL-based systems where query speed, search accuracy and performance were inadequate due to the complexities and the number of terms in the large OCR data populations.