



Condensing Multiple Collections: A Case Study in the Use of Equivio Technology

Introduction

This case study describes the use by H&A Forensics of Equivio near-duplicate detection technology in a matter encompassing multiple document collections. Use of the Equivio technology facilitated the reduction of the ultimate document set by more than 25%.

Background

The matter involved a single plaintiff and multiple defendants from the same organization. Each defendant was represented by different counsel, each of whom undertook to independently compile their affidavit of documents. The documents consisted of both ESI and paper.

All of the documents were produced as TIFF images and load files of varying format. Some productions included single-page TIFFs with corresponding single-page OCR text files, while other productions were composed of multi-page TIFF images, some with and some without OCR text files. Each production also contained at least one data file, although these were also in a variety of formats, and there was no consistency in the fields produced or their names.

Given that all the documents were from individuals within the same organization, it was expected that many documents would be duplicated between production sets. H&A was retained to consolidate and de-duplicate the productions into a single production, along with the appropriate data and load files so that the information could be imported into CT Summation.



Process

Although identifying duplicate ESI documents is a fairly simple process, involving the comparison of MD5 hash codes, the configuration of the data in this case precluded this method, as the documents were presented as TIFF images, many of which originated from scanned paper documents. A cursory review of the TIFF images and associated OCR text files (where available) indicated that different scanning software was used on the various productions, resulting in TIFF images at different resolutions, and varying quality of OCR text output. A simple MD5 Hash comparison of the TIFF images or OCR text files would not have identified any duplicates.

It was decided to employ Equivio to identify near-duplicates within the entire population of documents. The threshold for the definition of a near-duplicate would be set such that the identified documents would, for all intents and purposes, be considered identical.

As the document comparison would be based on the contents of the documents as opposed to their TIFF images, a single OCR text file of each document was required. Using a variety of scripts and tools, the single-page OCR text files were combined into single files per document, and missing OCR text files were produced from their TIFF images.

All of the OCR text files were processed through Equivio, which identified unique documents and their similar counterparts. The results of this process were:

Total documents	38,673
Unique documents without any "Near-Duplicates"	23,812
Unique documents with "Near-Duplicates"	3,588
Total unique documents	27,400
Number of duplicate documents	11,273

The 27,400 unique documents were extracted from the original population. These were grouped according to custodian, and the appropriate DII load files were created to import the TIFF images and OCR text summaries into Summation. The data files were normalized and combined into a single data load file.



Conclusion

Although standard MD5 Hash comparison was not appropriate for this matter, the near-duplicate detection afforded by Equivio enabled H&A to identify duplicates within and between the production sets. This resulted in a reduction of the document population by approximately 29%.

About Equivio

Equivio enables the management of data redundancy in content-centric business processes. Equivio's technology zooms in on unique data, allowing you to read less, think more, win big™.

With products for grouping near-duplicates, capturing email threads and automating prioritization, Equivio powers a broad range of business applications, including e-discovery, records management, email archiving, data retention and intelligence. To learn more about winning with Equivio, email info@equivio.com or visit www.equivio.com.

Equivio Inc.
5260G Nicholson Lane, Suite 150
Kensington, MD 20895
USA

www.equivio.com