CASE STUDY

# Legal Harmony Across Datasets:

## Helping Prosecution and Defense Document Sets Come Together

## SITUATION

Two AmLaw 100 firms, well-known in their fields for commercial litigation practice, needed a solution that would marry and reconcile two large sets of data, one from each the prosecution side, one from the defense.

The goal was to merge them into one database, with a single copy of each file (no duplicates, but no missing files either) and end up with something easily usable, portable and duplicative for the two legal teams involved.

## CHALLENGE

Valora needed to merge these two giant datasets requiring a significant, non-Hashing de-duping process. Also, the datasets were previously processed and stored in two very different systems, which resulted in data sets that were entirely inconsistent with each other. Almost all of the documents were missing their native files.

## SOLUTION

Valora solved these challenges by using PowerHouse to create a consistent and high performance OCR process over 1.6 million pages to ensure a consistent level of quality across the files. Next, PowerHouse created a rich set of applicable metadata fields to help tag each file with key attributes. Multiple levels of near-duplicates were identified by a text and attribute comparison of the files' newly-generated text and rich metadata.

## RESULTS

Valora issued a mid-project report indicating about 50% duplication in the populations, accounting for only half of the total set. Ultimately, the documents from both sides were successfully de-duped and merged, resulting in 70,000 unique documents and over 950,000 pages.

Once the whole process was complete, one of the firms in question was so thrilled by the results that they have since standardized on Valora's techniques for all dual data sets received in IP litigation.

### SOLUTIONS APPLIED:

- AutoClassification
- Document Analytics
- Electronic File Processing
- OCR & Text Extraction
- Analytics & Data Mining

### PRODUCTS USED:

- PowerHouse