Legal Harmony Across Datasets

Helping Prosecution and Defense Document Sets Come Together

It is not often that the prosecution team and the defense team find something to agree on in the middle of a large case. 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 side. 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. The catch? No file metadata to speak of, and no Hash de-duping – both sets were independently created in differing PDF file format. That's when Valora Technologies came onboard.

Valora immediately recognized several challenges:

- Merging two giant datasets required a significant, non-Hashing de-duping process
- The datasets were previously processed and stored in two very different systems. This resulted in data sets that were entirely inconsistent with each other, even though many of the files were actually the same. Almost all were missing their native files.

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.

The biggest surprise of the project happened when Valora issued a mid-project report indicating about 50% duplication in the populations, accounting for only half of the total set. As both firms thought they were working off the same dataset as opposing counsel, you can imagine the dialog that ensued!

The Hero of this Case Study: PowerHouse

PowerHouse is a fully functional, fully featured automated services delivery platform for eDiscovery, Records & Information Management and Information Governance. PowerHouse natively supports the following services:

- Autocode/AutoTag = Rich Metadata Creation
- AutoUnitization
- AutoTranslation for 105 languages
- AutoRedaction of sensitive info
- AutoReview (like TAR, but even more automated)
- AutoClassification for tagging and disposition of files for Retention, Security, "ROT," File Location/Storage and Privacy

Ultimately, the documents from both sides were successfully de-duped and merged, resulting in 70,000 unique documents/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.

VALORA technologies