Case Study

Catalyst Multi-Language Support Speeds International Arbitration

Global businesses operate in multiple languages. That is essential for business, but an obstacle for e-discovery. Such was the case when our client, a U.S. corporation, became embroiled in an international breach-of-contract case involving its former subsidiary, based in Hong Kong. The case required review of a 350 GB collection of hundreds of separate PST back-up files. The files contained some 3.36 million documents in either English or Chinese or a mix of both.

With review targeted to begin within days, the client wanted the multi-language data culled for documents relevant to the dispute. The pool of documents then had to be sorted by primary language, so that English documents could be routed to a review team in the U.S. and Chinese documents to a team in Hong Kong.

Catalyst’s Language Expertise Enables Rapid Processing

When Catalyst received the data, the first step was to process and index the files. Chinese-language documents present unique challenges. For one, many of the files were not in Unicode format; rather they were created using Big-5, a proprietary encoding method. Even more challenging, Chinese documents contain few punctuation marks or spaces to distinguish words and sentences. This makes them nearly impossible to index by traditional English-language methods.

Fortunately, the Catalyst Fast Track processing system is designed to work with both Unicode and the typical proprietary encodings found in Asian documents (e.g. Shift-JIS, JIS and Big-5). In addition, Catalyst’s FAST search engine supports search in more than 270+ languages including Chinese, Japanese and Korean. Catalyst CR can properly tokenize Chinese, which allows it to breaks documents into individual words for proper indexing and search. Other systems simply index individual character symbols, which makes search more difficult and less accurate.

A further challenge here was to group the documents for review either by English-language or Chinese-language review teams. The FAST indexer automatically identifies a document’s primary language and codes it accordingly. Here, however, the client wanted greater granularity in sorting the documents by language. To accomplish that, Catalyst came up with a method for extracting the text of each document.

Client Snapshot: U.S. Corporation

- International breach-of-contract case
- 3.36 million documents in English and Chinese
- U.S. and Hong Kong review teams
- Multi-language support allows rapid processing

With review targeted to begin within days, the client wanted the multi-language data culled for documents relevant to the dispute.
and precisely analyzing it for the occurrences of English and Chinese words. Documents could then be routed to the appropriate teams of reviewers.

The Catalyst Advantage: Multi-Language Support for Global Litigation

In just two days, the first batch of documents was ready for review. Catalyst was a forerunner in providing multi-language search and review for e-discovery and large-scale projects. Our embedded FAST search engine is the leader in supporting multi-language search. Search more than 70 languages with a single query. We specialize in the difficult CJK (Chinese, Japanese and Korean) languages, along with Russian, Arabic and Hebrew and Western European languages.