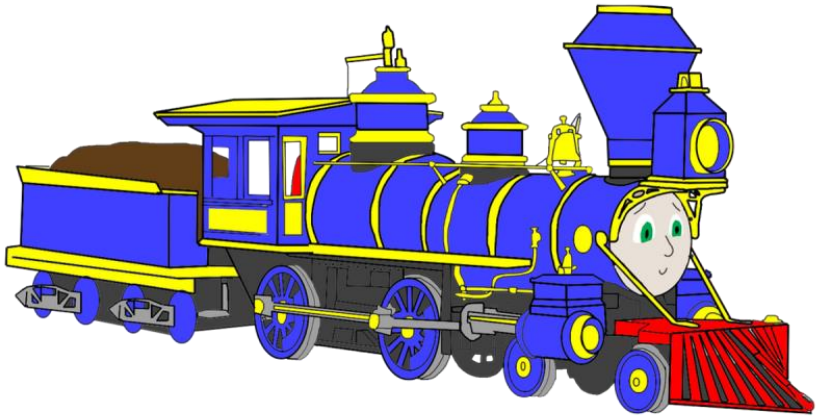
The Little Engine That Could!



When it comes to parsing and extracting data from mega PDF files, many well-known candidates run into deep problems. The challenge that we recently faced was to process a huge PDF file that had defied numerous attempts of processing with "known solutions".

It went like this:

We have a PDF file which has **1.9 million** pages and a file size of **3.4 GB**. It consists of **519,000** government correspondences of between 1 and 21 pages each.

The client wishes to parse this file, identify each correspondence, and extract indexing information such as date of the correspondence, correspondence number, recipient name, and mailing address.

The problem was that most PDF engines would either refuse to open the file or run into trouble soon after processing a few pages.

PDF provides a great medium for archiving collections of similar documents. By sharing graphics, images, fonts and other resources among documents, the storage requirement can be drastically reduced. It is also much easier to manage a single document than 519,000 individual files. Mega PDF files that were produced for this purpose are common in corporate and government institutions.

Parsing of such large PDF files for extraction of index data, protection of sensitive information, or performing analytics has always been challenging, until now...

The new PDF engine that we have developed at Opait Software was put through its paces in trying to work through our mega PDF file. The 64-bit version of the engine managed to pull through this file in a few hours with the CPU utilization never exceeding 15%!

Although we didn't focus on the processing time, the existing parallel processing framework in a multi-tasking environment can significantly reduce the processing overhead.

Our **FREE**, lean & mean, PDF viewer, can process such mega PDF files without hesitation. It is available for download from:

Opait Viewer - Give it a try!

Opait Viewer is probably the World's smallest full-featured PDF viewer with NLP enhancements, including automatic reading order detection (**section 508**), tokenizer, sentencizer, summarizer, and document title or abstract detection. The download size for the 64-bit version is less than 3.0 MB.