Cullen | Dykman

Simple Solutions: Optical Character Recognition with ABBYY FineReader 11

October 15, 2012

In modern litigation, the discovery process is often lengthy and expensive, depending on the complexity of the issues and the amount of documents that may need to be produced. Frequently, e-discovery vendors are necessary to help organize and sift through the potential millions of pages of documents that both your client delivers to you or opposing counsel produces. Sometimes, however, a single Portable Document Format ("PDF") document contains all the evidence.

Nevertheless, the same problem that arises in large discovery projects may reveal itself in those smaller projects. That is, the option to search the document may not be available. We recently covered how to search the text of a PDF document, but that solution will not suffice if the attorney receives a file that is unsearchable. In those cases, it may be crucial to make the document searchable in order to save time and money for your client. For large discovery projects, this service is usually included in the e-discovery vendor's software, but for cases with minimal discovery, a costly e-discovery platform is unnecessary. Rather, a stand alone software package may be enough to meet the needs of the smaller discovery project.

Saving a client's money by obtaining indispensable e-discovery and litigation software has always been one of our firm's top priorities. So, when confronted by a discovery project that does not require an e-discovery vendor, we feel that it's our duty to seek out other more cost-efficient solutions, which is exactly what we did when we found software that could make a PDF or static image searchable. Generally, in order to make an image or PDF searchable, the software must be capable of optical character recognition ("OCR"). Optical character recognition is the electronic conversion of images into machine-encoded text. It is used by all the major e-discovery vendors to create searchable text, which can then be stored more compactly, displayed in a viewer, and used in the vendor's in-house developed software platforms.[1] For the most part, certain versions of Adobe Acrobat Reader can OCR a PDF file, but for projects with large PDFs, other image formats, or multiple files, software that specializes in OCR is the wiser solution.

After researching and reading multiple reviews, the software package that our litigation department chose for smaller e-discovery projects was ABBYY's[2] FineReader (http://finereader.abbyy.com/). ABBYY is not specifically an "e-discovery" vendor or an e-discovery software developer. Rather, it has developed software that has exceptional OCR capabilities. According to its website, "ABBYY FineReader is an award-winning professional OCR software that offers a broad range of functionality for various needs of business, academic and government environments. It helps to streamline document processing, turn scans, PDFs into searchable and editable

formats. Intuitive use and one-click automated tasks let your workgroup achieve more in fewer steps."[3]

The software has multiple versions and purchase options. Unlike the "professional" version, the "corporate" version will allow your litigation support team to batch process documents and set up program tasks that will export files to a specific format. The export feature of the program will allow the user to determine how the final document is to be produced, such as exporting each page of a multi-page PDF file into its own text document. Our litigation department has recently used the latter feature on a small discovery project.

In that case, we received multiple PDF documents that totaled approximately 20,000 pages. In order to make everything searchable, we opened the PDFs in ABBYY FineReader, exported each page into its own text file, uploaded those files to a web server, and then created a PHP script to import each text file into a table we created in a MySQL database. (For the more tech-savvy users, the PHP script is below. The script takes a text file and inserts the contents and file name into a row in the database. Feel free to use it.) Once the data was in the database, we could run queries and perform Boolean searches for all the information.

As firms continue to grow their e-discovery departments, it is important to acquire software that can perform mundane tasks and develop a process to review documents more efficiently. With budget constraints often looming over litigation departments, it is important to consider simple solutions to the often protracted discovery process.

A special thanks to Sean R. Gajewski for helping with this post and creating the e-discovery procedure described above. Sean is a law clerk in our litigation department at Cullen and Dykman.

PHP CODE:

```
andlt;?php

$handle = opendir('.');

if(is_resource($handle)) {

$dsn = 'mysql:host=DATABASE_HOST;dbname=DATABASE_NAME';

$login = 'DATABASE_USER';

$password = 'DATABASE_PASSWORD';

$dbh = new PDO($dsn, $login, $password);

while (false !== ($entry = readdir($handle))) {

if ($entry != "." andand $entry != "..") {

$contents = file_get_contents($entry);

$sql = "INSERT INTO leehr (docTitle, docContent) VALUES (?, ?);";
```

```
$sth = $dbh-andgt;prepare($sql);

$sth-andgt;bindValue(1, $entry, PDO::PARAM_STR);

$sth-andgt;bindValue(2, $contents, PDO::PARAM_STR);

set_time_limit(20);

$sth-andgt;execute();

}

}

closedir($handle);

}

?andgt;
```

1. **[1]** *See generally* Wikipedia contributors, “Optical character recognition,” Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/Optical_character_recognition (last visited October 10, 2012). ?
2. **[2]** **Disclaimer:** Cullen and Dykman is in no way affiliated with ABBYY; rather, we simply recommend their product as we use it daily in our litigation department. With that said, ABBYY FineReader may not be the right solution for your business or firm and you should consult with your IT or litigation support department before purchasing their product. The views expressed in this post are not endorsed by, or do they in any way reflect, the opinions and/or positions of ABBYY. ?
3. **[3]** *ABBYY FineReader for Business*, ABBYY, http://finereader.abbyy.com/corporate/platform/ (last visited October 10, 2012). ?