

Reporting Requirements of U.S. Congress Leads The National Institutes of Health to Implement PDFTextStream

The National Institutes of Health (NIH) is the primary Federal agency for conducting and supporting medical research in the U.S. With an annual budget of over \$28 billion, the NIH awards almost 50,000 competitive grants every year to more than 212,000 researchers at over 2,800 universities, medical schools, and other research institutions in every state and around the world.



Challenge

The sheer number of grant awards naturally leads to an overwhelming amount of documentation. Grant applications are lengthy documents on their own—and as they enter the evaluation process, review committees append numerous comments and notes, ultimately creating a large file for each application. As a result, the NIH has accumulated millions of heterogeneously formatted grant-related documents over the years, most of which are not stored electronically.

In an effort to generate more accurate funding reports for the U.S. Congress, the NIH needed better access to this storehouse of archived grant application data. To that end, the agency recently embarked in a series of projects to convert physical application-related documents—using an Optical Character Recognition (OCR) system—into PDF files. Once in PDF, the NIH development team employed a text extraction engine so that their content analysis application could categorize the grant application data.

Unfortunately, what came out of the initial OCR conversion process was a set of documents that were readable to the human eye but difficult for text extraction engines to process. The engines' extraction fidelity was not high enough to properly derive critical font and document attributes from the original documents—wreaking havoc on the content analysis system.

To further complicate matters, the NIH also had to enable indexing and search within certain specific sections of grant application documents to help users narrow down their searching. However, not only were the extraction engines having difficulty interpreting the actual text, but they were also having trouble identifying section boundaries within documents.

"We feel [Snowtide Informatics] truly went the extra mile to make us happy—and their responsive, knowledgeable tech support was an unexpected benefit."
-- Shailender Chohan, Senior Developer, NIH

CHALLENGE

A large Federal government agency needed a way to rapidly and accurately extract text from archived PDF documents. The text extraction tool also had to enable fast and accurate indexing of specific sections within the PDF documents in order to feed a Java-based content analysis application used for generating funding reports for Congress.

SOLUTION

PDFTextStream—an enterprise-class PDF text and metadata extraction API for Java applications and Web services.

RESULTS

- **Fast and accurate text extraction from thousands of heterogeneously formatted PDF documents**
- **Automatic identification of document section boundaries enabled more relevant document searches**
- **Hundreds of man-hours saved to date; thousands more expected.**

Solution

The development team at NIH searched diligently for a Java PDF text extraction API (Application Programming Interface) with a high enough extraction fidelity to enable adequate indexing and search. Besides evaluating Snowtide Informatics' PDFTextStream, the team also tested various open source APIs such as iText, PDFBox, and JPedal. In the end, NIH chose PDFTextStream.

"We ended up not using any of these open-source APIs because they could not provide the functionality and the quality technical support we needed," said NIH Senior Software Engineer, Mark Yu.

Results

PDFTextStream enabled the NIH development team to more rapidly and accurately extract text from the grant-related PDF documents. More importantly, it allowed them to accurately extract text based on specific section boundaries within these documents.

Said Yu, "PDFTextStream did a very good job of identifying the specific page and spatial location of each section within the PDF documents. This allowed us to 'bucket' the content more easily and selectively extract text within these section boundaries."

"The impact has been tremendous. PDFTextStream has kept us from having to manually specify sections within thousands of PDF documents. This has allowed us to automate extraction at a very high speed, saving us hundreds of hours of tedious work. By the time the project is complete, it will have saved us thousands of man-hours, while also improving text extraction accuracy," Yu added.

Regarding the implementation, Shailender Chohan, Senior Developer at the NIH, said, "We found PDFTextStream to be an easy API to use; it didn't take long to implement. And our expectations regarding Snowtide's service were exceeded."

"In fact," he added, "Snowtide was extremely responsive and helpful in resolving the few issues we did find. They even created new, helpful features especially for us. We feel they truly went the extra mile to make us happy—and their responsive, knowledgeable tech support was an unexpected benefit."



"By the time the project is complete, [PDFTextStream] will have saved us thousands of man-hours, while also improving text extraction accuracy."

-- Mark Yu, Senior Engineer, NIH



What's Your PDF Problem?

Snowtide Informatics helps organizations of all kinds solve their most pressing and demanding PDF content extraction problems. Specifically, we excel in automating the extraction, conversion, and cataloging of content held in PDF documents by developing and applying high-performance software components and systems.

Snowtide is based in Northampton, Massachusetts. PDFTextStream, our flagship product, is a software component library for Java and .NET environments that has been built from the ground up to rapidly and accurately extract text and metadata held in PDF documents.

If you have a challenging document extraction problem, we can probably help.

243 King Street, Suite 248
Northampton, MA 01060
USA

Phone/Fax: +1 877.733.8980
www.snowtide.com