Success Story
Medical Education

ABBYY®

Automated OCR Helps Pioneering Medical Archive Create Searchable Index of 1,000,000 Files

Jeghers Medical Index comprises over one million medical journal articles, the vast majority of which were paper. To create a searchable digital archive from them meant scanning over 5 million pages into TIFFs, then converting them into archival PDFs. The task that was completed accurately and automatically using ABBYY Recognition Server.

> *"We ran tests for accuracy, speed and cost of several OCR software products. ABBYY Recognition Server had the most accurate OCR for our documents."*
>
> *- Lori Gawdyda, Medical Librarian, Jeghers Medical Index*

An invaluable resource for improving patient care and learning

Believing that effective medical education evolved from the care of patients, Dr. Harold Jeghers, M.D. founded the Jeghers Medical Index (JMI) in the early 1930s. His goal: Provide students and practitioners an easily searched repository of articles excerpted from medical journals to help them keep up with advances in medicine.

To achieve this, Dr. Jeghers created a classification system based on the everyday medical language of physicians. Over the decades, the JMI library grew to over 1 million articles and today supports 20,000 searches every year.

Reaching the limits of a 20th Century archive

The JMI archive contains articles from the 1920s to the late 1990s – putting over 70 years of medical history at the disposal of those seeking guidance from past research and treatment.

However, by 1997 many of the library's paper articles were beginning to decay and JMI undertook to preserve their content against this. Over 100,000 articles were scanned, indexed, added to an electronic database as TIFFs and made available online. But nearly a million more remained as paper. And as Lori Gawdyda, JMI's Medical Librarian, describes, manually accessing them could be a time consuming process:

"The articles were located in 165 filing cabinets and 44,000 folders – and there were more than

About Jeghers

Jeghers Medical Index provides the medical community with vital insights into the history and progress of 20th Century healthcare. Its unique collection of over one million articles have helped medical residents and community hospital practitioners improve patient care and treatment – and serve as an invaluable resource to medical students and historians. Learn more at www.jeghers.com

1 million articles. There was a database that listed the folder title, and the folders were arranged by body system. So for example, you would first have to search the database for, say, umbilical hernias. Then you'd have to retrieve the folder and look through the articles to find the answer to your question."

With most of its content exclusively available onsite and retrievable only through time-consuming manual searches, the full benefits of the JMI's resources remained unrealized. So JMI began the task of bringing the entire library into the 21st Century.

Seeking a solution for the digital age

To enhance its value to researchers and preserve its role in furthering medical education, JMI decided that all of the library's content should be made available online as a searchable digital archive. To prepare for the transition, the library began scanning articles into tiffs in 1997. And as the scanning process continued, JMI established its goals for the project – including requirements for a state of the art search and retrieval system for the new archive.

To ensure full searchability and longevity, the Jegher team decided on the PDF/A-1a file format for the archive's articles. They then engaged with Thunderstone to create an SQL relational database management system integrating a customized search device. And, after careful research, contacted ABBYY about a solution for converting the scanned articles into PDF/A-1a files. "They recommended DocuSyst, an ABBYY partner local to us," recalls Gawdyda. "And we engaged with them to convert the files."

Bringing seven decades of medical history into the 21th Century with ABBYY Recognition Server

For JMI's new archive to deliver highly accurate search results, its digital files had to be converted into PDFs as precisely as possible. And for DocuSyst, this meant using ABBYY Recognition Server. "Accurate automated document conversion was our only option," says Eric Posa, President of DocuSyst. "And for that, ABBYY Recognition Server is our go-to solution. Jeghers demanded the highest accuracy. We wouldn't have considered anything else."

According to Posa, the scale and complexity of the project required a great deal of research and planning. "Just scoping and planning the project with JMI took an entire year – resulting in a 108-page project plan." From there, DocuSyst's team began the process of converting files from the JMI library, provided on hard disk. "We started with an initial test batch of 25,000 files chosen at random by the JMI team," says Posa. "After verifying it exceeded the agreed accuracy benchmarks, the remaining articles, nearly a million of them, were sent over in a series of 7 batches over the course of a year."

And as Posa describes, quality control was stringent across the entire project. "Every batch completed was subjected to verification, based on random sampling of the converted files. Recognition Server never let us down."

The results

Conversion of JMI's tiffs began in August 2012 and was completed in May 2013. Since then, the resulting archival PDFs have been integrated into JMI's digital archive – and their new search appliance powered by Thunderstone's Texis relational database management system. "The appliance," explains Posa, "uses an algorithm to take the content of a document and identify it based on the OCR results. It crawls all the PDFs, collects all the information, then tags it so that everything is properly indexed in the system."

And what can users expect from the new system? According to Lori Gawdyda there's a night-and-day difference between JMI's new automated search process and the old one: "It is much faster, more comprehensive and offers many more options. The archive can be searched by author, title, subject, journal, publication date, by words in an article, by adjacency (meaning the words can be in the same sentence), paragraph, page, wildcards, truncation and much more."

Gawdyda concludes, "We are very pleased. And DocuSyst was very customer oriented. They finished the project on time, below budget and well within the defined quality parameters. It was a pleasure working with them."

Learn more at www.ABBYY.com/recognition_server

The Challenge:

Enable JMI to transform over one million paper medical articles and TIFFs into a searchable online archive, and eliminate tedious manual searches for users.

The Result:

ABBYY Recognition Server facilitated accurate and automatic conversion of articles totaling over five million pages into searchable PDF/A files for use in JMI's new online digital library.

"ABBYY Recognition Server is our go-to conversion solution. Jeghers demanded the highest accuracy. We wouldn't have considered anything else."

Eric Posa, President, DocuSyst

ABBYY USA
880 N. McCarthy Blvd.
Suite 220
Milpitas, CA 95035, USA
Tel 408.457.9777
Fax 408.457.9778
sales@abbyyusa.com

www.ABBYY.com

DocuSyst
550 Fillmore Avenue
Tonawanda, NY 14150, USA

www.DocuSyst.com

DocuSyst

ABBYY®