ABBYY FineReader® Engine Transforms Scientific Papers into Digital Knowledge

Customer Overview

CNKI 中国知网
www.cnki.net
中国知识基础设施工程

Name

China National Knowledge Infrastructure

Location

China

Industry

Education

Challenge

Digitize and organize scientific materials in a variety of languages while preserving their layout.

Solution

Implementation of a solution based on ABBYY FineReader Engine.

Results

- Increase in speed and accuracy
- Preserved structure and layouts of the documents

China National Knowledge Infrastructure (CNKI) is an e-publishing project supported by the Government of China — Ministry of Education, Ministry of Science and Technology, Propaganda Ministry and General Administration of Press and Publications. The project provides over 90% of China knowledge resources, the widest in title, type and geography coverage and deepest in year coverage in the country. The database covers journals, dissertations, newspapers, proceedings, yearbooks, reference works, encyclopedia, patents, standards, S&T achievements and laws and regulations in multiple scientific areas.

Mass digitalization of knowledge resources in China in late 90's initiated a creation of the most comprehensive system of China academic knowledge platforms. Thus, in 1999 Tsinghua University and Tsinghua Tongfang Holding Group built China Integrated Knowledge Resources Database and introduced a standardized system of Chinese academic journals. Today every scientist in China uses the platform, and every dissertation or scientific research is based on its resources.

Challenge

With the focus on education, CNKI has a massive library of books, documents, journals, doctoral dissertations, newspapers etc. both chinese and foreign ones in paper form, which needed to be digitized and organized into easy-to-search knowledge database — thousands of titles in an archive and hundreds of new ones being added every day.

Apart from the huge number of materials, the other issue to cope with was multiple languages, including Chinese, Vietnamese, Thai, most European languages etc.

Besides, the specifics of scientific works and dissertations is the abundance of illustrations, tables, schemes, graphics, diagrams etc, which also have much value and need to be preserved. Moreover, all the materials needed to be searchable and saved in special CAJ format (China Academic Journals).

Solution

With all the specifics mentioned above, manual digitizing turned out to be very hard and a big burden for CNKI, not mentioning huge time waste in this case. Therefore, the organization implemented an OCR solution by a local Chinese vendor to automate and fasten the process. The results were definitely better and faster than manual retyping, but poorer than expected.

First, the system supported only Chinese language, not covering a significant quantity of materials. Second, recognition quality was quite low, taking too much time and effort to verify the results. Third, the solution captured only the text and did not preserve the layout and other elements.

In order to replace the core OCR solution, CNKI addressed Shanghai Tai bi Information Technology — golden partner of a world-leading OCR and data capture technology vendor ABBYY.

To digitize the backlog of materials in the shortest period, Tai bi offered to use ABBYY FineReader Engine — OCR SDK, which enabled deep and seamless integration with the CNKI's existing environment.

At the first stage of processing, ABBYY FineReader Engine recognized the full text of documents, and at the second stage, it captured index values (metadata) from the content of documents. Those metadata were then used to perform fast and efficient search through the digitized materials across the knowledge database.

In comparison to previous OCR solution, ABBYY FineReader Engine allowed to preserve the original layout of documents and thus export the processed documents into Microsoft® Word, Excel®, searchable PDF/A and local Chinese format CAJ to comply with the national standards.

100% accuracy of search result was ensured by just one operator, who quickly and easily verified ABBYY OCR recognition results.

"ABBYY is an international famous OCR technology vendor, their OCR accuracy, even the Chinese OCR accuracy is far beyond my expectations. ABBYY technologies save us much time and improve our productivity, we are looking forward to the further cooperation in our new workflow."

Mr. Wu, Technical Director, CNKI

Results

With the implementation of ABBYY OCR technology, CNKI has significantly improved processing speed and accuracy and reduced human control.

Smart document analysis by ABBYY FineReader Engine has helped to preserve the structure and layout of the exported documents, which is important for further use and storage in terms of CNKI project.

By using multiple processing cores, they have increased speed. In the past, the same tasks could have taken several weeks, but now, just a couple of days.

Thanks to automation of the process, the organization has released tens of people, who previously performed manual digitizing and verification, and involved them into other projects. The productivity has grown much higher.

However, the most important result of this large-scale digitizing project consists in the increased comfort of usage. Users of the global platform now find the necessary information much faster and the results of this search are more accurate. Thanks to ABBYY's digitalization solution, nation-scale knowledge of China became more accessible and workable, which completely coincides with ABBYY's main mission — to action information.

ABBYY®

www.ABBYY.com