ABBYY FineReader® Preserves the National Library of Latvia for Future Generations

Customer Overview

N A T I O N A L
L I B R A R Y
O F   L A T V I A

Name

National Library of Latvia

Location

Riga, Latvia

Industry

Education

Industry

lnb.lv

Challenge

Turn the texts of the National Library of Latvia into searchable archives

Solution

ABBYY FineReader Engine

Results

- 4 million pages of books & periodicals processed in less than a year;
- Library materials are now accessible online.

As gateways to knowledge and culture, libraries shape the new ideas and perspectives that are central to a creative and innovative society as well as ensure an authentic archive of knowledge created and accumulated by past generations.

Challenge

The National Library of Latvia (NLL) has amassed 4.5 million paper units, including special collections — rare books, manuscripts, Letonica (i.e. books on the history of Latvia and Latvians), the Baltic Central Library, maps, scores, sound recordings, graphic documents, small prints, periodicals.

On the one hand, since its establishment in 1919 some of the oldest editions kept in the library have started deteriorating; on the other hand, the library fund has accumulated tons of valuable and popular materials. In other words, there arose a task to preserve these materials for the future and make them more accessible for the public today — a task accomplished by creating a digital archive.

Solution

The Internet has created tremendous opportunities in terms of accessing collections of the world's greatest libraries. Large-scale digitization of NLL, however, had yet to be realized. The first phase of the project included the scanning and creation of image-only PDFs, which wasn't good enough as the texts were impossible to work with.

In order to convert the materials into searchable formats the library needed OCR technology. But there another pitfall awaited: few OCR solutions could provide high quality of Latvian scripts recognition, to say nothing of support of ancient Latvian and European fonts. However, after a while the solution was found, and the second phase of archive digitization included a small pilot project with the use of ABBYY OCR technology. This project was conducted by Content Conversion Specialists (CCS).

To provide some background, CCS has been involved in developing special software solutions for the Cultural Heritage community since 2000. As a result, a new software tool for structured digitization docWorks, based on ABBYY FineReader Engine technologies, was brought to life in 2003 and afterwards used for NLL project.

abstract from a periodical from the NLL website;
illustration of font and image quality

At the beginning the library chose materials that were either physically damaged and thus had to be "saved" at least in a digital form, or that were popular among readers, or were considered historically important. The approximate scope of work included 2.5 million pages of periodicals (equal to about 1000 titles of full sets of periodicals) and 1.5 million pages of books (equal to about 7000 books).

ABBYY FineReader Engine, an integral part of CCS docWorks solution, was used to perform optical character recognition of historic texts in as many as 20 different languages. The near-perfect support of Latvian and Russian scripts — with up to 100% accuracy — played a special role in the choice of OCR provider for the project.

It should be noted that the texts contained rare gothic fonts which have fallen out of use and are not supported by most modern optical character recognition solutions. However, both Antiqua and Fraktur groups of fonts with special ornamental design were easily handled by ABBYY FineReader Engine technology.

"The National Library of Latvia has been involved in a large-scale digitization project with the aim to process and make available online about 4 million pages of historic books and periodicals. ABBYY Finereader engine has been an integral part in the project, providing very high accuracy OCR results. Most of the texts in the project were processed with precision close to 100%. This result allows our users to both make use of high quality OCRed text and do full-text searches in the periodicals portal: www.periodika.lv"

Joachim Bauer, Head of docWorks Group, CCS

Results

It took a little more than a year to process 4 million pages of ancient books and modern periodicals. Driven by the enthusiasm of a noble goal, 60 operators worked daily in three 8-hour shifts during the project's peak.

After the processing, the documents were exported into various formats (PDF, JPEG, XML) and imported into the periodicals portal (periodika.lv), where they became available to scientists, researchers, professors, students and general public. Due to copyright protection, most materials are accessible only from the network of Latvian libraries, although all periodicals published before 1941 are available with no restrictions, and public domain books (i.e. with expired copyright) are also available to all internet users.

About ABBYY

ABBYY is a leading global provider of technologies and solutions that help businesses effectively action information.

ABBYY®

www.ABBYY.com