ABBYY®

CASE STUDY

ABBYY® Recognition Server

UNIVERSITY OF Southampton

University of Southampton Improves Access to its Collection by Digitising Vast Quantities of Printed Material

Background

The University of Southampton Library supports a community of approximately 35,000 students and staff, providing access to an expansive collection in excess of 1.5 million books and many millions of pages of archive material. The Library recently embarked on programme to digitise a large number of its key texts through the Library Digitisation Unit (LDU). The LDU is a flagship enterprise in the academic sector and specialises in the digital capture of a range of materials for repositories or web distribution via URL links in the Library catalogue. The University Library's approach is to provide open access to the digital material that it creates, wherever this is appropriate and permissible.

ABBYY FineReader, a software product designed for ad-hoc scanning and digitisation, had been used by the Library for a number of years to idigitise small numbers of pages, and occasionally full books, with the addition of text recognition or OCR (optical character recognition). The Library quickly realised however that in order to automate the OCR process with a high throughput and achieve their goal of digitising half a million pages per year, they would need a more robust product capable of automatically processing large volumes of documents.

Solution

In order to find a solution to process printed documents into searchable formats, such as PDF and PDF/A for digitising archives or records creation, the Library evaluated various options on the market. They examined the following criteria across a number of products: Speed and quality of OCR, range of output formats and compressions, and API/workflow integration possibility. After looking at a number of products the Library selected ABBYY Recognition Server as a best-fit solution.

The Library selected Recognition Server because it perfectly matched their requirements – delivery of high quality OCR on printed texts; a broad variety of output options; and an open API for easy integration with other programs. The Library wanted to integrate Recognition Server with the Unit's Intranda GmbH workflow software, Goobi. The Goobi Production Workflow software is a web application that manages and tracks the Library's digitisation projects. Additional considerations in ABBYY's favour were the level of after-sales support and the affordable maintenance.

The LDU currently uses up to six book scanners and one high-end line scanner to digitise texts and images from their collection. ABBYY Recognition Server's XML ticketing API was used to integrate it with the Goobi Workflow. After the printed materials are scanned, Goobi manages the queuing of jobs to Recognition Server and then monitors the output. Character coordinated output can be ingested into a presentation layer for indexing and access. Thus as soon as the scanner operator completes the digitisation of a book, the files automatically move through each stage of the workflow

Two of the larger digitisation programmes of the University of Southampton Library stock included JISC (Joint Information Systems Committee) funded projects. The JISC inspires UK colleges and universities to use digital technologies in innovative ways and helps to maintain the UK's position as a global leader in education. Those two projects each generated over 1 million digital images:

About University of Southampton

The Library Digitisation Unit of the University of Southampton offers a scanning service to libraries, archives and the commercial sector. Since 2003, we have specialised in the digital capture of images and text from bound and unbound materials in a conservation environment.

Contact

www.soton.ac.uk/library/ldu

ABBYY Europe GmbH
Elsenheimerstr. 49
80687 Munich, Germany
Tel: +49 89 51 11 59 - 0
Fax: +49 89 51 11 59 - 59
sales_eu@abbyy.com
www.ABBYY.com

ABBYY UK Ltd.
Abbey House, Grenville Place
Bracknell RG12 1BP, UK
Phone: +44 1344 392 610
Fax: +44 1344 392 611
sales_UK@abbyy.com

Digitisation of 18th century Parliamentary Papers: Coverage includes the Journals of the House of Lords and Commons, Parliamentary Registers, Session Papers of the House of Commons, Acts, Bills and Local and Personal Acts from the 1700s to 1834. Examples can be found at: **http://www.southampton.ac.uk/library/ldu/parl18c.html**

Digitisation of 19th century pamphlets: Over 23,000 19th century pamphlets from UK research libraries that cover the socio-political and economic landscape in Britain were digitised by the LDU and converted into searchable PDFs. Project details and catalogue entry is available at: **http://www.britishpamphlets.org.uk/**

Other in-house projects include:

Digitisation of Doctoral Theses: Scanned copies of 20,000 University of Southampton awarded theses were converted to searchable PDFs and are now being made available through the university's institutional research repository, Eprints Soton. Eprints Soton contains electronic copies of research output, including journal articles, book chapters, conference papers and theses. It also includes unpublished manuscripts and papers. The full text of many of these items is freely available to be used in accordance with copyright and end-user permissions. **http://eprints.soton.ac.uk/**

Digitisation of Parliamentary Papers relating to Ireland (EPPI): Approximately 15,000 Parliamentary Papers from the period 1801 to 1922, were digitised and made available as searchable PDFs via links in the University catalogue.

Digitisation of Course Texts: The Library converts printed items into searchable PDFs in order to support courses with high text demands. Searchable PDFs are made available to restricted University users via the library catalogue within the licence from the UK Copyright Licencing Agency (CLA).

The Richard Rutt Collection: A collection of 19th century knitting books from the Winchester School of Art Reference Library have been digitised and made available as searchable PDFs via the web. **http://www.southampton.ac.uk/library/ldu/wsa.html**

Conclusion

A powerful and accurate OCR is an integral component of the Library Digitisation Unit's activities. Thanks to the increased throughput offered by ABBYY Recognition Server, Library staff are freed from the tedious work of manual OCR and millions of pages of documents from the University of Southampton's collection are available online in digital formats to its students and to the wider world.

The successful completion of the digitisation can also be attributed to the product's manageability and smooth integration with existing Library processes. "The ability to integrate ABBYY Recognition Server into our workflow was critical," states Julian Ball, Unit Manager at the Library Digitisation Unit. "The installation of ABBYY Recognition Server was quick. Initial feedback from ABBYY and their support team regarding any queries was rapid with good follow-up. We have been very happy with the results and look forward to using the product to continue to achieve our digitisation goals."

About ABBYY

ABBYY is a leading developer of document recognition, document conversion, data capture and linguistics technologies.

ABBYY's products include: ***FineReader*** *and* ***PDF Transformer*** *– end-user applications for document conversion;* ***Recognition Server*** *– a server-based OCR and PDF conversion solution;* ***FlexiCapture*** *– data capture programs for processing forms, semi-structured and unstructured documents;* ***FineReader Engine SDKs*** *that provide a full spectrum of ABBYY's recognition technologies; and* ***Lingvo*** *– a line of dictionary software.*

More information about ABBYY at www.ABBYY.com

ABBYY®

ABBYY Europe GmbH
Elsenheimerstr. 49
80687 Munich, Germany
Tel: +49 89 51 11 59 - 0
Fax: +49 89 51 11 59 - 59
sales_eu@abbyy.com
www.ABBYY.com

ABBYY UK Ltd.
Abbey House, Grenville Place
Bracknell RG12 1BP, UK
Phone: +44 1344 392 610
Fax: +44 1344 392 611
sales_UK@abbyy.com