Case Study

Education

DL Consulting Is Cutting the Cost on Digitizing the News with ABBYY®

Customer Overview

Name

DL Consulting

Location

Australia

Industry

Professional services

Challenge

Digitize world's aging newspapers stored on microfilms and make them available online

Results

The cost-per-page decreased dramatically from $1 to just few cents

A new cost-effective alternative for large-scale newspaper digitization is being launched by DL Consulting, the company behind the Veridian collection management software and related newspaper digitization services as well as an online search portal for the world's historical newspaper archives at www.elephind.com.

Challenge

Since 2002 DL Consulting has been assisting organizations around the world digitize millions of pages of historic newspapers and deliver the content as a digital collection. Based in Hamilton, New Zealand, with an office in Honolulu, Hawaii, DL Consulting also hosts digital newspaper archives for a diverse array of global cultural institutions including the National Library of New Zealand, Singapore National Library and a number of US States and universities among many others.

Managing Director Stefan Boddie said that current digitization projects underway across the globe have only scratched the surface, and there are still hundreds of millions of newspaper pages on microfilm still waiting to be digitized. He believes a new cost-effective, standards compliant digitization platform it has developed based on ABBYY's Recognition Server will be the key to providing broader online access to these untapped archives.

There are many major international projects to digitize newspapers. Many libraries and government organisations seek to digitize the content to provide free access, and there are many other commercial projects seeking to digitize archival newspaper content and charge for access.

The National Digital Newspaper Program underway in the US for the past 10 years has spawned a range of XML markup standards for the formatting and presentation of digitized newspapers on the Web.

METS and ALTO are the names of these XML standards which are maintained by the US Library of Congress. The METS standard is a flexible schema for describing a complex digital object (like a digitised newspaper issue). METS describes the structure of the object but does not encode the actual textual content of the object.

The ALTO standard fills this void by encoding the textual content of a digitized page in great detail, including styles and layouts. As well as encoding the digitized text itself ALTO encodes the spatial coordinates of every column, line, and word as it appears on the page.

“It’s always going to be difficult for OCR technology to deal with poor quality aging microfilm, however there are still millions upon millions of newer newspapers that can take advantage of the platform we have developed and employ a completely automated process. It is also quite applicable for older newspapers that have been scanned from the originals or have been microfilmed quite recently.”

— Stefan Boddie,
Managing Director
at DL Consulting

“While we are based in New Zealand most of the customers undertaking digitization projects for major libraries and cultural institutions are based in the US.”

— Stefan Boddie,
Managing Director
at DL Consulting

About ABBYY

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

ABBYY Australia
Citigroup Building, level 13,
2 Park Street, Sydney, NSW,
2000, Australia
Tel:+61 (02) 9004 7401
sales@abbyy.com.au

Solution

The combination of METS and ALTO (often written METS/ALTO) is the current industry standard for newspaper digitization used by hundreds of modern, large-scale newspaper digitization projects, for example:

- Chronicling America from the Library of Congress.
- The British Newspaper Archive from the British Library.
- Trove from the National Library of Australia.
- Papers Past from the National Library of New Zealand.
- NewspaperSG from the National Library of Singapore.
- Papers of Princeton from Princeton University Library.
- Columbia Spectator Archive from Columbia University Libraries.

“Traditionally we have been reliant on a small number of vendors who can produce data that complies with these METS/ALTO XML formats,” said Boddie. ”Over the past few years at DL Consulting we have been working to develop a pipeline that enables us to do it in a cost-effective fashion which we have now achieved using ABBYY’s Recognition Server.”

Out of the box, the ABBYY Recognition Server software includes ALTO support which Veridian has built on to enable the creation of METS compliant XML data. The pipeline also generates images that must comply with a standardised JPEG2000 profile. It toyed with converting basic PDFs and using ColdFusion to manipulate data into the right format.

“There were a series of different iterations trying different technology and eventually we settled on ABBYY around two years ago. The ALTO XML functionality attracted us and also the high profile of the ABBYY OCR engine which is underlying a lot of the platforms we have worked with. Recognition Server also offers some nice possibilities for process automation compared to other off-the-shelf OCR packages.”

Results

Digitizing aging newspaper archives can be a challenging task if the microfiche itself is decaying and the original digitisation process was not done at high quality.

“Traditionally the alternatives are quite expensive and it’s really only the high profile projects that have been able to afford these solutions,” said Boddie. “Faced with costs upwards of $1 per page, many projects have chosen to simply scan and OCR the newspaper microfiche to create a searchable PDF which can be done for a fraction of the cost but is not optimized for Web access. Our new solution will allow these projects to consider an ALTO and METS compliant alternative, which will only cost a few cents per page.”

The Veridian solution utilizing ABBYY Recognition Server has been built entirely in the Amazon cloud and is able to completely automate the process and remove manual operators from the workflow. Most competing alternatives employ operators in low wage countries to assist with validation, which is still required in projects that require individual articles on the newspaper page and advertisements to be segmented.

Originally published by IDM Magazine: https://idm.net.au/article/0010461-cutting-cost-digitising-news