Case Study
Digital Archiving

ABBYY

Historic Hansard brought to life in SEMA digitization project

Customer Overview

Name

SEMA Group

Location

Australia

Industry

Digital archiving

Challenge

Digitize all pre-1980 original Australian Hansard documents and make them accessible online.

Results

All Hansard original records successfully digitized, classified, validated and ingested into ParlInfo web site.

Sir Robert (Bob) Menzies, the 12th and longest serving Prime Minister of Australia, once said, “Never forget posterity when devising a policy. Never think of posterity when making a speech.” However Menzies’ place in posterity has since been assured, and all of his many speeches to parliament will soon be just a click away to all Internet users at the completion of a massive project by The SEMA Group to digitize all pre-1980 Hansard documents.

Challenge

This project will transform hard copy document collections, including original Hansard transcripts from Australia’s federation in 1901, into digital images that will be accessible online.

The origins of today’s Hansard can be traced back to the early years of nineteenth century London, when Hansard reporters were commissioned to produce an authentic and accurate account of parliamentary proceedings.

Since 1980 all Australian federal Parliament Hansard reports have been available in digital format, but there is a huge archive of pre-1980 parliamentary material that has not been available for viewing online until now.

Solution

The SEMA Group, an Australian IT services organization specializing in providing outsourced solutions for document-centric sensitive processes, developed a full end-to-end solution for the Australian Government Department of Parliamentary Services. The solution delivered both software and hardware to digitize the large paper-based archives held at Parliament House.

The images produced will be linked into the Parlinfo search engine. The search engine allows access to Australian Parliamentary resources including Hansards, Bills, Senate Journals, newspaper clippings, publications and much more. Parlinfo is accessible to anyone with an Internet connection and so the inclusion of the pre-1980 collection will give Australians a unique insight into historical events such as our federation at the touch of button.

One of the complexities of the project was the fact that prior to 1953 Hansards of the Senate and House of Representatives were physically integrated, day but day, in terms of numbering and sequence, however after 1953 there were two separate series. These are

as follows: House Of Representatives 1953 onwards (121 Volumes, 157,412 pages [x2 for images]), Senate 1953 onwards (88 Volumes 106678 pages [x2 for images]) and Pre 1953 (222 Volumes 346,444 pages [x2 for images]).

The physical Hansard pages are extracted from the original volumes (after guillotining the spine of the book away), and the page size for scanning is approx. 150mm x 200mm. Scanning is being undertaken using Kodak Model i780 and i1440 scanners with Hansard pages captured and processed as single page TIFFs.

SEMA is utilizing dual OCR Interpret Engines from ABBYY and RecoStar to ensure the highest integrity of capture.

All TIFFs are batched into Hansard records grouped by Day, then these are converted to PDF/A files. These PDF files are transferred to the Department of Parliamentary Services (DPS) along with the associated XML data for the Hansard Batch {Multi-page PDF}. After the DPS team performs QA on generated output, PDFs and XML data is loaded into the ParlInfo system and is made available to the public via the web.

Tony Smith, Software Product Manager at SEMA Group said there were huge challenges in dealing with the physical condition of Hansard volumes that dated back to 1901. “The quality of paper from this period was variable, and some of the individual pages and even whole volumes had deteriorated over this time and had effectively gone a very rusty orange colour,” said Smith.

“We were able to obtain the best possible image using software to clean and optimize these images, by understanding the raw characteristics the early Hansard documents. To achieve a successful result in imaging and OCR capture and Indexing, it effectively all starts with good images.”

Additional challenges faced by SEMA included the variable print quality, typeface changes and differing format and layouts used over time. “These all factored into further challenges for templating,” said Smith.

Results

A vast array of business rules had to be applied to control the indexing/tagging of the OCR Hansard information found upon each page, in order to ensure that the data was correctly tagged and mapped into the associated XML file structure.

This required the building very large reference libraries with distinct keywords [Hansard defined tags] that needed to be captured and tagged with information [metadata] found in that location.

“With such a large set of business rules for image treatment and metadata tagging, there was an equally large amount of effort expended in the testing to ensure the integrity of the output meet the requirements of the DPS. All business rules were successfully implemented,” said Smith.

Hansard Digitization Workflow

Scan
Validation
Snippet and indexing of tables
Capture and count the ayes, noes and pairs from Hansard pages
QA spell checking
QA paragraphs are correctly captured
Exception queue management
Release
Generate text searchable PDF
Generate XML
XML marked up for ingestion into ParlInfo web site
Members speeches are correctly identified.
Questions and debates correctly classified.

Originally published by IDM Magazine: https://idm.net.au/article/008293-historic-hansard-brought-life-sema-digitisation-project

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