

TREND



Accessing the Contents of Archived Documents with A2iA

The digitization of archives and old documents has been identified as a necessity to preserve their integrity and overcome the scarcity of space. But how can the contents of these documents be accessed once they have been scanned? How can specific information be found amongst the terabytes of images?

It would seem that digitization is merely the first step in a multi-step process that consists of qualifying, classifying and indexing images of archived documents so that their informational wealth can be fully utilized.

While ordinarily the process of classifying and indexing images can be automated, the task becomes more complex when dealing with archived documents. These files have characteristics that make them challenging to automate with existing electronic

document management (EDM) solutions, such as: documents that are primarily handwritten; contain writing that is old-fashioned; mixed formats that do not follow any predefined template; possess a quality that has been diminished by time; and poor storage conditions, to name a few.

These factors, compounded by the volume of images to be processed, make the process overly complex and are often quite substantial. Consequently, the manual indexing of documents becomes a long, painstaking process that results in numerous errors. Because of this, many organizations have chosen to outsource this process, but not without risk, notably in terms of data confidentiality and quality control.

A2iA, the world leader in handwriting recognition and automatic document processing, still operates one of the world's first private research centers specializing in the recognition of cursive handwriting, and has developed one-of-a-kind software that can automatically classify archived documents and extract pertinent information.

A2iA Processes 1.3 Million Files Dating As Far Back As World War I

To give every citizen the opportunity to access the French collective memory, the French Ministry of Defence has entrusted A2iA with processing the 1.3 million files of "Those Who Died for France" during the war of 1914-1918. According to the legislation, files containing information of a medical nature should remain confidential. As the only technology that can read old handwritten documents, A2iA's technology made it possible to automatically extract those files that contained medical data and thus protect their confidential character. To do this, A2iA analysed all the files to: locate the field to recognise, detect the heading "Type of death", search for medical key words, then categorise the file as "normal" or "medical". The Ministry of Defence insisted on high-quality detection with a maximum rate of error of 0.5% among the accepted documents. A2iA exceeded the ministry's expectations with a 0.2% rate of substitution for the processed files and data input savings of nearly 70%.

A2iA DocumentReader™:

- Recognizes handwriting, both modern and old-fashioned styles.
- Categorizes complex documents (documents that are old or unstructured, or that occur in varied formats, etc.).
- Extracts all types of information, whether printed or handwritten.

A2iA DocumentReader™ analyzes the images of scanned documents based on both their geometry and content. It then performs a literal transcription of the handwritten and/or typewritten areas and in the following steps, extracts key words or expressions. Then, for each analyzed image, the software defines the document-type and locates, extracts, and converts key information into digital data that can become searchable and reportable, with the same level of flexibility as printed or digital data.

***A2iA DocumentReader™* Indexes Pre-1908 French Census Data**

In order to make archived resources held privately or by local and regional authorities available via an Intranet to genealogists, the genealogy firm, Coutot-Roehrig, digitized nearly 20 terabytes of images, amounting to more than a half-billion documents (civil status, population counts, etc.). To facilitate the search functionality within these archives, Coutot-Roehrig wanted to automate image indexing, a process that if done by hand, would have required several dozen years of work. The problem was not a simple one, as the documents were over 100 years old and not always in good condition. After several tests, the firm selected *A2iA DocumentReader™* to index 350,000 pages of census tables dating from pre-1908. The software recognizes, and then extracts, the first and last names and family relationship between individuals. In light of the positive results achieved, the firm plans to entrust A2iA with indexing of additional census data, as well as civil status.

Thanks To *A2iA DocumentReader™*, Yale University Puts Its Botanical Specimens Online

To allow scientists around the world to access its collection of herbaria, Yale University sought to upload several thousand botanical specimens to the Internet. The task was far from easy: some documents were almost 160 years old, most of them were handwritten and the format and placement of information varied from one page to the next. Yale University chose *A2iA DocumentReader™* to re-transcribe the handwritten information from the specimens. *A2iA DocumentReader™* is, in fact, the only software capable of recognizing antiquated handwriting and processing unstructured documents. *A2iA DocumentReader™* analyses the pre-scanned images of the specimens in progressive steps. It locates, segments and then extracts the handwritten information, which it then converts into digital data using a dictionary containing 6-million scientific terms. This data, along with their corresponding images, are then transmitted to the Yale server, and published to the Internet.

CONTACT

EMEA : + 33(0) 1 44 42 00 80
AMERICAS : +1 (917) 237 0390
mail : contact@a2ia.com
web: www.a2ia.com

