# Handwriting Recognition Streamlines Historical Document Conversion

A document classification and data extraction toolkit's ability to recognize handwriting helps Coutot-Roehrig digitize 100-year-old census records and private archives.

by Ken Congdon

Coutot-Roehrig used A2iA DocumentReader to recognize and extract handwriting from old census documents.
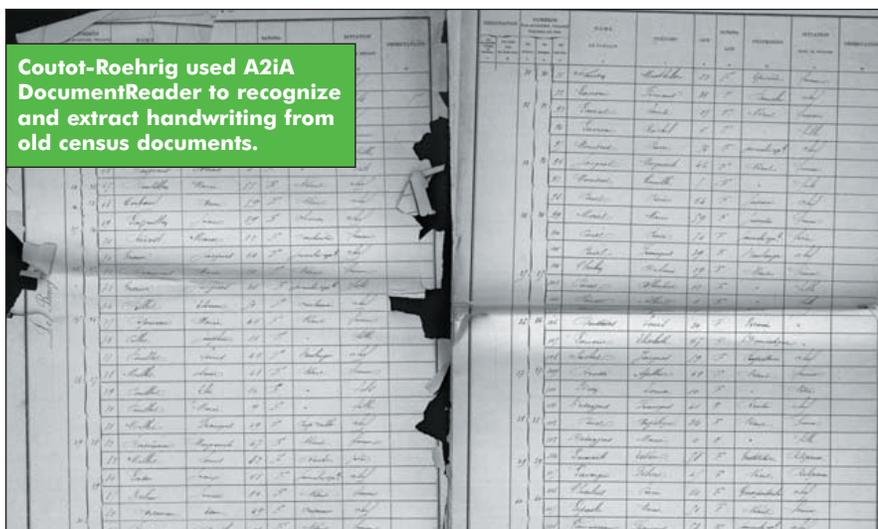
Data capture and extraction technology has come a long way in the past decade. Once only able to identify structured data in specified fields on a form, today's data capture technologies are highly advanced — capable of identifying words and phrases anywhere on a page, and using this data to automatically index and classify documents. While these capabilities are impressive, one area where some data capture products still fall short is in their ability to recognize handwriting as opposed to machine print.

This shortcoming can pose a serious challenge to many government agencies, healthcare facilities, and private corporations seeking to digitize the older, handwritten records (e.g. deeds, old census forms, patient charts, etc.) that make up their archives. One organization that recently discovered the impact effective handwriting recognition technology can have on improving operational efficiencies is Coutot-Roehrig — a European genealogy company that specializes in finding unknown heirs, missing beneficiaries, and real estate owners.

OLD, HANDWRITTEN DOCUMENTS MAKE INDEXING TREACHEROUS

Coutot-Roehrig manages more than half a billion documents (e.g. public records, census lists, city/municipality resident information, etc.) that have been digitized and posted on an intranet site for convenient search and review by genealogists. Most recently, the company decided to digitize the records held by the departments and municipalities in France. Partnerships were established with many local French communities that agreed to make their records available. The local communities would benefit from the use of the digitized public archives, while Coutot-Roehrig assumed full responsibility for the costs associated with the digitization process and maintains a set of all digitized images for their private intranet.

In order to optimize the search functionality for these records, a team of full-time employees was initially dedicated to manually indexing the scanned municipality records. It was quickly realized that if the entire project were to be processed in this fashion, it would have required a dozen years or longer to complete. The reason? The vast majority of these documents were handwritten, more than 100 years old, and images varied greatly in quality because they were scanned by multiple service providers. Because of this detail, the manual indexing process was extremely time consuming, costly, and error prone.

Learn how handwriting recognition helped a telecom provider at http://bit.ly/bxH6Yc.

AUTOMATIC DATA EXTRACTION DOES THE JOB WITH 1/5 THE LABOR

Coutot-Roehrig sought salvation from this manual indexing nightmare through implementation of automated data extraction technology.

Imaging And Capture

However, many of the data capture products the company tested were not able to address the complexity of the project, which included handwritten documents dating back to the 19th century, old-fashioned writing styles, and mixed file formats.

After an exhaustive evaluation process, Coutot-Roehrig was finally able to find a technology solution that addressed its needs — A2iA

> **According to Coutot-Roehrig, the A2iA DocumentReader solution automatically extracts handwritten content with 70% to 80% accuracy.**

DocumentReader. A2iA DocumentReader, a document classification and data extraction toolkit, was installed to recognize and extract the handwritten data from the French municipality documents and classify these documents based on layout of each document and the content contained on each form.

Specifically, the technology automatically extracted the first and last name of each individual, as well as the family relationship. First, A2iA DocumentReader optimized the quality of each document image (e.g. despeckled, deskewed, optimized brightness/contrast, etc.) and separated double-page spreads into single pages. Then, the A2iA engine detected the columns on each page and segmented them into boxes. Once a box was identified, the software accessed a customized dictionary to automatically identify the words located inside each box. The extracted words were then converted into useable electronic data and transmitted to Coutot-Roehrig's intranet system as an XML file, making these historic records and civic archives searchable and available to genealogists and the public.

According to Coutout-Roehrig, the

A2iA DocumentReader solution automatically extracts handwritten content from these dated French documents with 70% to 80% accuracy. With this high rate of automation, the company was able to index more than 350,000 pages of census tables with 1/5 of the keying labor that was necessary prior to implementation. Through effective deployment of handwriting recognition technology Coutot-Roehrig was able to reduce labor requirements and costs, accelerate the pace of indexing, and allow for these historic documents to be searched online by both genealogists and the public. ∎