



CASE STUDY: YALE UNIVERSITY PEABODY MUSEUM OF NATURAL HISTORY

Media Contact:

Wendi A. Klein
Director of Marketing & Communication, North America
+1.917.237.0390 x4034
wendi.klein@a2ia.com

A2iA Corporation:

24 West 40th Street, 3rd Floor
New York, NY 10018 USA
+1.917.237.0390 office
+1.917.237.0391 fax

A2iA DocumentReader Drives Records Management Solution for Yale University

OVERVIEW

Specimen collections are the primary research archives documenting the biological diversity of plants and animals on this planet. Information from specimens is fundamental to biodiversity research, education and natural resource management. Digital technology greatly facilitates information retrieval from globally distributed natural history collections. Currently, however, most of the 2.5 billion collections worldwide remain unavailable in the electronic domain. More than 100 million botanical specimens in herbaria are no exception, with less than five percent data records digitized.

Specimens in the Yale University Herbarium collection date as far back as 160 years, making it an important national historical collection. Yale required a solution that would make 350,000 specimen labels searchable by scientists and educators at a global level, without incurring vast amounts of manual labor to key the historical data.

**SCIENTIFIC INTELLIGENCE EXISTS IN
YALE'S COLLECTION OF HERBARIUM
SPECIMEN LABELS. YALE NEEDED TO
UNLOCK THIS VALUABLE INFORMATION
AND MAKE IT GLOBALLY ACCESSIBLE.**

This project is being conducted as part of Yale's HERBIS project, a collaboration with the New York Botanical Garden and University of Illinois and Urbana-Champaign, and is funded through the National Science Foundation. HERBIS is making informatics tools, including A2iA DocumentReader, available for specimen image and data capture.

CHALLENGE

Approximately 350,000 specimen labels make up Yale University's Herbarium collection. The labels are highly unstructured; they vary in size, format and layout, and do not contain static fields. Furthermore, they contain very old cursive handwriting, some of which dates back 160 years. These factors present several challenges in converting the information on the labels into electronic data. Traditional OCR applications would not be able to recognize the cursive handwritten information. And manual keying of the information would take years to complete.

Yale required a solution that would make the labels electronically searchable by scientists and educators globally. Due to the high volume of specimen labels and their unstructured nature, "a significant amount of manpower would be required to manually transcribe the information contained on the hundreds of thousands of specimens in Yale's collection and hundreds of millions of plant specimens worldwide," said Dr. Reed Beaman, associate director for informatics at the Yale Peabody Museum.

Yale University wanted to make the information computer-accessible while creating significant efficiencies compared to methods in current use. To do this, they needed a robust records management solution capable of converting the paper-based specimen labels into electronic images and extracting the data to make the archive searchable.

SOLUTION

The team at the Yale University Peabody Museum of Natural History, headed by Dr. Beaman, who is responsible for overseeing this project, turned to A2iA Corporation for an application that would

SOLUTION (CONTINUED)

overcome the challenges presented by different sized specimen labels, unstructured formats and very old cursive handwriting. A worldwide leading developer of Intelligent Document Recognition and Intelligent Word Recognition technologies, A2iA technologies classify paper documents of all sizes and extract key information from them – even text written in unconstrained handprint and cursive handwriting.

**A2iA DOCUMENTREADER IS
ENABLING EASY AND EFFICIENT
ELECTRONIC SEARCHING AND
VIEWING OF THE ARCHIVE.**

Yale University selected A2iA DocumentReader for transcribing and searching information contained on herbarium specimen labels. By combining advanced classification and recognition technologies to mimic a human speed-reader, A2iA DocumentReader quickly sorts unstructured documents and searches handwritten content, making it possible to mine data previously unavailable in electronic form.

A2iA DocumentReader quickly locates, segments and processes the information contained on the Peabody Museum’s plant specimens, including the scientific name of the plant, where it was collected and who collected it. It then converts the words into electronic data, matching the scientific names against a museum-compiled lexicon. The lexicon is made up of thousands of taxonomic names, several million place names, and other specialized biological terms.

Specimen label images are captured and then processed through a cascading set of web services to extract machine-printed and handwritten text. The final data results are loaded into collection management software along with the corresponding specimen images. A2iA DocumentReader uses tools that wrap image processing, convert images to text, and employ data markup capabilities in distributed interoperable web services.

RESULTS

A2iA DocumentReader achieves greater efficiency, portability and scalability than manual input and other less-advanced recognition technologies. The human labor requirements are seconds instead of minutes for capturing data from each specimen.

By capturing data from specimen labels, A2iA DocumentReader is enabling easy and efficient electronic searching and viewing of the archive. It is unlocking valuable cursive handwritten information, previously unavailable in electronic form, for scientists, researchers and educators worldwide.

“We are pleased that A2iA DocumentReader is meeting the needs of Yale University for this project, the first official installation of the application, which was released in 2005,” said Jean-Louis Fages, president and chairman of the board of A2iA Corp. “A2iA DocumentReader is providing the Peabody Museum with searchable access to information previously only available through cumbersome manual research. Our recognition engine is extremely flexible and intuitive. We have fine-tuned the recognition algorithms for language and era-appropriate handwriting – in this case, the American handwriting styles from 160 years ago.”

**THE BEST TECHNOLOGIES
ARE THOSE THAT MAKE
INFORMATION MORE ACCESSIBLE
AND HUMANS MORE PROFICIENT.
A2iA DOCUMENTREADER DOES BOTH.**

This records management application is being easily adapted to meet similar needs in educational, scientific, government and business environments – in the United States and abroad.