Document Imaging Report

Business Trends on Converting Paper Processes to Electronic Format

4003 Wood Street ● Erie, PA 16509 ● PH (814) 866-2247 ● http://www.documentimagingreport.com

February 20, 2015

Solving the Unstructured Document Conundrum

A few weeks ago, I was interviewing the CEO of a growing document imaging and management ISV who told me that although his company offered software for capturing data from structured forms, there really wasn't much demand. He explained that although his clients certainly did a lot of scanning, the documents they were capturing weren't structured enough to benefit from traditional forms processing.

Axis Technical Group may have a solution. The Anaheim, CA-based ISV has developed technology for classifying and extracting data from completely unstructured documents. "People in the ECM industry love to say that their systems are made to handle the 80% of a business' information that is unstructured," said Kevin Ells, a capture and ECM industry veteran who was recently brought on board as Axis' director of marketing. "When you are only automating capture of structured forms, it makes it very hard to achieve true ECM.

"A good majority of content still exists on unstructured documents. When you take out a mortgage for example, you sign 30 documents. You've got a stack of paper an inch thick with paragraphs full of legalese. Getting information off that and into line-of-business systems can be very challenging. It typically requires that a human read the documents and key the data.

"However, relying on humans can be time consuming, expensive, and prone to errors. There isn't really any technology out there that can be used to automatically extract a legal description from a deed. Most organizations that need to do this send the work out of the country to be keyed—and its complexity increases the chances for errors. If you screw up and list a piece of property as 1,100-feet wide, compared to 11,000 feet, that can be a big deal. There is a whole industry that has developed related those kinds of mistakes; it's called mortgage insurance, and it's designed to ensure that buyers get what they think they are getting."

Axis is especially attuned to working with real estate documents, as, in a previous job, Axis' SVP of product management and R&D architected a document imaging system for more than 5.5 billion documents related to U.S. properties. "Axis has created a solution designed to think like humans," said Ells. "You know how you can hand a person stack of documents and ask them to read it over and find information. Our software is designed to work the same way.

"Initially, we are going to focus on the vertical markets we have the most experience in. These include energy, mortgage and title, and healthcare. They all face the challenge of unstructured documents, and we want to leverage our subject matter expertise. We don't want to dilute our focus and our ability to do a great job in those industries."

How the software works

Founded in 2002, Axis provides IT advisory services focused in the aforementioned industries. Its customers include big names like **Alcoa, Lending Tree, TransAmerica, Green Dot**, and **Yamaha**. Axis' business involves working with developers in India—one of whom has developed natural language processing (NLP) technology that is now being brought to market as Axis AI (Artificial Intelligence).

Axis AI works on textual information. The software can receive documents from a variety of input sources, including document capture systems, as long as they are TIFFs to which it can apply OCR. "The software understands language in context—what's a noun, what's a verb, what comes before or after a certain phrase, etc.," said Ells. "It reads all the words and weighs them. It does not use image or shape recognition—which isn't very practical for the text intensive documents that we are focusing on."

Because of this contextual understanding, Axis AI is currently optimized only to work with English language documents. Ellis noted, however, "The core technology is mathematics and pattern matching, so it will recognize patterns of any language and successfully extract

a pretty good amount of the data. But, the technology also relies on facets of words and sentences; for example, names, dates, grammar, etc. All these facets would have to be converted to support a different language other than English."

The classification demo I saw was fairly fast and straightforward. Ells fed the application two sample sets of different document classes and then a batch of several hundred documents that included three classes, including the two that were sampled. The application successfully classified all the documents that fell under the two sampled classes and let the rest unclassified. Once a third sample set was entered, the remaining documents were classified under that.

It took only a few minutes to complete the entire process—from training to classification. Ells did note that Axis AI doesn't necessarily do auto-separation, and the batch he uploaded was already separated into individual document files.

To train Axis AI for extraction, Ells worked with each document class separately. For energy bills, which all had a similar structure, he merely highlighted the account numbers and dates (which were the two fields he wished to capture) on a few samples, and Axis AI was able to automatically capture them from the rest of the bills in the batch. A similar process was used to capture property descriptions in a second class of documents—but the difference was that the descriptions varied in location, format, and length. Nonetheless, less than 10 samples delivered a 100% success rate in extraction from more than 100 forms in the batch.

The third class of documents was a release form, from which a name or set of names had to be extracted. The same highlighting process was used, but on the first try the desired data could only be found on 87% of the forms. An exception process was executed through which it was discovered that the missed names all appeared in a similar type of table format. A couple examples were captured and the documents were re-run to a 100% success rate.

Although the set-up and training didn't appear overly challenging, Ells indicated that Axis would prefer to perform these steps for the customers. "We are basically asking the customers to send us samples of their documents with the data that they want to capture highlighted," he said. "We will train the system for them."

He indicated the in addition to the NLP inherent in the product, Axis can set up rules or parameters to increase accuracy. "When looking for fields on its own, the software can weigh information like the fact that a property description always starts with a certain set of words," Ells. "But, if you know a certain field is only going to appear in a specific alphanumeric format, we can set up rules to assist the software. We can also do look-ups against databases of known values."

Document Imaging Report
Business Trends On Converting Paper Processes To Electronic Format

DIR is the leading executive report on managing documents for e-business. Areas we cover include:
1. Document Capture
2. Image Processing
3. Forms Processing/OCR/ICR
4. Enterprise Content Management
5. Records Management
6. Document Output
7. Storage

DIR brings you the inside story behind the deals and decisions that affect your business.

Vol. 25, No. 4

Editor: Ralph Gammon
4003 Wood Street
Erie, PA 16509
PH (814) 866-2247
FX (412) 291-1352
ralphg@documentimagingreport.com

Managing Editor:
Rick Morgan
PH (814) 866-1146
rickm@scandcr.com

DIR is published 23x per year, on the 1st & 3rd Fridays of the month, by:
RMG Enterprises, Inc.
4003 Wood Street
Erie, PA 16509
PH (814) 218-6017
http://www.documentimagingreport.com

There are also standard data extraction features like confidence level feedback. "We also have the ability to untrain the software if a user is getting false positives," said Ells. "The software may include AI and NLP, but it's not magic. We own the code so we can always go back to the code level and fix something."

Ells stressed that Axis is looking for customers with high volumes of documents that it would consider as long-term partners. "We're not going to sell the software to someone, say good luck, and be on our way," he said. "And unlike most software vendors, our business model does not include selling professional services to supplement our software license sales.

"Our plan is to charge our customers based on the number of guaranteed fields of data they are receiving from us. The set-up fees will be incorporated in our price.

"And we don't plan on making customers buy a certain volume up front. Rather, they will only pay for what they use. This helps make our solution scalable—especially when compared to dealing with people. It's very easy to scale our software up and down vs. having to train new workers every time you have a spike in volume, let them go, and then train another new set when your volume spikes again."

For more information:
http://axistechnical.com/advanced-data-extraction/data-extraction-axis-ai/