athento
understanding documents

Digitization and Smart Capture of Documents

August 2013

Being able to keep your documents accessible from any point on the planet, and being able to use the information contained in those documents, has become critical for many businesses. To do that, you first need to have the documentation stored in electronic format. The digitization of documents is the process which converts paper documents into digital formats (by using scanners and other hardware).

However, **digitization, by itself, isn't of much use to businesses.** Having thousands or millions of documents in a filing system, or cross-referenced in a database, isn't sustainable, especially because recovering a determined documents turns into an overly-complex process. That's where document management systems come in: they maintain control of, and accessibility to, those documents.

Capture is the process by which digitized documents are sent to the document management or ECM system.

Additionally, even sending digitized documents to the document management system is still too much work for the users to do. Tasks such as naming documents in the system and adding metadata which permit descriptions of the content and make subsequent searches easier; or saving document in a pre-determined location according to their type; or initiating work flows. This work **can be made easier by smart capture**, which automates certain tasks such as data extraction and recognition of document types or document classification.

In Figure 1, you can see the complete process of smart document capture. Following that, we'll describe this process, step by step, once documents have been digitized:

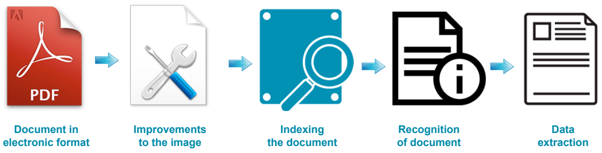


Figure 1. The process of smart document capture.

1. Obtaining the documents in electronic format

This process is carried out by connected scanners to the system. In Athento's case, it's possible to scan a document from the platform. It's also possible to capture vast amounts of documents. This can be done in two ways:

- **Massive loading of documents from the platform.** It's possible to upload various documents to the platform by selecting them from a local disk (see Figure 2). These documents can be processed in a pre-programmed or an immediate way.

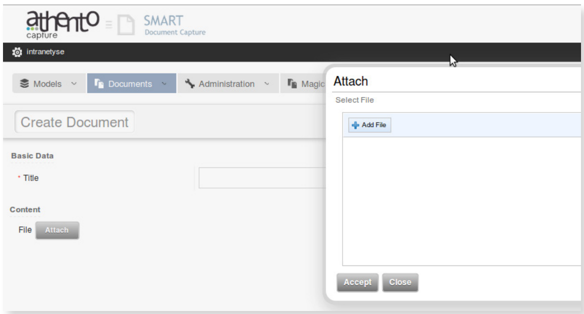


Figure 2: Simultaneous capture of various documents

- **Hot Folder:** It's possible to connect Athento to a folder so that it can be monitored (so that each time a document is added via the scanner, Athento can process it.) This allows the capture process to take place without requiring one person to be in charge of it.

With Athento's smart capture, it's not necessary to scan each document separately to produce separate files for each document.

Documents can be scanned in lots or batches which the system then automatically separates into individual documents.

Other systems allow you to carry out this operation by introducing bar codes for each document in the batch that has been scanned. Athento can also read bar codes, but given that this process is costly for users (who have to generate the bar codes, print them out and stick them onto each document), Athento offers the possibility of carrying out the division of documents by analyzing the structure of the documents and identifying types of documents. That way, users don't have to intervene in the process.

2. Improvements to the image

The idea behind improving the image is to make sure that scanned documents have the necessary quality to be processed and stored.

When we scan or digitize a document, we come up against quality defects such as documents that are not in the correct position (they're off-center) or which have white or black borders. One of the outstanding features that Athento has is the capability to correct defects in digitization such as (see Figure 3):

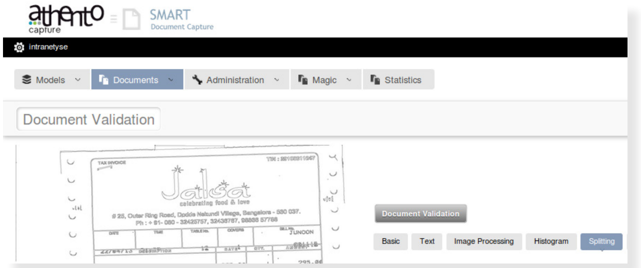


Figure 3: Options for improving the image

- **Rotation:** Re-orients the document by turning it a certain number of degrees. Can be done automatically or manually.
- **Binarization:** Applies improvements in contrast to the image.
- **Dynamic threshold:** Converts an image into black and white.
- **Borders:** Cuts off unwanted black or white borders. Can be done automatically or manually.
- **Flipping:** Allows users to flip the position of documents as if they were looking at a mirror image.
- **Smoothing:** refers to the reduction of noise in the image, such as “salt-and-pepper noise” -- those little black dots that sometimes appear after you’ve scanned an image.

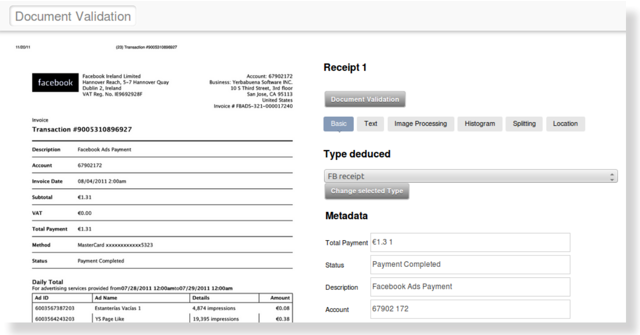
3. Indexing documents

With most ECM systems, only the title and the description of the document are indexed. This means that the title and a document description are fields into a database which then allows users to find the document (by using the database.) Normally, the indexing of the title (but not the description) is “full-text”. That means that, to find the document by its description, we have to look for the complete description exactly as it was entered into the database.

In contrast, when the title is fully indexed, you can search for the document by looking for any of the words contained in the title. Athento goes one step further: thanks to its OCR, Athento indexes each word of the content of the document and saves it in a database so that we can search for documents according to the words that the document contains.

4. Recognition of the document

Athento can be trained to recognize different types of documents. This recognition is carried out by using the application together with various different technologies. For example:



11/20/11 (23) Transaction #9005310896927

facebook
Facebook Ireland Limited
Hannover Reach, 5-7 Hannover Quay
Dublin 2, Ireland
VAT Reg. No. IE9692928F

Account: 67902172
Business: Yerbabuena Software INC.
10 S Third Street, 3rd floor
San Jose, CA 95113
United States
Invoice # FBADS-321-000017240

Invoice
Transaction #9005310896927

Description	Facebook Ads Payment
Account	67902172
Invoice Date	08/04/2011 2:00am
Subtotal	€1.31
VAT	€0.00
Total Payment	€1.31
Method	MasterCard xxxxxxxxxxxx5323
Status	Payment Completed

Daily Total
For advertising services provided from07/28/2011 12:00amto07/29/2011 12:00am

Ad ID	Ad Name	Details	Amount
6003567387203	Estanterías Vacías 1	4,874 impressions	€0.08
6003564243203	YS Page Like	19,395 impressions	€0.38

Type deduced
FB receipt

Metadata

Total Payment	€1.3 1
Status	Payment Completed
Description	Facebook Ads Payment
Account	67902 172

Figure 4: Data extracted from an invoice

- **Neural networks:** The system is shown a sample of a document of a determined type. By using neural networks, Athento compares the structure of the captured documents with those that belong to the samples, and assigns it a percentage of similarity
- **Histogram:** The system is shown a sample of a document of a determined type so that it can analyze the color structure. From then on in, any captured document is compared with that colour structure o histogram, and is assigned a percentage of similarity.
- **Regular expressions:** Athento can search for the appearance of certain terms, phrases or numbers associated with a document type.

For example, when the system sees the words "Invoice" or "Tax Registration Number," those are indications that the document type of this document would be an invoice.

This step is very important because it allows the system to keep documents in a specific location or to start a work flow for revision or approval.

5. Data extraction

By taking a sample document and indicating the location of the data to be extracted, Athento can obtain data from the content of the said document. Figure 4, for example, shows the data extracted from an invoice.

These data can be validated by a user in order to guarantee that the extraction is 100% correct.

Additionally, the way to indicate the location of these fields is done in a completely user-friendly way, designing a template visually in the system.

This template can be designed by any user, without any technical knowledge required. Other mechanisms can also be used for extracting data from the document, in the case of unstructured document types.

Once the data has been obtained, it is sent as metadata to the preferred document management system or ECM platform.

As with the case of separating documents, other products use the barcode method to extract data beforehand and codify them with barcodes which are attached to document, so that the system can read them and use them as desired. Again, we should point out that this process is costly for users: although some ERPs have the capability of generating barcodes, any documents that come from outside the organization, this process of generating and sticking barcodes has to be done manually.

Processes such as the digitization of individual documents, done separately, the classification of document or the extraction of data are costly in terms of the time for employees of the business who have to do these tasks.

Smart capture of documents solves this problem by automating these processes.

Athento and its Smart Document Capture can be integrated into any document management system or ECM platform which uses the CMIS standard.

athento

understanding documents

askourteam@athento.com

Twitter: @yerbabuenasoft

www.athento.com