athento
understanding documents

Digitalización y Captura Inteligente de Documentos

Mayo 2013

Poder mantener accesibles los documentos desde cualquier punto del planeta y utilizar la información contenida en ellos se ha vuelto crítico para muchas empresas. Para ello, en primer lugar, se requiere tener la documentación en formatos electrónicos. La digitalización de documentos es el proceso por el cual, a través de escáneres y otro hardware, se convierten documentos en papel a formatos digitales.

Pero **la digitalización por sí sola no es de gran utilidad para las empresas.** Tener miles o millones de documentos en un sistema de ficheros o referenciados en una base de datos no es sostenible, sobre todo porque recuperar un determinado documento se vuelve un proceso demasiado complejo. Es allí en dónde aparecen los **sistemas de gestión documental** que mantienen el control y la accesibilidad de los documentos.

La captura es el proceso por el cual los documentos digitalizados son enviados al sistema de gestión documental o ECM.

Por otro lado, aún enviando los documentos digitalizados al sistema de gestión documental queda demasiado trabajo por hacer para los usuarios. Trabajo como nombrar documentos en el sistema y añadir metadatos que permitan describir su contenido y faciliten las posteriores búsquedas, o como guardar los documentos en una ubicación determinada según su tipo o iniciar flujos de trabajo. Este trabajo **puede ser facilitado por la captura inteligente**, que automatiza algunas tareas como la extracción de datos y el reconocimiento de tipos documentales o clasificación de documentos.

En la ilustración 1 puede verse el proceso completo de captura inteligente de documentos. A continuación, vamos a describir este proceso paso a paso, una vez digitalizados los documentos:

Ilustración 1: Proceso de Captura Inteligente.

1. Obtención de documentos en formato electrónico.

Este proceso se lleva a través de la conexión del sistema con escáneres. En el caso de Athento, es posible escanear un documento desde la plataforma. También es posible capturar grandes cantidades de documentos. Esto se puede hacer mediante dos mecanismos:

- **Carga masiva de documentos desde la plataforma:** Es posible subir varios documentos a la plataforma seleccionándolos desde un disco local (Ver la ilustración 2). Estos documentos pueden ser procesados de forma programada o de manera inmediata.

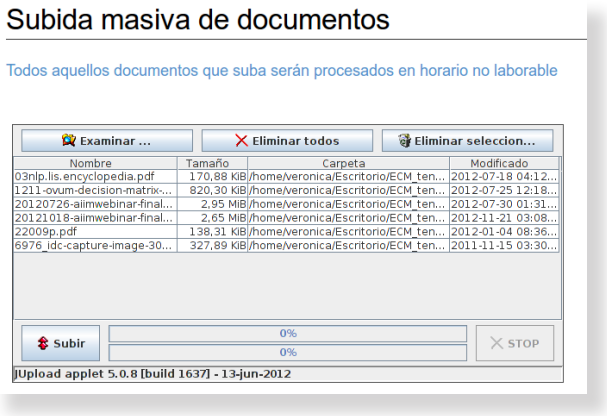

Ilustración 2: Captura de varios documentos al mismo tiempo.

- **Hot Folder:** Es posible conectar Athento a una carpeta para que la monitorice. Es decir, para que cada vez que un documento sea añadido por el escáner Athento lo procese. Esto permite que ninguna persona tenga que ocuparse del proceso de captura.

Con la captura inteligente de Athento, no es necesario que cada documento se escanee por separado para producir archivos únicos por documento.

Pueden ser escaneados en lotes de documentos que el sistema separará en documentos individuales de forma automática. Otros sistemas permiten realizar esta operación mediante la introducción de códigos de barras por cada documento escaneado dentro del lote. Athento puede leer también códigos de barras, pero dado que este proceso es costoso para los usuarios (que tienen que generar los códigos de barras, imprimirlos y pegarlos a cada documento), Athento ofrece la posibilidad de llevar a cabo esta división de documentos analizando la estructura de los mismos e identificando su tipología, de forma que los usuarios no tengan que intervenir.

2. Mejoras de la imagen

Mediante las mejoras de la imagen se busca que los documentos escaneados tengan las características de calidad necesarias para su almacenamiento y procesado.

Cuando escaneamos o digitalizamos documentos, nos encontramos con defectos de calidad como por ejemplo que los documentos no se encuentran en una posición correcta (no están rectos), tienen bordes negros o blancos, etc. Algunas de las posibilidades que brinda Athento para corregir estos defectos de calidad de la digitalización son (Ver ilustración 3):

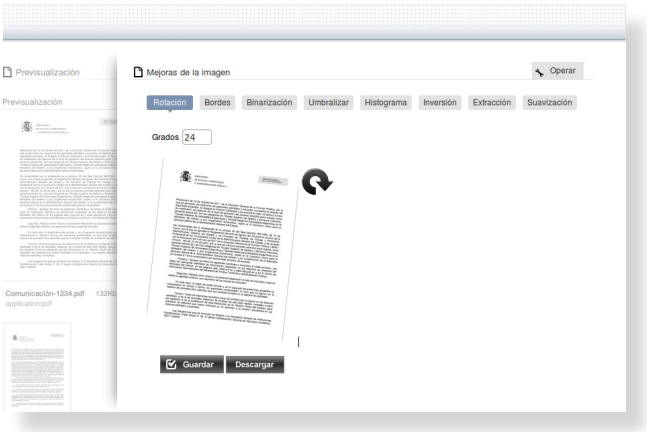

Ilustración 3: Opciones de mejora de la imagen

- **Rotación:** Se trata de re-orientar el documento, aplicando grados de rotación al mismo. La re-orientación del documento puede hacerse de forma automática o manual.
- **Binarización:** Aplica mejoras de contraste a la imagen
- **Umbral dinámico:** Convierte una imagen a blanco y negro.
- **Bordes:** Recortes de bordes blancos o negros no deseados. Puede realizarse de forma manual o automática.
- **Inversión:** Nos permite girar la posición de documentos como si se tratase de una imagen espejo.
- **Suavización:** Se refiere a la reducción de ruido en la imagen, por ejemplo el ruido "sal y pimienta", que son aquellos puntitos negros que aparecen en ocasiones tras escanear o digitalizar una imagen.

3. Indexación del documentos

En la mayoría de sistemas ECM sólo el título y la descripción del documento son indexados. Esto quiere decir que el título y su descripción se introducen en una base de datos para que mediante consultas a la misma, el documento pueda ser encontrado. Normalmente, la indexación del título es "full-text", pero no así la de la descripción. Esto significa que para encontrar un documento por su descripción, tenemos que buscar por la descripción completa del documento tal cual se introdujo.

En cambio, en el caso de la indexación full-text del título, podemos encontrar el documento buscando por palabras incluidas dentro de él. Athento va más allá. Gracias a su OCR, Athento indexa cada palabra del contenido del documento y lo guarda en una base de datos para que podamos buscar un documento por las palabras incluidas en su contenido.

4. Reconocimiento del documento

Athento puede ser entrenado para reconocer tipos documentales. Este reconocimiento se realiza mediante la aplicación conjunta de diversas tecnologías, por ejemplo:

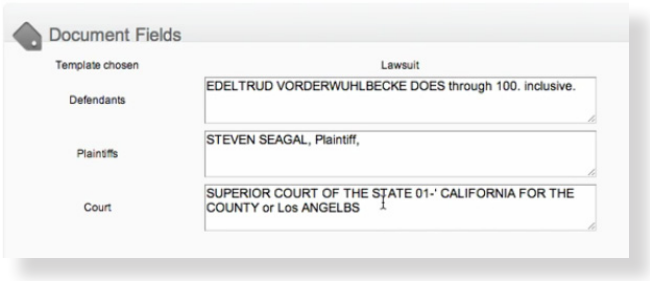

Ilustración 4: Datos extraídos de un documento de demanda

- **Redes neuronales:** Al sistema se le enseña una muestra de documentos de determinada tipología. Mediante las redes neuronales, Athento compara al estructura de los documentos capturados con aquellos pertenecientes a las muestras y arroja un porcentaje de similitud.
- **Histograma:** Al sistema se le enseña una muestra de un documento de cierta tipología para que analice su estructura de color. En adelante, cualquier documento que se capture se compara con dicha estructura de color o histograma y se arroja un porcentaje de similitud.
- **Expresiones regulares:** Athento puede buscar la aparición de ciertos términos, palabras, frases o números asociados con una tipología.

Por ejemplo, para el sistema la aparición de un CIF y/o la palabra “Factura” es un indicio para considerar que el tipo documental de ese documento es una factura.

Este paso es muy importante, ya que permite que el sistema pueda guardar por sí mismo los documentos en una determinada ubicación o iniciar un flujo de trabajo de revisión o aprobación.

5. Extracción de datos

Indicando la ubicación de los datos que se quieren extraer en un documento de muestra, Athento puede obtener datos del contenido de dicho documento. Por ejemplo, en la ilustración 4 se muestran los datos extraídos de una demanda legal.

Estos datos pueden ser validados por un usuario para garantizar que la extracción ha sido 100% correcta.

Por otro lado, la forma de indicar la ubicación de estos campos se hace de forma totalmente user friendly, diseñando de manera visual una plantilla en el sistema.

Esta plantilla puede ser diseñada por cualquier persona, sin ningún conocimiento técnico requerido. Otros mecanismos pueden ser utilizados también para la extracción de datos de documentos, en el caso de tipos documentales desestructurados.

Una vez obtenidos los datos, estos son enviados como metadatos al sistema de gestión documental o plataforma ECM que se prefiera.

Al igual que en el caso de la separación de documentos, otros productos usan el método de extraer los datos previamente y codificarlos en códigos de barras que se pegan a los documentos para que el sistema pueda leerlos y hacer con ellos lo que se requiera. Pero de nuevo, hay que decir que este proceso es costoso para los usuarios. Aunque algunos ERPs son capaces de generar códigos de barras, para los documentos que provienen de fuera de la organización este proceso de generación de código de barras y pegado sobre los documentos debe hacerse de forma manual.

Procesos como la digitalización de documentos individuales por separado, la clasificación de documentos, o la extracción de datos son costosos en términos de tiempo para los empleados de las empresas que tienen que realizarlos.

La captura inteligente de documentos soluciona este problema automatizando estos procesos.

Athento y su captura inteligente de documentos puede ser integrada con cualquier sistema de gestión documental o platorma ECM que soporte el estándar CMIS.

askourteam@athento.com
Twitter: @yerbabuenasoft
www.athento.com