Repositorio Institucional

Procesos de Digitalización Documental

Documento Base para la Digitalización de Documentos

Indice

Introducción

El presente documento tiene como objetivo introducir los **fundamentos y conceptos necesarios de los procesos de Digitalización Documental** de modo tal que, estos reflejen con fidelidad la información contenida en esos documentos de índole administrativa o académica de esta Universidad.

El proyecto de digitalización debe justificar la necesidad de digitalizar cualquier material, basándose en los beneficios que otorgará según las actividades a realizarse por los usuarios; por ejemplo, el aprendizaje, la enseñanza y la investigación. Y, por otro lado, garantizar las necesidades del organismo mediante una mejor conservación de los originales y para generar copias de seguridad o respaldo de documentos de archivo con valor histórico, científico o cultural, y habilitarlos en sustitución de los originales en caso de que se presenten siniestros que afecten a los archivos o documentos de conservación permanente.

Los procesos de digitalización implican que sean trabajados y abordados desde dos grandes aristas:

- ➔ El proceso de captura de la información en sí, en cuanto a las referencias técnicas de escaneado del documento.
- ➔ El proceso de guarda y de resguardo de esos documentos digitalizados en un repositorio documental seguro y accesible.

Universidad Nacional del Litoral.

Secretaría General.

Dirección de Informatización y Planificación Tecnológica.

-Santa Fe, octubre de 2017.-

¿Qué es el Modelo de Gestión de Documentos?

Modelo de Gestión de Documentos es el esquema teórico que debe facilitar la comprensión y homogénea implantación de la gestión documental de una organización.

Este Modelo de Gestión de Documentos y de administración de archivos es de aplicación a todo tipo de documentos, en cualquier formato o soporte, creados o recibidos por una organización o institución, pública o privada, en el curso de sus funciones o actividades, durante toda la vida del documento y, por lo tanto, para toda clase de archivos.

El uso del término documento, en el contenido y el contexto de este Modelo, se referirá siempre como sinónimo del término documento de archivo, es decir, el testimonio material de un hecho o acto realizado en el ejercicio de sus funciones por personas físicas o jurídicas, públicas o privadas, de acuerdo con unas características de tipo material y formal.

La Gestión Digital Documental

La Gestión Digital Documental es el conjunto de normas técnicas y prácticas usadas para administrar el flujo de documentos de todo tipo en una organización, permitir la recuperación de información desde ellos, determinar el tiempo que los documentos deben guardarse, eliminar los que ya no sirven y asegurar la conservación indefinida de los documentos más valiosos, aplicando principios de racionalización y economía. .

Las TIC impulsan esta gestión de construcción o representación de Modelos de Gestión Digital Documental que:

- ➔ Aceleran los tiempos de recuperación de la información contenida en esos documentos digitalizados.
- ➔ Transparentan la fiabilidad del contenido de los mismos, esto es incorporando tecnologías que garanticen que los mismos no han sido adulterados en el proceso de digitalización.

➔ Posibilitan la guarda y almacenamiento de los mismos cronológicamente, ya que el formato papel es sujeto a daños o deterioros por agentes externos o por el paso del tiempo.

El proceso de captura de la información

Una de las patas fundamentales en gestionar Modelos Digitales Documentales con las TIC es el proceso de captura del documento.

Se entiende como **proceso de captura** al conjunto de técnicas empleadas en el traspaso de la información contenida en un formato analógico como el papel a un soporte digital. Es en la calidad de la captura del documento papel en donde reside el éxito de una gestión digital optima.

Esquema de la captura de documentos

Llamamos **“captura de documentos” al proceso por el que se convierte un formato “papel” a un formato “digital”.**

Esta captura se realiza generalmente mediante un dispositivo de entrada llamado escáner, una aplicación de software que ayuda a clasificar e indexar el papel como un Repositorio Digital Documental.

Procesos de digitalización

El proceso de digitalización consiste en la captura de los documentos que se encuentran en formato papel. Al definir este proceso deben tenerse en cuenta los siguientes aspectos:

1. Conseguir el escáner adecuado para el trabajo, basándose criterios como tipos de documento y requerimientos de proceso.
2. Fiabilidad, soporte técnico, calidad de imagen y capacidad de alimentación son factores a tener en cuenta para elegir un buen escáner. Las características técnicas a tener en cuenta serán: tamaño del sensor, lentes, perfil de color.

3. Definir un modelo documental de información adecuado que pueda representar con mayor fiabilidad y confianza el contenido y la metainformación de los documentos digitalizados.
4. Trabajar con el gestor documental en función de las necesidades de cada organización o de cada proyecto documental.

Control de calidad

El control de calidad es un elemento importante en cada una de las etapas de un proyecto de digitalización. Sin este trabajo no será posible garantizar la integridad y consistencia de los ficheros de imágenes. El control de calidad se necesita para revisar los documentos digitalizados y así asegurarse de que cumplen el estándar del original.

Recuerde que al digitalizar un documento con las herramientas de captura adecuadas, con el dispositivo hardware escáner y el software de procesamiento óptimos, puede realizar un mejora sustancial en el realce y la calidad de la imagen.

Las características a tener en cuenta en un buen módulo de control de calidad son:

1. Que disponga de opciones como añadir, insertar, reemplazar, girar, voltear y borrar páginas de un documento.
2. Que permita la edición de las imágenes, con opciones como alineamiento, eliminar bordes negros, cortar, recortar, cambiar la orientación de la página, limpiar puntos negros, etc.
3. Que admita la revisión de imágenes automática, vistas en miniatura, reordenación de páginas y zoom de las imágenes.

Herramientas de Digitalización

Desde la Dirección recomendamos los siguientes programas para realizar las tareas de post-procesamiento antes descriptas:

ScanTailor[1]: Herramienta para el alineamiento de hojas, eliminar bordes negros, cortar, recortar, cambiar la orientación de la página, limpiar puntos negros y otras

[1] http://scantailor.org/

deformaciones que pueden ocurrir en el proceso de captura. Es de software libre y puede utilizarse tanto en sistemas operativos Windows como Linux.

PDFSam[2]: Esta herramienta permite la edición de archivos pdfs, reordenación de páginas, rotarlas, unir pdfs, etc. Es de software libre y puede utilizarse tanto en sistemas operativos Windows como Linux.

Como alternativa también se puede utilizar:

PDFShuffler[3] que posee características similares aunque solo puede utilizarse en sistemas operativos Linux.

gImageReader[4]: Es una interfaz simple realiza tareas de Reconocimiento Óptico de Caracteres (OCR es su sigla en inglés) utilizando el motor tesseract. Permitiendo generar archivos pdf con los caracteres reconocidos, o exportar este reconocimiento a archivos hORC. Es de software libre y puede utilizarse tanto en sistemas operativos Windows como Linux.

El proceso de guarda y de resguardo de los documentos digitalizados

Desde la Universidad se ofrece el servicio de gestión digital documental a partir del uso del repositorio institucional, proyecto RDDI, con tecnología Nuxeo.

La plataforma de gestión documental es donde se almacena y se resguarda la información digitalizada, obtenida de los procesos de gestión digital documental. Es en función de la gestión de esta herramienta desde donde se propone un Modelo de Gestión Digital de Documentos en el ámbito de la UNL.

[2] http://pdfsam.org/
[3] https://pdfshuffler.sourceforge.io/
[4] https://github.com/manisandro/gImageReader

Tipos de soporte de textos

Las fuentes de los textos a digitalizar pueden estar alojadas en diferentes soportes.

- ➔ **Textos manuscritos sobre hojas sueltas o encuadernadas**. Es el caso, por ejemplo, de los típicos apuntes escolares. Las soluciones de reconocimiento automático de texto en este caso son nulas. Lo máximo que se puede obtener por ahora es una imagen digitalizada que puede ser leída por el usuario (siempre que su visión se lo permita, claro está), ya que la función de reconocimiento de texto la realiza el cerebro de la persona que lo lee.
- ➔ **Textos impresos en hojas sueltas o encuadernados**. Suelen ser la mayoría de los casos. Varían desde una simple carta, factura, etc., hasta un libro convencional. La herramienta más importante a tener en cuenta en este caso es el escáner o la cámara con los que obtener la mejor imagen posible a digitalizar.
- ➔ **Textos impresos con textos manuscritos.** Es una combinación de los textos anteriores. Se podrá obtener un reconocimiento automático sobre el texto impreso y
- ➔ **Textos impresos con gráficos**. Son documentos que pueden contener imágenes con o sin texto impreso y dichas imágenes poseen una calidad fotográfica. Por lo que es de interés que el documento digitalizado posea una calidad similar.
- ➔ **Archivos de texto.** Estos últimos hace referencia a los archivos digitales. Si tenemos acceso a los mismos podemos obviar el proceso de digitalización del documento impreso.

Proceso de Captura

Es en este proceso de captura, donde debe establecerse la calidad de digitalización. Dependiendo del tipo de escáner, se puede definir la calidad configurando los DPI *“Dot Per Inch”* o PPP *“Puntos Por Pulgada”* ó la resolución de la captura establecida en pixeles.

Además se pueden configurar si la imagen digitalizada es en **blanco y negro** (imagen binaria), **escala de grises**, ó **color**.

Para **textos impresos y manuscritos** digitalizar con una calidad de **300 DPI** y en **blanco y negro.** Si bien recomienda configurar la captura en **escala de grises** y realizar el paso a **blanco y negro** en el post-proceso.

Para **textos con gráficos** digitalizar con una calidad de **600 DPI** en **escala de grises** o a **color,** dependiendo de los gráficos originales.

Todos estos parámetro van a influir en el tamaño de documento digitalizado. Siendo la digitalización de **300 DPI** en **blanco y negro** la de menor tamaño y la mayot la de **600 DPI** a **color.**

Elementos físicos de digitalización (Hardware)

- **Escáner de cama plana.** Los escáneres de cama plana son los más comunes, y se utilizan para copiar documentos, hojas sueltas, fotografías de diferentes tamaños, hasta un máximo de tamaño (generalmente una hoja de tamaño Letter, Legal u Oficio).
    - Se recomienda su uso para textos en hojas sueltas.

- **Escáner con alimentador automáticos de documentos o ADF (*Automatic Document Feeder*)**. Es un dispositivo que permite tomar varias páginas y alimentarlas hoja por hoja al escáner. Esto permite escanear documentos de múltiples páginas sin necesidad de colocar página por página. Inclusive existen otros alimentadores capaces de escanear de los dos lados de la hoja en una sola pasada.
    - Es el más recomendable para textos en hojas sueltas

➔ **Escáneres de libros.** Existen de distintos tipos pero en general digitalizan capturando imágenes con una cámara de fotos de las hojas del libro. Algunos tienen una única cámara para ambas hojas y otros una cámara por hoja.

- Se recomienda su uso para textos encuadernados y que no puedan ser deshojados para pasar por alguno escáneres de los mencionados anteriormente.

Post-Procesamiento

La gran variedad de dispositivos y software que se pueden utilizar para procesar documentación impresa u obtener el texto contenido en imágenes hace que sea imposible determinar unas pautas genéricas de uso. No obstante, existen factores comunes que influirán en la calidad de los resultados. En primer lugar, hay que tener en cuenta el soporte del texto a digitalizar, puesto que sus características, calidad, influirán decisivamente en los resultados. Por ejemplo, el texto obtenido tras digitalizar una hoja impresa con tinta de buena calidad será idéntico al original, pero el resultado de escanear un papel manuscrito será texto ininteligible. El tipo de papel también será determinante para la obtención de buenos resultados. Por ejemplo, para escanear revistas, folletos o manuales impresos en papel fotográfico o satinado, deberá prestarse atención al brillo, ya que normalmente será superior al que encontraremos en documentos procedentes de impresoras convencionales. En estos casos, la presentación de las propias publicaciones puede suponer un problema añadido, puesto que es posible que los programas de OCR no interpreten correctamente la disposición de los elementos en los documentos, así como la presencia de figuras, tablas complejas, etc. Es frecuente la necesidad de escanear fotocopias o documentos similares y es muy habitual que la calidad de las mismas no sea la idónea, ya sea porque no están bien hechas o porque la tinta de la fotocopiadora no sea de buena calidad. En estos casos, en la medida de lo posible deberá recurrirse a documentos originales. A la hora de escanear libros o material encuadernado es necesario prestar atención a la correcta colocación de los mismos sobre la superficie del escáner. Normalmente, la mayoría de programas de OCR permiten escanear dos páginas a la vez. No obstante, si la colocación es incorrecta o la propia encuadernación impide que toda la zona impresa entre en contacto con el cristal del escáner, se producirán pérdidas de datos.

Pasos sugeridos de post-procesamiento

1. renombrar las imágenes;
2. rotar las imágenes pares e impares;
3. alinear las páginas;
4. recortar los márgenes;
5. pasar a blanco y negro (opcional).

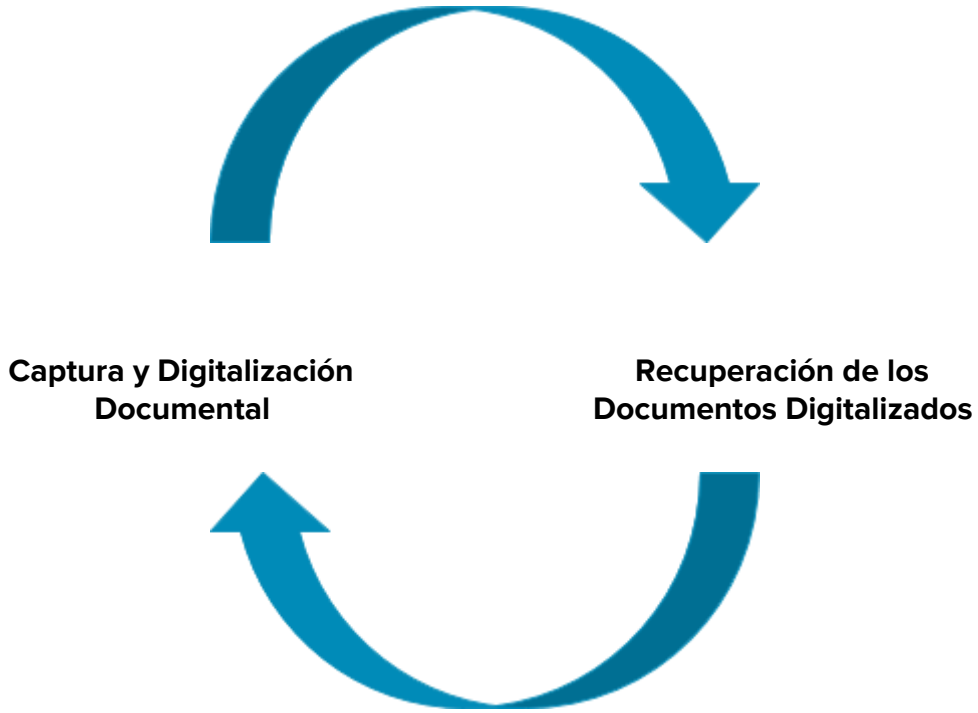
La herramienta sugerida, desde la dirección, para realizar estos pasos es **ScanTailor**

Caso práctico - Actas Históricas de Alumnos

Para ello, proponemos el siguiente modelo de trabajo:

1. Las oficinas de alumnados digitalizan las actas de exámenes y guardan esa imagen digital en formato .pdf
    a. La tarea de captura se debe realizar de acuerdo a las especificaciones técnicas que se registran en este documento.
    b. Estas especificaciones detallan la calidad de los archivos que se generarán, el tamaño, la escala de colores, entre otros.
2. Los miembros de las oficinas de alumnados pueden subir al espacio de trabajo asignado en el Repositorio Institucional de su U.A. los archivos generados identificándolos por U.A. y número de acta. Estos, la U.A. y el número de acta, será el identificador único que posea dicho documento.
3. Una vez que el documento se encuentre en el espacio de trabajo correspondiente, el Responsable del Dpto. de Alumnado deberá controlar la calidad de la captura del mismo y será el único usuario que podrá aprobar o rechazar la publicación definitiva de ese documento. Este proceso se avala firmando digitalmente el documento escaneado.
4. En la gestión de **“Control de Actas” desde el sistema SIU-GUARANÍ**, los operadores de los departamentos de los alumnados al realizar una gestión de alumno que requiera normativamente el control de las actas impresas, en la operación de impresión de la Historia Académica con el detalle de actas de un alumno, el personal de alumnado deberá poder recuperar desde el repositorio institucional el documento digitalizado, publicado y firmado digitalmente coincidente (identificado por UA y nro. de acta) y contrastar el contenido del mismo con la información que remite SIU-GUARANÍ.

Con este particular caso de ejemplo queremos representar que, en este proceso de digitalización documental, existen dos etapas bien definidas:



Entonces, desde el trabajo conjunto y acordado con estándares comunes y, por supuesto, con la normativa respectiva que lo avale, podremos establecer los procesos de resguardo y centralización de la información sensible que aún se encuentra almacenada en formato papel y que es necesaria para ejecutar acciones delicadas y costosas ya que la información se encuentra diseminada y, aún en un soporte que puede sufrir deterioros o alteraciones.