Guía de procedimientos para edición de documentos digitales

Autores:

Marcela Coria

Carolina Unzurrunzaga

Revisión:

Mariana Pichinini

Versión 0.1 Beta

La Plata,

Noviembre 2012

1. Generalidades

La puesta en línea de los archivos en el repositorio requiere un tratamiento propio de acuerdo con los requerimientos del software y de los criterios adoptados. Los criterios que se siguen en esta guía son los adoptados por Memoria Académica, el repositorio de la FaHCE-UNLP.

Principalmente, debemos saber que por cada documento a incluir son necesarios según la política descripta previamente, dos archivos en los siguientes formatos:

- .pdf que permite la descarga del documento, y
- .html que permite la puesta en línea del documento y/o búsqueda a texto completo.

1.1. Identificación de los documentos

Cada documento individual se identifica mediante el código alfabético correspondiente a la colección de pertenencia, seguido de un número secuencial correlativo. Esta identificación es unívoca y es la que llevan en su nombre los archivos de salida (doc, html, pdf, etc.), y las carpetas que los contienen. En Memoria Académica (en adelante MA) los identificadores correspondientes a las colecciones de Eventos, Tesis y Artículos de revistas son ev, te y pr, respectivamente. Es puntualmente necesario que el nombre y número asignado a los archivos html y pdf sea exactamente igual al nombre de la carpeta que los contiene.

2. Digitalización documento original en soporte papel

En el primero de los casos, cuando el documento llega a la biblioteca en formato papel y se agotan las posibilidades de conseguirlo en algún tipo de formato digital se dispone a la digitalización del mismo. Para ello, se sigue lo expuesto en la G*uía de Digitalización* confeccionada ad hoc por BIBHUMA.[1]

La digitalización se realiza escaneando tales documentos impresos siguiendo los criterios y especificaciones impuestos tanto por las cuestiones de preservación como de manipulación y edición de los posteriores archivos digitales. Las características que deben conservar e incluir los archivos resultantes de la digitalización son las siguientes:

- Resolución de escaneo: 600 dpi
- Si se trata de texto debe escanearse en blanco y negro. Si el orginal es antiguo y se pierden características propias en la digitalización blanco y negro, entonces se hace en escala de grises. Si el original contiene imágenes coloreadas el escaneo se hace a color.
- Por cada página se debe originar un archivo independiente.
- El formato de las imágenes obtenidas debe ser tiff.

Estas especificaciones se configuran desde el software empleado para el escaneo. Dependiendo del scanner que tengamos el programa para su gestión será diferente, y por consiguiente las formas de configuración. Sin embargo, debemos investigar el mismo a fin de ajustar las especificaciones necesarias.

Una vez que se han obtenido el total de imágenes como páginas consta el documento original se procede al proceso de Reconocimiento Óptico de Caracteres (OCR) con un programa que lo posibilite, a partir del cual se obtiene un archivo de texto, preferentemente con extensión .doc. Este archivo de texto obtenido puede conservar aún algunos errores de reconocimiento de caracteres que se deben revisar manualmente en un procesador de texto. Este paso es sumamente necesario cuando el archivo .html que se creará en base a este texto va a estar visible. En caso de que el .html resultante se vaya a utilizar solo para búsqueda interna, se puede obviar un análisis tan minucioso.

[1] Disponible http://www.memoria.fahce.unlp.edu.ar/memoria/envio-de-trabajos/guia-de-digitalizacion

3. Procesamiento documento original en soporte digital

Los documentos a incluir en las diferentes colecciones del repositorio pueden llegar al la biblioteca en diferentes formatos: .doc, .rtf, .pdf y en la minoría de los casos en .html.

Teniendo en cuenta lo dicho al principio de esta guía respecto a la necesidad de generar un archivo .pdf y un .html se describirán las prácticas necesarias para su obtención de acuerdo a esta diversidad de archivos que se puedan obtener desde las unidades generadoras de la facultad.

4. Generación de archivo en formato PDF

Si el original digital obtenido es un archivo .doc se exporta como archivo .pdf

Si el original está en .pdf pero formando parte de un pdf mayor (p.ej. conteniendo el número entero de la revista o las actas completas de un evento) es necesario extraer las páginas correspondientes a cada uno de los trabajos y guardar cada trabajo así extraído por separado. El pdf que sufrió las extracciones se guarda sin cambios, y con cada uno de los pdf generados se procede como si originalmente hubieran sido originales pdf independientes.

Si los archivos originales que se poseen son las imágenes obtenidas del escaneo, tales imágenes se compaginan formando un archivo .pdf por documento. Es decir, se crea un archivo .pdf que contenga todas las imágenes correspondientes a un mismo documento. Este archivo recientemente creado debe someterse al proceso de OCR provisto por el programa editor de pdf, este proceso permitirá que se puedan hacer búsquedas dentro del archivo. Por ultimo, debido al origen de este archivo creado a partir de imágenes queda de un peso demasiado grande para ser subido en línea, por ello también a través del editor de pdfs se baja la resolución del mismo.

Pasos según Adobe Acrobat 9 Pro

- Traspaso de .doc a .pdf:

Ir al menú Archivo > Imprimir y seleccionar el formato pdf.

- Extracción de páginas en archivo .pdf:

Ir al menú Documento>Extraer páginas; o Documento>Páginas>Extraer

- Compaginar un archivo .pdf desde varias imágenes:

Ir la menú Archivo > Crear PDF > Combinar archivos en un solo pdf y allí seleccionar los archivos que forman parte de este documento.
Tener en cuenta que conserven el orden correcto en relación al original.

- OCR en pdf:

Ir al menú Documento > Reconocimiento de texto OCR > Reconocer texto usando OCR

- Bajar resolución de pdf:

Ir al menú Documento > Reducir tamaño de origen

4.1. Enriquecimiento e identificación del archivo en formato PDF

Una vez obtenida la versión en formato pdf del documento se realizan los siguientes pasos propios del procesamiento de MA.
- Insertarle la carátula de MA correspondiente, que permite identificar al repositorio, al documento y la licencia con que este es distribuido
- Incorporarle los metadatos descriptivos en las propiedades del archivo, lo que facilita la indización de los buscadores
- Asegurar el archivo mediante una contraseña, esto hace que no se pueda modificar el contenido del documento y los metadatos asignados.

4.1.1. Insertar carátula

Todos los archivos .pdf para descarga deben llevar insertada la carátula de MA como primera página. En trabajos que contienen más de un .pdf, cada uno llevará su carátula también.

La carátula es otro archivo PDF de una sola página con formularios en forma de cajones de texto donde se ingresan los datos bibliográficos del documento y una cita sugerida.

Pasos según Adobe Acrobat 9 Pro

- Ir al menú Documento>Insertar páginas; o Documento>Páginas>Insertar, y seleccionar la carátula correspondiente según el tipo de documento.
- Editar los campos de autor, título, etc. Esto se trabaja como con cualquier editor de textos ya que los textos a modificar son todos editables.
- Se deben respetar los tamaños de letras que llevan las carátulas, pero si es necesario reducir unos puntos el tamaño de los títulos, proceder así: seleccionar el cajón de texto del título, hacer doble click sobre él y en la ventana que se abre modificar, dentro de la pestaña Aspecto, el tamaño de letra.
- Si se necesita ampliar la cantidad de líneas del texto (p.ej. para autores), cliquear dos veces sobre el cajón de texto previamente seleccionado. Abre una ventana donde se elige pestaña Opciones, y se tilda la opción Multilínea.
- Si se necesita mover un cajón de texto entero para componer mejor la página, usar la herramienta Campo de texto (ubicada en menú Herramientas o Documento, según la versión)
- Si se necesita mover la línea ubicada bajo el título, usar la herramienta Objeto de la Edición avanzada (ubicada en menú Herramientas o Documento, según la versión)

4.1.2. Insertar metadatos descriptivos en las propiedades

En el archivo .pdf correspondiente al documento a incluir se agregan ciertos metadatos descriptivos correspondientes a título, autor, temas y tipo documental del documentos en la sección propiedades del archivo (Menú Archivo > Propiedades).

Estos metadatos pueden ser obtenidos a través de la base de datos. En MA se ha confeccionado un modo de visualización especifico que permite que se muestren estos metadatos de la forma directa en que se deben volcar en las propiedades del archivo pdf.

Pasos según Adobe Acrobat 9 Pro

- Ir al menú Archivo --> Propiedades. Se abrirá una ventana.

- Completar los metadatos teniendo en cuenta las siguientes especificaciones:

***Título**:* Cargar el título del documento, en el idioma en que esté escrito el documento. Omitir ñ y acentos. Debe ser n y las letras sin acentuar.

***Autor/es**:* Cargar los autores del documento separados por punto y coma espacio(;). Ingresarlos en forma directa (Nombre Apellido). Omitir ñ y acentos. Debe ser n y las letras sin acentuar.

***Asunto**:* Cargar el tipo documental y tema del recurso separados por punto y coma espacio (;). Omitir ñ y acentos. Debe ser n y las letras sin acentuar. Empezar cada término con mayúscula.

***Palabras claves**:* Cargar los descriptores asignados al recurso separados por punto y coma espacio (;). Omitir ñ y acentos. Debe ser n y las letras sin acentuar. Empezar cada término con mayúscula.

4.1.3. Asegurar con contraseña

Una vez caratulado e insertos los metadatos, a cada archivo PDF se lo asegura contra copiado y adulteración mediante una contraseña.

Pasos según Adobe Acrobat 9 Pro

Ir a menú Avanzado o Documento, según la versión:
Seguridad>Ver configuración de seguridad>Seguridad con contraseña
O Proteger > Codificación con contraseña.

Configurar las siguientes variables:
- Compatibilidad: Acrobat 5.0 y posterior
- Componentes del documento que desea codificar: Codificar todo el contenido del documento.
- Permisos: Restringir la edición e impresión del documento.
- Impresión permitida: Baja resolución (150 ppp)
- Activar copiar texto, imágenes y otros contenidos.

5. Obtención de archivo html

5.1. Documento html para visualización en pantalla

Si el archivo original obtenido es .doc se deben seguir estos pasos:

- Descartando título, subtítulo, autorías, filiaciones y todo otro dato ya incluido en la base de datos y que no forma parte del trabajo propiamente dicho, seleccionar el resto, copiar y pegar en un archivo nuevo (en blanco) de un programa editor de html.
- Luego ir cotejando con el original digital o en papel los formatos finales de cada elemento (p.ej. citas, tablas), y hacer algunas otras modificaciones generales.
- Las citas al pie, por su parte, deben ser copiadas y pegadas todas con su número correspondiente al final del documento.
- Si es necesario, buscar y reemplazar caracteres especiales tales como comillas simples y dobles y guiones especiales, según se indica en el archivo caracteresespeciales.doc
- Finalmente, guardar el documento en formato de archivo .html

En el caso de que el archivo original obtenido es .pdf se debe exportar a un archivo .html y proceder con el formateo descripto recientemente.

5.1.1. Marcado de html en documentos extensos

Dada la extensión de algunos trabajos que se quieren visualizar en pantalla y pueden sobrepasar las 200 o 300 páginas, en pantalla se presenta un índice o tabla de contenido de manera tal de ir leyendo por partes, aunque el archivo html sigue siendo uno solo.

El html se prepara igual que como en artículos y eventos, pero se le agregan marcas de secciones a fin de que el software Greenstone pueda generar la tabla de contenido. Aquí hay que ver el índice del trabajo y evaluar cuál división es mejor (capítulo, parte, etc., o contribuciones en caso de obras en colaboración)

El marcado se puede hacer a medida que se va armando el archivo html o una vez que se ha terminado, y consiste en:

1. Agregar al principio del documento las etiquetas

```
<!--
<Section>
<Description>
<Metadata name="Title"> Tabla de contenido </Metadata></Description>
-->
```

2. Y al final del documento la etiqueta de cierre:

<!--
</Section>
-->

3. Marcar cada sección considerada como tal al evaluar el índice, agregando a su inicio las etiquetas
<!--
<Section>
<Description>
<Metadata name="Title"> Título de sección </Metadata></Description>
-->

reemplazando "Título de sección" por el nombre del capítulo o la parte que figura en el índice. Es muy importante respetar exactamente la forma de los títulos.

4. Y al final (tras las notas) la etiqueta de cierre:
<!--
</Section>
-->

5. Proceder con las restantes secciones de la misma manera hasta que todas queden marcadas en su inicio y su final, incluidos anexos, bibliografías, etc.

5.1.2. Insertar imágenes

En los html para pantalla, las imágenes (gráficos, fotos, etc.) deben ser insertadas como imágenes de archivo, es decir que hay que generar con cada una de ellas un archivo de imagen.
Se trabaja entonces copiando la imagen del original digital (sea éste .doc o .pdf), y pegándola o abriéndola en un editor de imágenes, p.ej. Paint. Una vez abiertas, se verifica su tamaño y su resolución, y se guarda como archivo de imagen en formato JPEG (para fotos y gráficos) o GIF (sólo para gráficos). La resolución de las imágenes debe ser la de pantalla (72 píxeles por pulgada), y el ancho no sobrepasar los 20 cm.

5.2. Documento html para indización (sin visualización)

En el caso de que se disponga incorporar solamente documentos en pdf, es posible, para no desaprovechar la funcionalidad del software Greenstone de recuperación en el texto completo del documento, generar un html al cual no se requiere formatearlo[2]. Solo se genera y se identifica.
Para ello es necesario no completar el metadato correspondiente para indicar que el .html no debe estar disponible para la visualización.

6. Formatos que se mantienen en el etiquetado html del documento en proceso

- variables tipográficas del texto (negrita, cursiva, subrayado, etc.)
- justificación del texto (a la izquierda, a la derecha, a ambos lados)
- marginación de párrafos con marginadores (pero no tabulaciones)
- viñetas
- imágenes insertadas como imagen

6.1. Estilos que hay que eliminar del etiquetado html de los documentos

- tipo de letra (fuente)
- tamaño de letra (cuerpo)
- interlineado
- líneas en blanco: agrega siempre 1 tras un salto de línea.
- tabulaciones y otros espacios en blanco
- formato de tablas
- celdas combinadas en tablas

[2] También se puede usar el documento pdf generado sin contraseña para la indización del texto completo, esta opción no se incluye en esta guía ya que Memoria Académica no la utiliza.