

Política de digitalización de Proyecto Humboldt Digital (ProHD)

v1.0.1

2020

Estado del documento

- 29-05-2020: redactado por Antonio Rojas Castro
- 29-07-2020: revisado por Grisel Terrón y Tobias Kraft

Introducción

La digitalización es “el proceso mediante el cual la información analógica se convierte en información digital, a través de un sistema informático y de unos dispositivos, el escáner óptico o la cámara digital, para su conversión en imagen o en texto” (Díez Carrera, 2012: 76). La imagen digital resultante se conoce como imagen “de barrido” o “ráster”. Desde un punto de vista técnico, la imagen digital obtenida con un escáner es un mapa de bits organizado en forma de retícula o rejilla y compuesto por píxeles.

Este documento tiene por objetivo facilitar la conversación entre los miembros de ProHD y servir como una política de digitalización. En ella se definen las decisiones, tareas, procesos y orden en que se llevará a cabo la obtención de digitalizaciones en la Casa Humboldt con el equipamiento adquirido en el marco del proyecto de cooperación entre la Oficina del Historiador de la Ciudad de la Habana y la Berlin-Brandenburgische Akademie der Wissenschaften.

El contenido y el orden de esta “Política de digitalización” es el siguiente: revisión de algunos conceptos básicos, descripción del equipamiento (escáneres y software), definición de calidad y el uso previsto en el contexto del proyecto, establecimiento de especificaciones de escaneado, caracterización de los formatos de ficheros que se espera obtener tras la digitalización, fijación de algunas convenciones sobre los nombres de los ficheros y, por último, exposición de las acciones y el orden que componen el flujo de trabajo que se realizará en la Habana.

Conceptos básicos

Resolución

La resolución de una imagen ráster indica el nivel de detalle: cuanto mayor sea la resolución, mayor será el nivel de detalle (Zhang y Gourley, 2008: 56). La resolución se mide contando el número de píxeles, es decir, un punto o cuadrado diminuto que ocupa una posición en una retícula ordenada. Cuantos más píxeles tenga una imagen, más próxima se considera al original.

La unidad de la resolución de una imagen ráster es “*pixel per inch*” (ppi) o bien “*dots per inch*” (dpi). Ambas expresiones se utilizan de manera intercambiable, aunque es más correcto hablar de “ppi” porque la segunda expresión refiere, en realidad, a la resolución de las impresoras. Así, pues, una resolución de 300 ppi significa que la imagen ráster tiene 300 píxeles por pulgada (*inch*).

Se suele distinguir entre una resolución de captura (*input*) y otra de salida (*output*). Cuando se captura una imagen con un escáner se define una resolución de captura. Por el contrario, cuando se visualiza en un monitor o bien se imprime en papel una imagen ráster se utiliza una resolución de salida. Por este motivo, es muy importante tener en cuenta la resolución de la pantalla con que se evalúa la calidad de una imagen ráster.

Profundidad de color

La profundidad de color (o resolución de la señal) es la cantidad de información en bits representada en un solo píxel (Zhang y Gourley, 2008: 58). Los datos digitales se componen de bits, que solo pueden tener dos valores: 0 y 1. El número de bits utilizados para definir los píxeles de una imagen digital determina la profundidad de color. Puesto que cada píxel tiene asignado un valor correspondiente a un color o tono, cuantos más bits tenga un píxel, mayor es la información que puede representarse y, por tanto, mayor será el número disponible de colores.

Los píxeles de una imagen ráster se crean mediante la combinación de los tres colores primarios (rojo, verde y azul, RGB) o canales de color. La profundidad de bit se expresa en “bits por píxel” (bpi) y es la suma de los bits representados en cada canal de color. Las profundidades de color más frecuentes para digitalizaciones son:

- 8 bits para imágenes en escala de grises o en color (16 colores disponibles).
- 16 bits para imágenes en color con una mayor cantidad de tonos (65.536 colores disponibles).
- 24 bits para imágenes consideradas “true color” (16.777.216 colores disponibles).

A medida que se incrementa la profundidad de color, es posible capturar mayor número de detalles y, por tanto, la calidad de la imagen es más elevada; sin embargo, el incremento de bits por píxel afecta a la resolución, el tamaño del fichero y la capacidad de compresión.

Compresión

La compresión es la reducción del tamaño de un fichero con el objetivo de procesarlo, almacenarlo o transmitirlo de manera óptima (Zhang y Gourley, 2008: 60). Los ficheros comprimidos son menores que su contrapartida sin comprimir; sin embargo, la técnica de compresión y el nivel de compresión pueden disminuir la calidad del fichero. La “ratio de compresión” mide la relación existente entre el tamaño del fichero sin comprimir y el fichero comprimido; algunos programas permiten al usuario establecer la ratio de compresión y, por tanto, controlar la calidad resultante.

Hay dos tipos de compresión:

- Sin pérdida (*lossless*): asegura que la información es preservada, ya que no se elimina información redundante. Si la imagen se descomprime, tendrá la misma calidad que el original sin comprimir.
- Con pérdida (*lossy*): los ficheros comprimidos resultantes son más reducidos pero a costa de perder información. Si la imagen se descomprime, no tendrá la misma calidad que el original sin comprimir.

Las *Directrices técnicas* de FADGI recomiendan la compresión sin pérdida para todos los usos y la compresión con pérdidas para usos específicos. En ambos casos, sin embargo, se deben evitar técnicas de compresión “que usan programas patentados o de propietario debido a sus problemas potenciales de sostenibilidad a largo plazo” (Rieger, 2016: 13).

Tamaño de los ficheros

Las dimensiones de las imágenes se suelen representar de manera convencional indicando la amplitud x la altura. Así, pues, una imagen de dimensiones 600 x 400 consiste en 600 columnas y 400 filas de píxeles, lo que equivale a un total de 240.000 píxeles. Para calcular el tamaño de un fichero se debe multiplicar el número total de píxeles por la profundidad de color en bytes; el resultado de esta multiplicación debe dividirse por 8 (Zhang y Gourley, 2008: 64-65). En síntesis:

tamaño de un fichero = (amplitud en píxeles x altura en píxeles x profundidad de color) / 8

Por ejemplo, si la imagen de 600 x 400 píxeles tiene una profundidad de color de 24 bits, el tamaño del fichero se calculará de la siguiente manera:

$$(600 \times 400 \times 24) / 8 = 720.000 \text{ bytes}$$

Para que esta cifra sea más fácil de entender, se puede dividir el resultado en bytes por 1.024 para obtener kilobytes (KB) y luego, de nuevo, por 1.024, para obtener megabytes (MG):

$$720.000 \text{ bytes} / 1.024 = 703,125 \text{ KB}$$

$$703,125 \text{ KB} / 1.024 = 0,69 \text{ MB}$$

Por tanto, si sabemos el tamaño del documento original en pulgadas y la resolución y la profundidad de color con las que se va a escanear, es posible calcular el tamaño final de la imagen digital según esta fórmula:

$$\text{tamaño del fichero} = [(amplitud \times ppi) \times (altura \times ppi) \times profundidad \text{ de color}] / 8$$

Según las *Directrices técnicas* de FADGI (2016: 81), “el tamaño del fichero de imagen, en términos de almacenamiento de datos, es proporcional a la resolución espacial (mayor resolución implica mayor tamaño de fichero para un tamaño de documento fijo) y al tamaño del documento escaneado (un documento mayor implica un tamaño de fichero mayor para una resolución espacial fija). Aumentar la resolución incrementa el número total de píxeles, y produce un fichero de imagen mayor. Escanear un documento mayor genera más píxeles lo cual aumenta el tamaño del fichero de imagen.”

A modo de recapitulación, para calcular el tamaño de un fichero, hay que tener en cuenta el tamaño del documento original, la resolución, la profundidad de color y la compresión. Además, si se desea obtener un documento multipágina, la extensión en número de páginas o folios el documento original también es relevante (más sobre los formatos de fichero multipágina en el apartado “Formatos de fichero”).

Formatos de ficheros

Los formatos de ficheros proporcionan un método estándar de organizar y almacenar los datos (Zhang y Gourley, 2008: 61). El uso potencial de un fichero depende del formato elegido: “A good object exists in a format that supports its intended current and future use” (NISO, 2007). Suelen distinguirse dos tipos de ficheros en función de su uso: ficheros de máster y ficheros de acceso. A continuación, se caracterizan con más detalle los usos de ambos tipos y los formatos más comunes.

Máster (o maestro)

Los ficheros de máster (o maestros) tienen dos usos: por un lado, sirven para archivar la información y preservarla a largo plazo; por el otro, sirven para producir archivos derivados con distintas finalidades y para migrar el contenido a nuevos formatos de ficheros. De esta manera, cuando una nueva tecnología emerge, se evita volver a escanear un documento. Es por este motivo que se

debe crear un fichero de máster con la mejor calidad posible y guardarlo en el formato más apropiado.

Las *Directrices Técnicas* de FADGI distinguen dos tipos de ficheros de máster:

- Máster archivístico o de preservación: representa la mejor copia producida con una escala tonal larga, una gama de colores amplia y retoques mínimos para ser neutrales al uso.
- Máster de producción: es un fichero creado a partir del máster archivístico y de una calidad equiparable que se diferencia por algunas correcciones estéticas.

El formato de fichero de máster más común y recomendado es TIFF (*Tagged Image File Format*) porque se trata de un formato estándar, no propietario, multiplataforma, que puede guardarse sin pérdida de información y con suficiente calidad para impresión y publicación. Además, TIFF puede contener varias páginas (es multipágina) e incorporar metadatos técnicos en la cabecera.

De acceso

También conocidos como ficheros de servicio, entrega, visualización o salida, los ficheros de acceso derivan de los ficheros de máster de producción y tienen como principal uso la visualización de la imagen digital en internet por medio de repositorios y archivos digitales. Aunque existen otros formatos de acceso, como PNG, JPEG 2000 o GIF, los formatos de ficheros de acceso más comunes son:

- JPEG (Joint Photographic Expert Group): adecuado para fotografías y pinturas con variaciones sutiles de tonos y colores; es compatible con técnicas de compresión con pérdidas.
- PDF (Portable Document Format): adecuado para documentos multi-página y para almacenar texto obtenido con tecnología OCR y, por tanto, *searchable*.

Ambos formatos de ficheros son estándares y pueden visualizarse en todos los navegadores web existentes actualmente a tamaño completo o bien como miniatura (*thumbnail*).

Equipamiento

Escáneres

ProHD ha adquirido a la empresa Zeutschel dos escáneres planetarios para llevar a cabo la digitalización de documentos en la Casa Humboldt de La Habana. Los escáneres planetarios tienen un sensor en el cabezal que se mueve captando la imagen línea a línea y se caracterizan por su capacidad de crear imágenes en

alta resolución. Son especialmente idóneos para escanear libros y documentos encuadernados, manuscritos, mapas y hojas sueltas, es decir, todo tipo de materiales delicados que no se pueden introducir en un escáner plano o de rodillo. Sin embargo, este tipo de escáneres demoran la captura de la imagen (ya que se requiere una acción humana para pasar las páginas), solo son eficientes con materiales que pueden mantenerse planos y requieren un software para corregir la curvatura del lomo.

Los escáneres planetarios de ProHD funcionan con una potencia eléctrica comprendida entre 110 - 230 V y una frecuencia de 50/60 Hz; asimismo, requieren una superficie estable, que no transmita vibraciones. Para su uso correcto y mantenimiento a largo plazo, las condiciones del espacio de trabajo son importantes: la temperatura de la habitación debe ser entre 18 y 35° C, mientras que la humedad del aire debe ser de un máximo de 80%, sin condensación.

La garantía de piezas de repuestos dura 2 años a partir de la instalación, siempre y cuando las condiciones de uso sean normales (10 horas máximo de uso por día). La garantía de las piezas de repuesto entra en vigor una vez Zeutschel recibe un informe de puesta en marcha completo y un escaneo de aceptación del Universal Test Target (UTT) después de la instalación.

OS15000 Advanced Plus

Escáner planetario fijo¹ que permite un modo de trabajo automático: posicionamiento automático del libro, apertura automática de la placa de cristal y descenso automático de las placas de soporte del libro después del escaneado, inicio automático del escaneado, presión del libro controlada electrónicamente, ajustable de forma continua en 5 pasos para la protección de los documentos.

Características principales:

- Formato de escaneo: 460 x 360 mm (>A3)
- Modo de escaneo: color, escala de grises, b/n; 4 canales, RGB- Color y Gris en total 43200 Píxeles
- Procesamiento de color verdadero: 42 bits de color / salida 24 bits de color; Escala de grises de 14 bits / salida Escala de grises de 8 bits; 1 bit b/n / salida 1 bit b/n
- Resolución: hasta 600 dpi
- Espesor máximo del libro: 100 mm
- Productividad: 8 escaneados a color a 300 ppi en tiff en 1 minuto
- Múltiples salidas: disco duro, red, email, FTP e impresora
- Dimensiones: 540 x 640 x 670 mm

¹<https://www.zeutschel.de/en/produkte/scanner/farbscanner/os-15000-advanced-plus.html>

- Superficie de instalación: 504 x 540 mm
- Requerimiento de energía: 100 - 240 V / 50 - 60 Hz
- Peso: 50 kg

Para utilizar el escáner planetario fijo, ProHD también ha adquirido un monitor Dell de 23" Widescreen TFT Display con una resolución de 1920x1080 y una ratio de 16:9, y una estación de trabajo compuesta por una torre Dell (Workstation OS12000/15000/16000 - Multilanguage - (GB - 1x1 TB SATA HDD / HX-4915.05) con sistema operativo Windows10 Pro (64 Bits).

zeta comfort

Escáner planetario móvil² concebido para la digitalización a demanda; consta de una pantalla táctil e interactiva con la que se puede editar las imágenes, guardarlas en distintos formatos y transferirlas a USB.

Características principales:

- Formato de escaneo máximo: 480 x 360 mm
- Modo de escaneo: color, escala de grises, b/n; 4 canales, RGB- Color y Gris en total 43200 Pixel
- Resolución: hasta 600 dpi
- Espesor máximo del libro: 100 mm
- Dos salidas USB 3
- Valores de conexión 100 - 240 V / 50 - 60 Hz
- Dimensiones con panel táctil: 870 x 633 x 629 mm
- Superficie de instalación con panel táctil: 818 x 530 mm
- Peso aprox.: 25 kg

Escáner A0

Escáner adquirido por la Oficina del Historiador a través de un proyecto de cooperación internacional con la Agencia Española de Cooperación Internacional para el Desarrollo (AECID) en el año 2009. Se trata de un escáner planetario cuyo funcionamiento consiste en el desplazamiento de una cámara lineal sobre el documento (sin ningún contacto) y digitaliza una o dos páginas al mismo tiempo. Los datos se transfieren poco a poco a un PC que tratará la imagen y la visualizará en una pantalla de alta definición.

Características principales:

²<https://www.zeutschel.de/en/produkte/scanner/farbscanner/zeta-comfort.html>

Obras aceptadas:

- formatos: DIN A5 a DIN AO formato vertical (2x DIN A1)
- altura y anchura máximas: 1250 mm (I) x 870 mm (h)
- grosor máximo: 120 a 500 mm. con opción de soporte libro, 40 mm. sin opción

Digitalización:

- cámaras intercambiables: 6000, 10200, 14400 píxeles en gama de grises y 3x6000, 3 x 10200, 3 x 14400 en color.
- 24 bits en color, 8 bits en escala de gris, 1 bit en binario
- resolución óptica máxima (la resolución real depende de la distancia entre la cámara y el documento y, por consiguiente, del grosor del mismo): con 6000 píxeles: A1 250 dpi, A3 400 dpi, A4 600 dpi con 10200 píxeles : A0 300 dpi, A1 400 dpi, A2 600 dpi con 14400 píxeles : A0 400 dpi, A1 600 dpi, A2 800 dpi
- corrección instantánea y memorización de la iluminación
- enderezamiento del texto
- corrección de la curvatura de la página
- balance de blancos

Formato de los ficheros:

- TIFF 3 x 8 bits no comprimido, JPEG o PNG en color
- TIFF 8 bits no comprimido, JPEG o PNG en escala de gris
- TIFF G4 en binario

Por logística u otras razones de organización o técnicas, se podrán utilizar otros equipos, escáneres o cámaras digitales, pertenecientes a la OHCH, siempre y cuando cumplan con los estándares de digitalización definidos en esta Política.

Otros

El equipamiento del proyecto también incluye un kit de calibración para ambos escáneres, una estación de trabajo (monitor y torre), un servidor Huawei (6 x HDD, 8TB, SATA), tres access points Huawei, 6 monitores EIZO, 4 memorias externas WD Elements Desktop de 8 TB, 5 memorias externas Intenso de 128GB, y 6 discos duros internos Fujitsu 1 TB SATA

Software

ProHD ha adquirido tres programas informáticos para gestionar el escaneo y procesar las imágenes resultantes de la digitalización: Omniscan 12 y OmniPro para el escáner planetario fijo y un módulo de OCR específico para el escáner planetario móvil.

Las licencias de los programas adquiridos por ProHD permiten obtener actualizaciones durante un año. Los costos solo incluyen la licencia para la actualización; la actualización misma debe realizarse en el sitio.

Omniscan 12

Omniscan 12³ es un programa multilingüe fácil de utilizar que permite capturar las imágenes de manera intuitiva gracias a la interfaz de usuario, guardar distintas configuraciones con el objetivo de estandarizar la realización de trabajos repetitivos y editar grandes cantidades de imágenes sin que la memoria se resienta. Con este programa se puede digitalizar un documento una sola vez y convertir las digitalizaciones a distintos formatos de fichero: TIFF, JPEG, PNG, PDF, multipágina PDF y multipágina TIFF, entre otros. También facilita la descripción de los documentos con metadatos en forma de etiquetas TIFF, la definición de los nombres de los ficheros e implementar algunos controles de calidad mediante pasos y tareas.

Para utilizar correctamente Omniscan 12 es necesario distinguir tres conceptos fundamentales:

- *Job* (Trabajo): las imágenes resultantes del proceso de escaneado y un fichero “job” con los parámetros seleccionados.
- *Scan* (Escaneado): los escaneados resultantes del proceso de digitalización.
- *Clip* (Imagen): las imágenes digitales derivadas de los escaneados.

A diferencia de otros programas de escaneo, en Omniscan 12 una imagen digital no es simplemente el resultado de un escaneado. Al contrario, un escaneado puede generar muchas imágenes digitales con distinta resolución, formato, recorte, etc. Por ejemplo, el escaneado de un libro ilustrado puede dar lugar a un clip con la página entera y varios clips con los que se seleccionan únicamente las fotografías (sin el texto) contenido en la página.

Los parámetros y especificaciones de un “job” pueden guardarse de tal manera que se puedan reutilizar en el futuro con otros documentos. En consecuencia, es útil crear distintos “jobs” para cada tipo de documento, si los parámetros cambian. Por ejemplo, se puede crear, en primer lugar, un “job” para escanear volúmenes impresos encuadernados y generar ficheros TIFF como ficheros maestros, y JPEG y PDF multipágina con OCR como ficheros de acceso. En segundo lugar, se

³<https://www.zeutschel.de/en/software/capturing-software/os-12.html>

puede crear otro “job” para escanear volúmenes manuscritos encuadernados para y generar TIFF como ficheros maestros y únicamente JPEG como ficheros de acceso. El software Omniscan 12 adquirido por ProHD incluye dos módulos adicionales: Perfect Book 3.0 y Imaging Kit.

Perfect Book 3.0⁴ sirve para corregir de manera automática la curvatura, separar las páginas, recortarlas, y eliminar la presencia de dedos. Este módulo es especialmente importante cuando se escanean documentos encuadernados.

Imaging Kit sirve para editar imágenes y llevar a cabo las siguientes acciones:

- Corrección automática del color
- Máscara “unsharp”
- Rotar
- Invertir
- Enderezar
- Recortar
- Insertar marca de agua
- Binarizar
- Reconocer el área de impresión
- Obtener texto con OCR

Este módulo es de uso general con independencia del formato de los documentos.

Omniscan 12 permite procesar las imágenes de cuatro formas: de manera individual y manual después del escaneado; de manera global (batch) y manual después del escaneado; durante el escaneado, de manera global (batch) y automática; o bien al terminar el “job” y de manera automática y global.

OmniPro

OmniPro⁵ permite procesar imágenes procedentes de otros escáneres como zeta comfort. Con OmniPro es posible controlar la calidad de las imágenes (como hace Imaging Kit), transformar un documento a varios formatos de ficheros y generar texto con OCR en formato METS/ALTO.

Módulo de OCR

El módulo de OCR para el escáner planetario portátil permite generar documentos PDF con OCR de manera directa, es decir, sin tener que transferir las imágenes a un USB y procesarlas con OmniPro.

⁴<https://www.zeutschel.de/en/software/integrierte-loesungen/perfect-book-3.0.html>

⁵<https://www.zeutschel.de/en/software/workflowloesungen/omni-pro.html>

Calidad y uso

El concepto de “calidad” es muy difícil de definir en proyectos de digitalización. Aunque se suele equiparar la resolución con la calidad (o, mejor dicho, con fidelidad) porque indica el nivel de detalle, en realidad, depende de varios factores de naturaleza técnica, ambiental y humana. Las *Directrices para proyectos de digitalización* editadas por IFLA dan buena cuenta de la complejidad de este concepto:

La calidad de la imagen durante la captura depende de la suma de resultados de la resolución aplicada al escaneo, la profundidad del bit de la imagen escaneada, los procesos de mejora y el nivel de compresión aplicada, el dispositivo de escaneo utilizado o técnicas usadas, y la preparación del operador del escáner (IFLA, 2002: 45).

Asimismo, tal y como reconocen las *Directrices técnicas* de FADGI, las condiciones físicas en que se lleva a cabo la digitalización de documentos también influyen en la percepción y evaluación de la calidad:

La estandarización del entorno de digitalización crea un espacio de trabajo donde las variables de la percepción visual pueden ser controladas. Sin estandarización, la percepción de la calidad de imagen puede variar dramáticamente. En un entorno de creación de imágenes donde la opinión humana sea un factor, estandarizar el entorno físico es de importancia crítica para mantener la consistencia (Rieger, 2016: 14).

Dado que la calidad es tan difícil de definir, conviene mejor reconocer su naturaleza relativa (se considera que algo tiene calidad óptima o suficiente, excesiva o defectuosa) y adoptar un criterio más pragmático: el uso. En otras palabras, una imagen digital tiene suficiente calidad si posibilita el uso previsto.

¿Cuál es el uso previsto de las digitalizaciones en el contexto de ProHD? Las imágenes digitales ráster obtenidas deberían satisfacer dos usos principales: preservación y acceso. Es decir, las digitalizaciones de calidad elevada (en formato de preservación) deben permitir generar otros ficheros en formato de acceso para su publicación en el repositorio digital del proyecto (más detalles sobre este tema en la sección “Formatos de fichero”). A su vez, estos ficheros en formato de acceso deberían permitir la obtención de texto con OCR en el caso de los impresos o la transcripción (manual o asistida) en el caso de los manuscritos.

En el sistema de evaluación establecido por FADGI es clave la relación entre calidad, el rendimiento técnico del operador y de los escáneres, y el uso previsto. Las *Directrices Técnicas* de FADGI definen cuatro niveles de calidad de imagen, que puntúan con estrellas (1 estrella = menor calidad, 4 estrellas = mayor calidad):

1. **Una estrella:** la creación de imágenes de una estrella se debe considerar informativa solamente, pues las imágenes no tienen calidad suficiente para

ser útiles en el reconocimiento óptico de caracteres u otras técnicas de procesamiento de información.

2. **Dos estrellas:** la creación de imágenes de dos estrellas es apropiada si no hay una esperanza razonable de obtener un rendimiento de tres o cuatro estrellas. Las imágenes tendrán valor informativo solamente, y pueden ser, o no ser, adecuadas para el reconocimiento óptico de caracteres.
3. **Tres estrellas:** la creación de imágenes de tres estrellas define una muy buena imagen profesional que admite casi todos los usos.
4. **Cuatro estrellas:** la creación de imágenes de cuatro estrellas define la mejor práctica en la actualidad. Las imágenes creadas con un nivel de cuatro estrellas representan el estado del arte en la captura de imágenes y son adecuadas para casi cualquier uso.

Con el escáner planetario fijo (OS15000 Advanced Plus) es posible lograr el nivel 3 de FADGI, mientras que con el escáner planetario móvil (zeta comfort) la calidad obtenida se sitúa entre los niveles 1 y 2⁶. El nivel 2, por tanto, sería el mínimo deseable para las digitalizaciones de ProHD. Con el objetivo de que las imágenes digitales tengan “valor informativo” y sean “adecuadas para el reconocimiento óptico de caracteres”, conviene incrementar el rendimiento en los otros factores que influyen en la calidad: el componente humano (personal con formación y experiencia) y el ambiental (condiciones de la habitación estandarizadas).

Especificaciones

Para obtener un nivel 3 de calidad según el sistema de evaluación FADGI de los siguientes tipos de materiales. Se aplicarán las siguientes especificaciones:

Tipo de documento	Fichero maestro	Copia de difusión
Libros raros, especiales o valiosos	600 ppi, 24 bits, TIFF, Color	300 ppi, JPG y PDF+OCR, Color
Manuscritos	600 ppi, 24 bits, TIFF, Color	300 ppi, JPG, Color
Impresos	400 ppi, 24 bits, TIFF, Color	300 ppi, JPG y PDF+OCR, Color
Prensa	300 ppi, 24 bits, TIFF, Color	300 ppi, JPG y PDF, Color

Formatos de ficheros

En el contexto de ProHD, se espera que el fichero máster archivístico en formato TIFF sea preservado a largo plazo como parte del plan de preservación de la Dirección de Patrimonio Documental. En cuanto a los formatos de acceso, como

⁶Sobre la dificultad de adherirse a las directrices FADGI y Metamorfose se recomienda la lectura de esta entrevista a John Barrett (Bodleian Library), en inglés: <https://www.genvusit.com/digitisation-standards-at-the-bodleian-library-by-john-barrett/>

contenido del repositorio digital del proyecto, se prefiere el uso de JPEG para todas las digitalizaciones tanto de manuscritos como de impresos, con una ratio de compresión baja para mantener una calidad suficiente que permita ver con nitidez los detalles de las grafías; la copia de difusión debe tener una resolución mínima de 300 ppi. Además, para los documentos impresos, se incluirá una copia en formato PDF multipágina con OCR.

Nombres de ficheros

Antes de la captura con el escáner, es necesario establecer un nombre con el que identificar el fichero resultante y guardarlo en un directorio específico del ordenador. Para evitar problemas al localizar ficheros, es una práctica estándar establecer un esquema que tenga sentido en el contexto del proyecto y que permita identificar de manera unívoca, persistente y consistente los ficheros de imágenes. Dicho esquema o convención también debería establecer relaciones con otros ficheros como, por ejemplo, la secuencia de las imágenes.

Zhang y Gourley (2008: 44) distinguen dos tipos de convenciones para nombrar ficheros “no-descriptiva” y “significativa”. La primera convención asigna nombres arbitrarios a los ficheros en forma de identificadores alfanuméricos como, por ejemplo, “7646”. En cambio, una convención significativa codifica información descriptiva en el nombre del archivo como el formato, el número de la página, el propietario, etc. Las *Directrices Técnicas* de FADGI (Rieger, 2016: 104) recomiendan el uso de nombres de ficheros significativos para facilitar la navegación y localización en proyectos que no utilizan una base de datos para organizar las digitalizaciones.

En el contexto de ProHD, es preferible utilizar nombres significativos y establecer antes de escanear un documento el nombre del fichero siguiendo las siguientes recomendaciones:

- Los nombres de ficheros significativos deben conectar la digitalización con el documento físico original y su procedencia.
- Siempre que sea posible, se usarán las siguientes siglas para identificar las bibliotecas y archivos cooperantes: aohch (Archivo de la Oficina del Historiador de la Ciudad de La Habana), bnc (Biblioteca Nacional de Cuba), uh (Universidad de La Habana), acc (Academia de las Ciencias de Cuba), anc (Archivo Nacional de Cuba), ill (Instituto de Literatura y Lingüística).
- Los nombres de ficheros deben ser breves para evitar introducir errores humanos. Además, deben ser homogéneos y consistentes. Para lograr esto, es recomendable utilizar solo minúscula, evitar el uso de acentos y utilizar un solo tipo de guion para separar partes (en lugar de espacios en blanco).
- Las secuencias de números deben iniciarse con un 0 para facilitar el ordenamiento numérico: c01c03, c10c22, etc. Para establecer el número

necesario de 0 iniciales, es necesario saber de antemano la escala del objeto digitalizado, es decir, si se trata de 2 cajas, 10 tomos o de un documento individual.

- No hace falta incluir información redundante como el formato de los ficheros. Los nombres de los ficheros de imágenes digitales ya contienen una extensión que identifica el formato como *.tif* para TIFF, *.jpg* para JPEG o *.pdf* para PDF.
- Es necesario incluir información sobre las distintas representaciones de un mismo objeto para distinguir entre el archivo de máster, el archivo de máster de producción, el archivo de acceso y el archivo de acceso en miniatura. Así, por ejemplo, se recomienda incluir un calificador al final del nombre como “aohch-ms-c01c03-06-master.tif”, “aohch-ms-c01c03-06-master-produccion.tif”, “aohch-ms-c01c03-06-acceso.jpg”, “aohch-ms-c01c03-06-acceso-miniatura.jpg”.

Flujo de trabajo

Para la elaboración del flujo de trabajo se han consultado las Directrices técnicas de FADGI (Rieger, 2016) y la tesina de Diplomado en Conservación de Grisel Terrón (2014).

Preparación

- Selección previa de los documentos por instituciones cooperantes.
- Acuerdo de sistema de trabajo en cada caso: en la Casa Humboldt con el escáner OS 15000 o bien en la institución cooperante con el escáner móvil zeta comfort.
- En todo caso se rellenarán los documentos de entrega establecidos en el proyecto y los que exija la institución cooperante; en estos documentos se establecerán la duración aproximada de la digitalización.
- Si la digitalización se lleva a cabo en la institución cooperante, se crearán las mejores condiciones para el trabajo, en dependencia de lo que disponga la institución (local, luz, electricidad, seguridad, mobiliario mínimo, etc.).
- Si la digitalización se lleva a cabo en la Casa Humboldt, se revisarán las condiciones del entorno de trabajo (iluminación e incidencias posibles en la captura).
- Valoración del estado de conservación de los documentos atendiendo a:
 - Papeles quebradizos
 - Libros quebradizos por el lomo

- Ejemplares con ataques graves de microorganismos o de insectos, que impidan la lectura
- Ejemplares con problemas físicos, como hojas desgarradas, sueltas o con pérdidas
- Ejemplares con mapas o grabados desplegados para seguir una política especial que tenga en cuenta su tipología documental y su tamaño
- Ejemplares gravemente deformados por la acción del agua o de la mala colocación que pueden presentar deformaciones importantes en la caja de texto
- Una vez creadas las condiciones de trabajo y revisados los ejemplares, se procederá a preparar el documento físicamente:
 - Valorar el grado de intervención puntual para prepararlo
 - Retirar presillas metálicas
 - Desdoblar los documentos y puntas plegadas
 - Limpiar el documento con brocha (limpieza mecánica) para eliminar el polvo
- Solo en casos excepcionales, donde la integridad de la información del documento esté comprometida, y previa consulta y aprobación documentada de la institución propietaria, el documento se someterá a un proceso de restauración, tras evaluar las complejidades de la intervención y las capacidades reales para acometerla. La restauración sólo pretenderá garantizar la legibilidad de la información.
- Limpieza del escáner de acuerdo a sus especificidades técnicas para evitar que partículas ajenas sean captadas.

Digitalización o captura

- Se deberá cuidar especialmente la manipulación del documento y el escaneado con el equipo correcto diseñado para la protección del original.
- El proceso de digitalización debe documentarse.
- La captura debe abarcar un área mayor que el documento aumentando la superficie captada en 1,5 cm por cada lado.
- La imagen digital resultante debe captar el margen interno y no debe despreñar ningún detalle del documento.
- De todos los documentos digitalizados (volúmenes encuadernados, no encuadernados y periódicos), se obtendrán ficheros de preservación (máster) en formato TIFF, en color, con una resolución mínima de 300 ppi y una profundidad mínima de 24 bits.

- Estos parámetros, sin embargo, pueden aumentarse (por ejemplo, la resolución puede incrementarse hasta 600 ppi para documentos con caracteres muy pequeños) tras una primera evaluación de la calidad y si el resultado obtenido no permite la lectura del texto.
- No se debe alterar el contenido semántico del documento digitalizado y se evitará optimizar con filtros o retoques el fichero maestro a excepción de la curvatura del lomo y la eliminación de dedos.
- Cuando el resultado obtenido sea satisfactorio, se exportarán o guardarán las imágenes digitales en formato de acceso: JPEG para todos los documentos y, además, para los documentos impresos, PDF con OCR.
- Se guardarán los ficheros maestros en una carpeta individual siguiendo el esquema de nombres de ficheros expuesto más arriba; en esta carpeta también se guardarán las copias de difusión y el resto de ficheros mencionados en este flujo de trabajo.

Post-procesamiento

- Se garantizará que tanto la imagen digital como los metadatos significativos obtenidos en el proceso de digitalización son fieles al documento original y cumplen con los requisitos de calidad establecidos cumplimentando una matriz con los siguientes aspectos y una puntuación (ilegible, mínimo, bueno, superior):
 - Integridad del documento
 - Legibilidad del contenido del documento
 - Nitidez
 - Uniformidad
 - Apariencia del documento digital respecto al original
 - Fidelidad de los colores
 - Calidad de color-brillo-contraste
 - El documento debe estar derecho
 - Nombre de los ficheros
- Una vez se apruebe la calidad de la digitalización, se procederá con la optimización de la imagen para la copia de acceso de acuerdo a:
 - Como en el momento de la captura, los bordes no se recortarán exactos al tamaño del documento sino que se escaneará dejando un borde por fuera de la página real, durante el control de la calidad se recortará lo que queda bordeando el documento de manera que las páginas queden exactamente como son en el documento físico.

- Si el documento tuviera alguna mancha que hiciera imposible su lectura, se limpiará hasta hacer legible el documento, no es necesario que desaparezca totalmente
- Las imágenes que vayan a ser procesadas con OCR podrán ser trabajadas para mejorar la efectividad del proceso
- Cuando se utilicen los escáneres de Zeutschel, el software de optimización será Omniscan 12; como alternativa, se puede usar Photoshop si se usan otros escáneres de la OHCH.
- Cuando se realicen cambios en la imagen digitalizada, siempre se conservará el “master” o “fichero maestro” sin ningún proceso de optimización. La copia “manipulada” u “optimizada” deberá revelar en todo momento que, efectivamente, ha sido sometida a tal proceso, y así se hará constar en los metadatos.
- La mejora de las imágenes digitales debe limitarse al balance de tonos y la eliminación de accidentes sobre el documento (rayas, puntos de oxidación, manchas de humedades y roturas leves) que dificulten la lectura o apreciación.
- La optimización de la imagen será minuciosa cuidando los detalles de la obra como sellos, miniaturas, adornos, o cualquier otro detalle artístico.
- Se trabajará todo el tiempo con el original a la vista, para controlar posibles borrados de contenido y alteraciones de color.
- No se harán restauraciones digitales sino solo pequeños arreglos que contribuyan a la legibilidad del documento.
- A continuación, se procederá con el reconocimiento óptico de caracteres (OCR) de los documentos impresos. Se utilizará el módulo de OCR de Omniscan 12 para obtener el texto.
- Se generará un documento PDF que contenga la imagen del documento y el OCR oculto para hacer búsquedas en él.

Creación de metadatos

- En todo momento, el operador tendrá a la vista el documento original para, en caso de duda, consultarlo y así describir con mayor certeza las imágenes digitales desde un punto de vista bibliográfico (metadatos descriptivos) y la estructura física del documento, esto es, la secuencia de páginas individuales y su relación con los ficheros. Asimismo, se consultará la ficha bibliográfica, si existiera (por ejemplo, un catálogo en línea) o bien si es proporcionada por la entidad cooperante.

- La estructura físico-lógica del documento distinguirá las secciones principales del documento: encuadernación, portada (página del título y autor), sección de ilustraciones (cuando aparecen todas juntas), prólogo, capítulos o división intelectual del contenido, apéndice, bibliografía y fe de erratas. Si el libro no tuviera alguna de estas partes, se crearán las divisiones de acuerdo al contenido propiamente o bien se representará únicamente una secuencia de páginas.

Copias de respaldo y almacenamiento (*backup*)

- Se guardarán los ficheros maestros en un servidor y una copia de los ficheros maestros en las memorias externas (back-up) del proyecto a fin de restaurar los ficheros en caso de pérdida o destrucción de los originales.
- Tras la devolución de los documentos a las bibliotecas y archivos cooperantes, se iniciará el proceso de ingesta, administración, almacenamiento y publicación de los ficheros de acceso en el repositorio digital del proyecto; para ello, se usarán los ficheros de acceso y el fichero de metadatos.
- Si algún día el formato TIFF es sustituido por otro formato estándar, los responsables de la preservación de los ficheros migrarán la información al nuevo formato estándar.

Referencias bibliográficas

- Díez Carrera, Carmen. 2012. *La biblioteca digital*. Madrid: Ediciones Trea.
- IFLA. 2002. *Directrices para proyectos de digitalización*. Ministerio de Cultura. <https://www.ifla.org/files/assets/preservation-and-conservation/publications/digitization-projects-guidelines-es.pdf>
- NISO. 2007. *A Framework of Guidance for Building Good Digital Collections*. <http://framework.niso.org/5.html>
- Rieger, Thomas. 2016. *Directrices técnicas para digitalizar materiales del patrimonio cultural. Creación de imágenes ráster*. Federal Agencies Digitization Guidelines (FADGI). http://www.digitizationguidelines.gov/guidelines/FADGI%20Federal%20%20Ag2016%20Final_rev1.pdf
- Terrón, Grisél. 2014. *La digitalización para la conservación del patrimonio documental en la Dirección de Patrimonio Documental de la OHC*. Tesina de diplomado. Colegio Universitario de San Gerónimo de La Habana.
- Zhang, Allison B. y Don Gourley. 2008. *Creating Digital Collections. A Practical Guide*. Oxford: Chandos Publishing.