15ª Jornada sobre la Biblioteca Digital Universitaria
"Acceso a la información: uso e impacto"
2 y 3 de noviembre de 2017

# Una experiencia de digitalización masiva en la Biblioteca Digital de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires

Autores:

Lic. Martín Williman, mwilliman@bl.fcen.uba.ar

Prof. Ana Sanllorenti, asanllorenti@bl.fcen.uba.ar

Biblioteca Central "Luis F. Leloir" de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires

**Resumen:**

En la construcción de bibliotecas digitales y repositorios, los procesos de digitalización son una de las fuentes principales de incorporación de documentos. La Biblioteca Central de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires, guarda la colección completa de tesis doctorales desde 1889, superando las 6000 tesis. A partir de 2010 se estableció la obligatoriedad de entregarlas en formato papel y digital.

Con la finalidad de poner en línea la colección completa de tesis doctorales entre otros objetivos, se presentó un proyecto de financiamiento ante el Sistema Nacional de Repositorios Digitales, Proyecto RDF3 "Mejoramiento cualitativo y cuantitativo de la Biblioteca Digital de la FCEN-UBA". Durante 2016 y 2017 se digitalizaron 3200 tesis doctorales de la Facultad. Se tomaron becarios que realizaron la captura de las imágenes, se adquirió equipamiento y se desarrolló un software de seguimiento de tareas.

El trabajo presenta el circuito completo de planificación del proyecto, su ejecución, la descripción de los problemas presentados y las decisiones tomadas en el marco de un proyecto de digitalización masiva de una colección.

## Introducción

La Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires (FCEN-UBA) es una de las instituciones argentinas con mayor calidad y cantidad de producción científica, en

gran parte publicada en revistas científicas a nivel internacional. Además de los más de 800 artículos anuales publicados en revistas con referato, los graduados elaboran y presentan tesis doctorales, los investigadores y docentes elaboran ponencias y escriben trabajos que se comunican a través de otros canales formales e informales.

En cuanto a las tesis que consagran la obtención del grado de doctor, la Biblioteca Central “Luis F. Leloir” es depositaria de la colección completa, desde el trabajo de Pedro Pando, “Esquisomicetas”, la primera tesis defendida en 1882, hasta las más de 200 tesis que actualmente se aprueban por año. La colección en papel suma más de 6200 tesis doctorales y constituye un tesoro histórico y actual de la producción de conocimiento científico en Argentina.

Desde 2005 y mediante una modificación del Reglamento de Doctorado (Resolución CD 2053/05) los tesistas envían a la Biblioteca Central una copia digital además de la de papel. En 2010 se inauguró la Biblioteca Digital de la FCEN-UBA con las características de un repositorio institucional, como una herramienta para la gestión, almacenamiento y preservación de lo que la comunidad académica produce, que sea soporte de la investigación, la docencia y el aprendizaje. Una de las colecciones principales del repositorio institucional es la colección de tesis que en la actualidad cuenta con 3020 tesis a texto completo en línea. En 2013 el Consejo Directivo autorizó a la Biblioteca Central a digitalizar y disponer en Acceso Abierto la colección de tesis doctorales hasta 15 años atrás del año en curso (Resolución CD 272/2013).

La colección de tesis digitales tiene un promedio de 16.000 consultas mensuales.

La Biblioteca Digital FCEN-UBA es compatible con las Directrices del Sistema Nacional de Repositorios Digitales. Por Resolución No. 007/13 del 29/01/2013 fue aceptada su adhesión al Sistema y hoy es uno de los 22 repositorios cosechados en el portal del SNRD, aportando el 3,70% de lo reunido en el Sistema.

Cuenta con recursos humanos, tecnológicos, jurídicos y administrativos que le dan un carácter sustentable. Tiene el respaldo de las autoridades de la Facultad y de su Consejo Directivo que mediante la Resolución 2533/09 aprobó el instrumento jurídico para la gestión de los derechos de autor y asignó a la Biblioteca Central la responsabilidad por la gestión del repositorio..

A principios de 2015 la Biblioteca Digital FCEN-UBA tenía 1200 tesis doctorales y era insuficiente la colecta y puesta en Acceso Abierto de artículos científicos producidos por investigadores de la Facultad. En ese momento se analizó y concluyó que era necesario una adecuación política, técnica y de producción a los mandatos de la Ley de creación Repositorios

Digitales, por lo que se identificaron ejes para fortalecer el repositorio y cumplir a cabalidad la Ley 26.899.

Con esa estrategia se elaboró y presentó un proyecto para obtener financiamiento del Sistema Nacional de Repositorios Digitales: “Mejoramiento cualitativo y cuantitativo de la Biblioteca Digital de la FCEN-UBA”, que fue aprobado por Resolución SACT Nº 008/15.

El proyecto tuvo entre sus objetivos:

1) Políticas y sensibilización: Elaboración de una política de Acceso Abierto para la FCEN-UBA
2) Colecciones: Incorporación de 1200 tesis doctorales a través de la digitalización; Incorporación de 500 artículos con referato; Normalización y mejora de la base referencial de 11.000 artículos de revistas con referato; Revisión de las políticas editoriales de las revistas en las que publican los investigadores de la Facultad respecto del Acceso Abierto, con fines de incorporar artículos en forma retrospectiva.
3) Infraestructura informática: a) incorporación de dos nuevos servidores, que incrementen la capacidad de atención simultánea de usuarios, otorguen mayor robustez en el funcionamiento y aumenten la capacidad de almacenamiento con miras a cumplir con los objetivos de crecimiento de los próximos cuatro años y de preservación a largo plazo; b) incorporación de un scanner para aumentar la capacidad de digitalización de documentos.

A fines de 2015 el SNRD otorgó la financiación para la contratación de becarios que realizaran las tareas de digitalización y marcado de las tesis y para la adquisición del equipamiento informático. Las demás actividades fueron sustentadas por la FCEN-UBA en carácter de contraparte.

El presente trabajo da cuenta de la realización del objetivo de digitalización de las tesis doctorales.

**Planificación**

Para la caracterización del material a digitalizar se realizó un análisis del estado físico de los ejemplares con la finalidad de determinar los requerimientos técnicos para la captura digital.

Para eso se tuvo en cuenta el tipo de encuadernación, el número promedio de páginas, el tipo de impresión, el tamaño del papel y el estado de conservación. Por otra parte se consideró la

existencia de ejemplares duplicados –un 76 % de las tesis lo tenían- y anexos con planos, mapas y otras láminas que superaran el tamaño de la página de texto. También se encontraron tablas, fotografías, fórmulas, firmas originales, anotaciones personales y correcciones manuscritas.

Como resultado, la colección fue dividida en 5 secciones, correspondientes a los distintos períodos en que fueron producidas. Por ejemplo, la sección 1 abarcó las tesis desde 1882 hasta 1912, que poseen una encuadernación cosida de tapas duras, con papel de imprenta, con 120 páginas de promedio, de tamaño menor que A4, buen estado de conservación, tipografía de imprenta y un 46% de tesis con ejemplar duplicado.

El estudio anterior permitió estimar los volúmenes a digitalizar y los parámetros para la captura digital. Se relevaron 3355 tesis en condiciones de ser digitalizadas, con un total de 503.393 páginas. De estas tesis, 792 tenían ejemplar único y 2563 contaban con un duplicado.

**Parámetros de digitalización y equipos de captura:** En base a lo evaluado, se decidió digitalizar con una resolución de 300 ppp con un espacio de color RGB en scanner plano, tamaño oficio. Para los ejemplares con duplicados se optó por su desarmado, guillotinado y digitalización con alimentador automático. Para los ejemplares únicos se utilizó la cama plana del scanner con soporte para mantener la integridad de la encuadernación.

Si los anexos mayores al tamaño de la página de texto se extendían en forma de secuencia horizontal o vertical, se los digitalizaba en partes. Los anexos que se desplegaban en ambas dimensiones, quedaron pendientes de digitalización posterior, con un scanner adecuado.

**Procesos pre escaneo:** Los procesos previos a la digitalización consistieron en la limpieza del material, desanillado si correspondía, eliminación de ganchos y tachuelas. En el caso de las tesis con duplicado, guillotinado de las mismas. Cada conjunto de páginas correspondientes a una tesis se colocó en un sobre de papel madera con su identificación.

**Estimaciones:** En la etapa de planificación se estimaron 1508 horas para la digitalización, a ser cumplidas en 8 hs por día en 9 meses y medio. El cálculo se basó en una proyección de 40 minutos promedio para las tesis con ejemplar único y 25 minutos para duplicados digitalizados con alimentador automático.

De acuerdo con el volumen a digitalizar y los parámetros de captura se estimaron 11,3 TB para las imágenes máster y 17 GB para los derivados[1].

**Formatos de archivos y esquema de nombramiento:** Se decidió utilizar formato TIFF para las imágenes máster de cada página con compresión sin pérdida GIII y GIV.

El formato elegido para los archivos derivados fue PDF con el OCR[2] correspondiente y optimizados para su descarga en la web.

Para el nombramiento de los archivos se aplicó una nomenclatura descriptiva, utilizando un número consecutivo para la tesis y el apellido del autor sin caracteres especiales.

**Almacenamiento:** Se optó por un sistema de almacenamiento en línea para los másteres y los derivados, en agrupamiento de discos duros con un sistema RAID 5. La adquisición de dos servidores iguales permitió la guarda de las imágenes y los datos en dos lugares físicos diferentes.

**Metadatos:** Se utilizaron los datos de la base bibliográfica de la colección de tesis, que la Biblioteca ya poseía y actualiza regularmente en formato BIBUN (para bibliotecas de la Universidad de Buenos Aires).

Estos datos son migrados a la Biblioteca Digital mediante un mapeo a un perfil de Dublin Core con calificadores propios.

[1] Se denomina master a la imagen digital que contiene toda la información obtenida durante el proceso de captura. El documento derivado se produce a partir de la imagen master, con características específicas para su utilización en la Web y que sólo posee la información pertinente a su función.

[2] OCR: Optical Character Recognition, Reconocimiento Optico de caracteres, proceso por el cual la imagen de un símbolo es identificada como un caracter

**Ejecución**

En el esquema siguiente se resumen las principales etapas y procesos del circuito de digitalización hasta la obtención de la versión publicable en la Biblioteca Digital.


**Limpieza:** Se trató cada tesis con aspiradora y pinceles para reducir el polvo de los ejemplares más antiguos. Durante este proceso se detectaron cuestiones de encuadernación o integridad de las obras que afectaban la captura. Por ejemplo, hojas sueltas, tesis sin carátulas, partes a reparar, entre otros.

**Desencuadernado:** Para las obras con ejemplares duplicados se destinó el duplicado para el proceso de digitalización con alimentador automático. Esto implicó el desencuadernado y guillotinado de las copias que, en este caso, alcanzó al 76% de las tesis a digitalizar. El proceso de desencuadernado consistió en retirar las tapas, espirales, ganchos, carpetas y luego guillotinar el margen izquierdo para eliminar las irregularidades que podrían atascar el alimentador automático.

**Ensobrado y Rotulado:** A fin de mantener la integridad de cada obra, las hojas resultantes del desencuadernado se colocaron en sobres de papel madera con una identificación unívoca mediante un rótulo adherido al exterior del sobre que tiene el número de de la obra.

**Software de Seguimiento:** En esta etapa comienza a utilizarse el Software de gestión de la digitalización. En la interfaz en línea, cada operador se registra y comienza su sesión de trabajo. Utiliza la información del rótulo para recuperar los datos de la obra que están precargados y que permiten confirmar la correcta identificación de la obra y dar comienzo al proceso de digitalización.

El operador indica al sistema la cantidad de imágenes estimadas para la captura y elige un perfil de filtros de imagen[3] para la obra en proceso. El operador fue entrenado previamente para la identificación de estos perfiles, que responden a características de color del fondo, contraste entre fondo y texto, presencia de manchas o irregularidades en el color del papel, grosor de la tipografía, entre otros. De este modo se determina la selección entre 8 posibles perfiles, combinables entre sí. Dado un conjunto de perfiles de filtros, las imágenes capturadas recibirán procesos específicos para mejorar la calidad de la imagen en los procesos de OCR.

**Captura:** Por alimentador automático para duplicados desencuadernados y de forma manual para ejemplares únicos.

**Controles Automáticos:** El software de gestión controla: los Parámetros de captura: Resolución y color. Nombramiento de los archivos. Coincidencia entre la cantidad de hojas estimada por el operador y las capturadas.

**Marcado:** Identificación de las hojas con imágenes para su correcta visualización en el documento derivado, excluyéndolas de las modificaciones que generan los procesos automáticos post-captura. Marcado del documento en sus secciones que permitirán construir las tablas de contenido navegables.

**Controles Manuales:** Evaluación de la calidad de las imágenes. Detección de hojas faltantes o repetidas. Verificación de los perfiles de filtros de imagen. Aceptación o rechazo para su corrección.

**Pre-OCR:** Aplicación de filtros de imagen para mejorar la separación entre fondo (papel) y figura (caracter). Eliminación de ruido y manchas, mejorando la lectura de la fuente.

[3] Cada perfil de filtro de imagen asignado indica los filtros que se corren sobre una imagen digitalizada para mejorar los resultados del OCR.

**OCR:** Se utiliza ScanTailor para la estructuración del documento y Tesseract como motor de OCR.

**Post-OCR:** Identificación de palabras equivocadas y corrección por el mejor término candidato posible utilizando diccionarios español e inglés, documentos y datos bibliográficos del dominio disciplinar, diccionario de abreviaturas y listas de términos científicos.

**Publicación:** Publicación de los trabajos de post-grado en la Biblioteca Digital de la FCEN-UBA. Disponibilidad de los documentos por protocolo OAI-PMH para su reutilización por cosechadores (SNRD, BASE, SISBI, etc.)

**Desarrollo de software de gestión:** Se desarrolló un sistema para la gestión de los procesos de digitalización: captura, OCR y construcción de archivos en formato portable (PDF) para su publicación en el repositorio. Está desarrollado en lenguaje Python 3, con el framework Django. Se complementó con el uso de la librería Celery para manejo de tareas, Tesseract como motor de OCR y ScanTailor para pre-procesado de las imágenes.

El software de seguimiento permite:

- Asignar operadores, revisores, administradores y puestos de escaneo
- Automatizar y controlar el nombramiento de archivos.
- Controlar la correcta asignación de los parámetros de captura y valida el formato del archivo
- Asignar perfiles de filtro de imágenes
- Marcar secciones y ubicar imágenes o gráficos en los archivos capturados.
- Controlar los trabajos realizados. Solicitar correcciones al operador correspondiente.
- Administrar los procesos automáticos de post-captura hasta la obtención del derivado para publicación: aplicación de los filtros de imágenes, ejecución de OCR, generación de documentos derivados (portada normalizada, encabezado del documento con los metadatos correspondientes y marcas que reflejan las secciones de la tabla de contenido)
- Asignar estados para los documentos en proceso (Digitalizando, Marcando, Pendiente, Aceptada; Rechazada, Terminada, etc.)
- Obtener datos estadísticos del avance de la tarea.

**Control de calidad**: En un proceso de digitalización masivo los controles de calidad de los documentos capturados no pueden realizarse de forma exhustiva ya que implicarían tiempos

similares o mayores que los de la captura. Por esa razón se determinó una forma de control dirigida a detectar patrones de hábitos en cada uno de los operadores de digitalización. Para ello se asignaron dos personas que con regularidad revisaron los materiales digitalizados en conjunto por un operador determinado. En esos controles se evaluó la calidad de la imagen, la integridad del documento, y los perfiles asignados. Los parámetros de captura y el formato de archivo fueron controlados en forma automática por el software de gestión. Luego de la evaluación se le entregó a cada operador un informe para que hiciera los ajustes necesarios en su rutina de trabajo.

**Recursos humanos:**

Personal contratado:

8 becarios, 12 hs. semanales cada uno, 9 meses. Funciones: digitalización, marcado de documentos, enriquecimiento de metadatos

Personal de la FCEN-UBA:

1 responsable del proyecto, para planificación y dirección del proyecto, dedicación full time

2 personas con 30 hs semanales de dedicación y 1 persona con 20 hs. semanales de dedicación, para preparación de materiales, supervisión y asistencia, enriquecimiento de metadatos

1 técnico informático y 1 asistente, con 20 hs. semanales de dedicación, para desarrollo y mantenimiento de software

1 especialista en imagen digital para la aplicación de OCR, que tipificó las características del material a digitalizar para aplicar los filtros más apropiados en cada caso. Esto permitió definir los perfiles de filtro de imágenes para mejorar la calidad del OCR. Contratación temporal específica.

## Resultados

* 3211 tesis digitalizadas, que completarán 5000 tesis doctorales en Acceso Abierto a fines de 2017

* Quedó constituida una infraestructura para la digitalización con dos puestos de captura y software para la gestión de esos procesos y se aumentó la capacidad de procesamiento con dos nuevos servidores de alto desempeño y la capacidad de memoria del repositorio en 16 TB.

* 6 personas capacitadas y con experiencia en digitalización y procesos de marcado y edición digital

Instrumentos para transferencia:

* Guía para la planificación de la digitalización retrospectiva de tesis en la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires.

*Sistema de gestión de los procesos digitalización: captura, OCR, construcción de archivos en formato portable (PDF) y publicación en el repositorio. Este software queda disponible para su transferencia, adaptación y reutilización por otras instituciones integrantes del SNRD. La Biblioteca Central de la FCEB-UBA se compromete a colaborar con los procesos de adaptación mediante la transferencia de información y experiencia en su utilización. La institución que lo reciba debe disponer de recursos para desarrollar la adaptación.

*Manual de Uso y Procedimientos para el operador de Digitalización de Tesis, aplicable al software de gestión descripto en el párrafo anterior.

**Reflexiones**

En un proceso de digitalización masiva es fundamental la planificación que incluya un estudio pormenorizado cuantitativo y cualitativo del material a digitalizar. En nuestro caso, el profundo conocimiento sobre las características y volumen de la colección permitió cálculos precisos sobre los tiempos de captura y el volumen del almacenamiento necesario. Asimismo el diagnóstico posibilitó una buena selección de las estrategias más adecuadas que redundaron en la abreviación de los tiempos de captura y el descarte de situaciones que estuvieran por fuera de lo previsto. Gracias a la planificación, la ejecución del proyecto se mantuvo dentro de lo pautado.

El 20 % de los tiempos de ejecución fue insumido por la preparación de los materiales, el 40% se consumió durante la captura de imágenes y el 40% restante en los procesos post-captura (enriquecimiento de los metadatos, el marcado de la estructura de los documentos e identificación de imágenes insertadas). El software de seguimiento se aplicó en los procesos de captura y post captura, e incidió fuertemente en la abreviación de los tiempos de la misma y

moderadamente en los procesos post captura ya que en estos últimos incluyen muchas tareas no automatizables.

La posibilidad de utilizar alimentadores automáticos en el 76% de la colección fue un factor determinante para reducir los tiempos de digitalización.

El trabajo de los 8 becarios con una dedicación de 12 hs. semanales distribuidas en tres jornadas de 4 hs., fue organizado en 3 puestos de trabajo. Este esquema, reforzado por el uso del software de gestión, resultó un modo eficiente de coordinar las actividades y maximizar la producción.

La aplicación de filtros de imágenes mejoró en forma notable la calidad de los procesos OCR, lo que redundó en la obtención de documentos derivados con buenos resultados en la indexación por parte de los buscadores en la Web.

La digitalización retrospectiva y puesta en Acceso Abierto del conocimiento producido por una institución universitaria adquiere valor científico e histórico. En la experiencia que se ha relatado, la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires pone a disposición de la comunidad uno de los canales principales de comunicación del conocimiento científico que ha producido en toda su historia.

**Bibliografía**

Argentina. Ley 26899: Creación de Repositorios Digitales Institucionales de Acceso Abierto, Propios o Compartidos
URL: http://repositorios.mincyt.gob.ar/recursos.php

Directrices para proyectos de digitalización de colecciones y fondos de dominio público, en particular para aquellos custodiados en bibliotecas y archivos / IFLA Traducidas por el Grupo de Trabajo de Colecciones Digitales de las Comunidades Autónomas y el Ministerio de Cultura de España.- Madrid : Ministerio de Cultura, 2005. Caps. 1 a 5, 8, Bibliografía y Apéndices. Cap. 8
URL: http://www.mcu.es/archivos/docs/pautas_digitalizacion.pdf

Federal Agencies Digitization Initiative (FADGI) - Still Image Working Group. Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files. 2010. 96 p. http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf

GARCÍA, Efraín y OSUNA, Rubén. Fundamentos de fotografía digital. http://www.uned.es/personal/rosuna/resources/photography/ImageQuality/fundamentos.

imagen.digital.pdf

Glosarios de términos sobre imágenes digitales: Glosarios, PADI: Preserving Access to Digital Information (Preservando el Acceso a la Información Digital) http://www.nla.gov.au/padi/format/gloss.html

IFLA. Directrices para planificar la digitalización de colecciones de libros impresos antiguos y manuscritos. 2014, 21 p.
URL: http://www.ifla.org/files/assets/rare-books-and-manuscripts/rbms-guidelines/directrices_para_planificar_la_digitalizacion_de_colecciones_de_libros_antiguos_impresos_y_manuscritos_-_enero_2015.pdf

JISC Digital Media. Still images, moving images and sound advice. http://www.jiscdigitalmedia.ac.uk/guides

RLG DigiNews contiene artículos que se centran en temas de vital interés para quienes gerencian iniciativas de digitalización de imágenes.
http://library.oclc.org/cdm/ref/collection/p15003coll29/id/27