15 Jornada sobre la Biblioteca Digital Universitaria, “Acceso a la información: uso e impacto”, Facultad de Derecho de la Universidad de Buenos Aires,
2 y 3 de noviembre de 2017

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

Lic. Martín Williman, Prof. Ana Sanllorenti,
Biblioteca Central “Luis F. Leloir”, Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires

La Biblioteca Digital FCEN - UBA

Producción anual de la FCEN-UBA
200 tesis doctorales
800 artículos en revistas con referato

Biblioteca Central "Luis F. Leloir" de la FCEN – UBA
Colección de 6200 tesis doctorales en papel, desde 1882
Desde 2009, Resolución CD No. 2533/09, responsable de la Biblioteca Digital

3020 Tesis
8 Publicaciones periódicas
477 fotografías
17 libros
19 Reportes técnicos

2005: Res. CD 2053/05: copia digital de las tesis
2009: Res. CD 2533/09: Licencia de depósito
2013: Res. CD 272/13: Digitalización de tesis
Res. SNRD 007/13: Adhesión al SNRD

Proyecto “Mejoramiento cualitativo y cuantitativo de la Biblioteca Digital FCEN-UBA”

Aprobado por Resolución SACT Nº 008/15

Objetivos

1. Elaboración de una **Política de Acceso Abierto** para la FCEN-UBA
2. **Colecciones:** Incorporar: 1200 tesis digitalizadas y 500 artículos. Normalizar la base de datos referencial de 11.000 artículos con referato; Revisar políticas editoriales para sumar artículos en forma retrospectiva
3. **Infraestructura informática**: 2 nuevos servidores y un scanner

Financiación del SNRD para:

- ✓ contratación de 8 becarios (digitalización y marcado de las tesis)
- ✓ compra de equipamiento

Planificación

Relevamiento y Estimaciones

Sección 1	Sección 2	Sección 3	Sección 4	Sección 5
1882 - 1912	1912 - 1952	1953 - 1984	1984 - 1992	1991 - 2004
46 41% c/dup.	647 86% c/dup.	1022 89% c/d	557 94% c/d	1083 95% c/d
120 pag. 5520 pag.	70 pag. 45290 pag.	120 pag. 122640 pag.	180 pag. 100260 pag.	170 pag. 184110 pag.
A4 Imprenta	A4 y oficio Mecanografiado	A4 y oficio. Mecanografiado. Impresora de puntos.	A4 Imprenta. Procesador de texto. Impresora de punto	A4 Procesador de texto. Impresora a chorro.
Volumen a digitalizar: **3355** tesis con **457.820** páginas (76% con duplicados)				
Parámetros de captura: RGB **300** ppp				
Almacenamiento: **11,3 TB** imágenes máster y **17 GB** derivados				
Tiempo estimado: **1508 horas**				

Planificación

Equipo de captura:
Tesis ejemplar único: cama plana de scanner con soporte para cuidado de integridad
Tesis con duplicado guillotinado: alimentador automático

Pre escaneo: limpieza, desanillado, eliminación de ganchos y tachuelas
Tesis con duplicado: guillotinado y guardado en un sobre con identificación

Formatos de archivo: Imágenes master: TIFF de cada página con compresión sin pérdida GIII y GIV
Derivados: PDF con OCR, optimizados para descarga en la Web
Esquema de nombramiento: nomenclatura descriptiva (número consecutivo para cada tesis y apellido del autor

Planificación

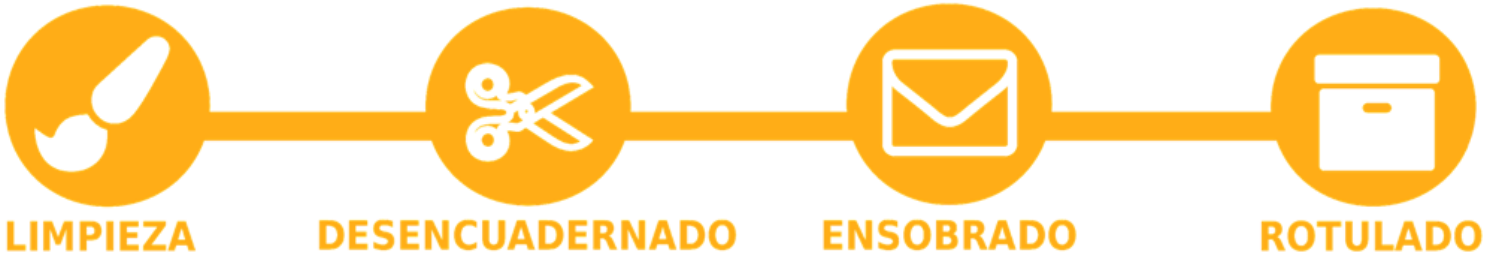
Metadatos: Datos preexistentes en la base bibliográfica en formato BIBUN
Migración a la Biblioteca Digital mediante mapeo a perfil Dublin Core con calificadores propios

Almacenamiento: Sistema en línea para archivos master y derivados, en agrupamiento de discos duros con un sistema RAID 5
La adquisición de dos servidores iguales permite la guarda de imágenes y metadatos en dos lugares físicos diferentes

Estimaciones de tiempo : 1508 horas para digitalización (8 hs. diarias en 9 meses y medio). Basado en 40 minutos promedio para tesis con ejemplar único y 25 minutos para tesis guillotinadas con alimentador automático
Estimación de espacio de almacenamiento : 11,3 TB para imágenes master y 17 GB para derivados

Ejecución

Principales Etapas y Procesos del circuito de publicación hasta la obtención de la versión para publiicación



Limpieza

Aspiradora y pinceles para reducir el polvo.

Se detectaron cuestiones de encuadernación o integridad. Ej. hojas sueltas, tesis sin carátulas, partes a reparar, entre otros.

Desencuadernado

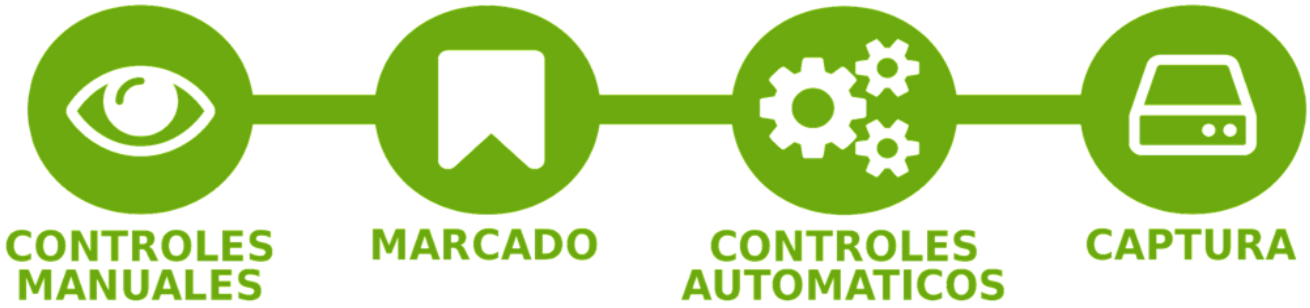
El 76% de las tesis tenían duplicado.

Se desencuadernaron las copias duplicadas: retirar las tapas, espirales, ganchos y carpetas.

Se guillotinó el margen izquierdo para eliminar las irregularidades que podrían atascar el alimentador automático.

Ensobrado y Rotulado

Las hojas resultantes del desencuadernado se colocaron en sobres de papel madera con rótulo.



Captura

Por alimentador automático para duplicados desencuadernados y en cama plana para ejemplares únicos.

Controles Automáticos

Parámetros de captura: Resolución y color.

Nombramiento de los archivos.

Hojas informadas por el operador vs. las capturadas.

Marcado

Identificación de las hojas con imágenes.

Marcado de las secciones para crear la tabla de contenido.

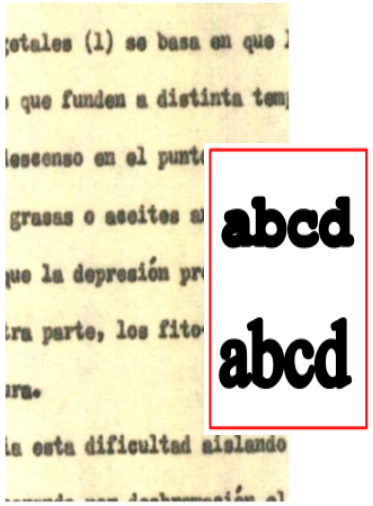
Controles Manuales

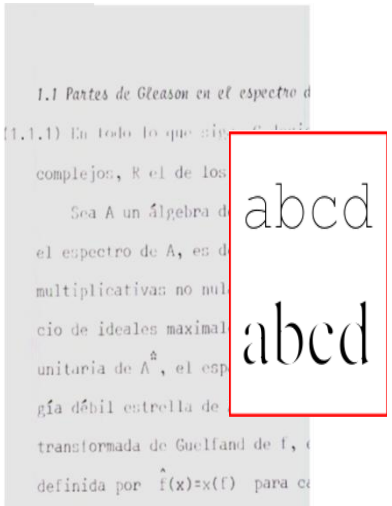
Evaluación de la calidad de las imágenes.

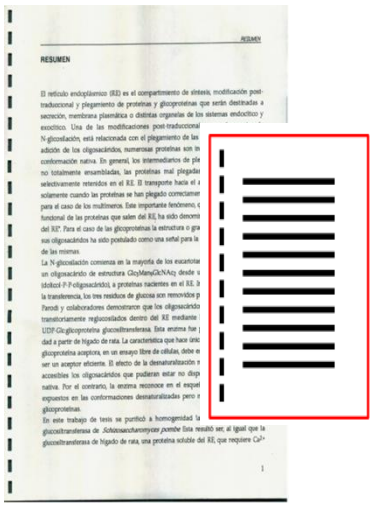
Detección de hojas faltantes o repetidas.

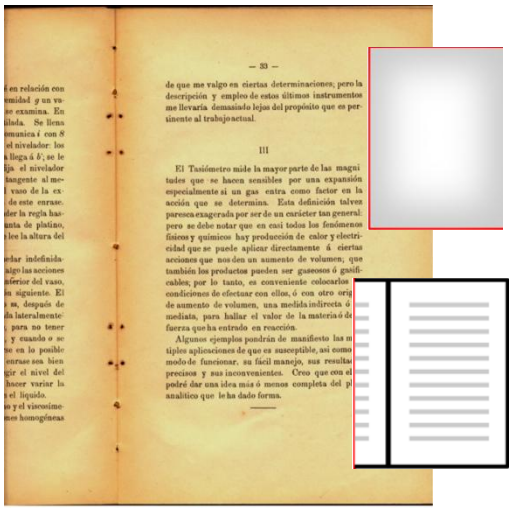
Verificación de los perfiles de filtros de imagen.

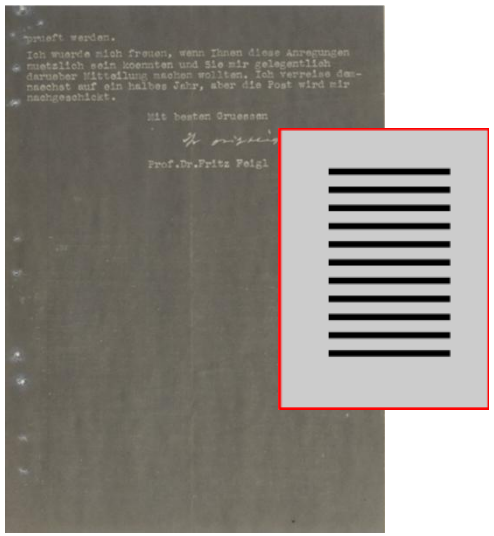
Asignación de Perfiles











Pre-OCR

Aplicación de filtros de imagen.

OCR

ScanTailor para la estructuración del documento y Tesseract como motor de OCR.

Post-OCR

Corrección automática de términos utilizando diccionarios, 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.

Protocolo OAI-PMH para su reutilización por cosechadores (SNRD, BASE, SISBI, etc.)

- Desarrollado en lenguaje Python 3, con el framework Django, librería Celery, Tesseract y ScanTailor
- Asignar operadores, revisores y administradores
- Automatizar y controlar el nombramiento de archivos
- Controlar los parámetros de captura
- Valida el formato de los archivos
- Asignar perfiles de filtro de imágenes
- Marcar secciones e imágenes
- Administrar los procesos automáticos de post-captura:
    - Aplicación de los filtros de imágenes
    - Ejecución de OCR
    - Generación de documentos derivados (portada normalizada, encabezado con metadatos y Bookmarks)
- Obtener datos estadísticos del avance de la tarea.

Ejecución

Recursos humanos

Personal contratado:
8 becarios, 12 hs. semanales cada uno, 9 meses. Digitalización, marcado de documentos, enriquecimiento de metadatos
1 especialista en imagen digital, 6 meses. Definición de los filtros de mejora de las imágenes para aumentar la calidad del OCR

Personal de la FCEN-UBA:
1 responsable de proyecto, tiempo completo. Planificación y dirección del proyecto
2 asistentes 30 hs. semanales y 1 asistente 20 hs. semanales. Preparación de materiales, supervisión y enriquecimiento de metadatos
1 técnico informático y 1 asistente 20 hs. semanales. Desarrollo y mantenimiento de software

Resultados

- ✓ **3211 tesis digitalizadas, que compondrán una colección de 5000 tesis a fines de 2017**
- ✓ **Dos puestos de escaneo**
- ✓ **Software para la gestión de los procesos de digitalización**
- ✓ **Dos nuevos servidores y aumento de la capacidad de almacenamiento (16 TB)y de procesamiento**
- ✓ **6 personas capacitadas y con experiencia en digitalización y procesos de marcado y edición digital**

Resultados

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: Disponible para su transferencia, adaptación y reutilización por otras instituciones integrantes del SNRD. 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**

Reflexiones

- ✓ **El estudio pormenorizado del material a digitalizar permitió calcular en forma precisa tiempos de captura y volumen de almacenamiento . También posibilitó seleccionar estrategias que abreviaron tiempos de captura y volumen de almacenamiento. La ejecución del proyecto se mantuvo dentro de los tiempos estimados**

- ✓**Distribución de los tiempos: 20%: Preparación de materiales; 40%: captura; 40%: enriquecimiento de metadatos, marcado de estructura, identificación de imágenes**

- ✓**El software de gestión redujo fuertemente los tiempos de captura y en menor medida, los procesos post captura**

- ✓**La posibilidad de utilizar el alimentador automático en el 76% de la colección redujo en forma determinante la captura**

Reflexiones

- ✓ **La organización del trabajo de los 8 becarios con 12 hs. semanales de dedicación, distribuidas en tres jornadas de 4 hs. en 3 puestos de trabajo y el software de gestión resultó un modo eficiente de coordinación y maximización de la producción**

- ✓**La aplicación de los filtros de imágenes mejoró notablemente la calidad del OCR, con buenos resultados en la indexación por parte de los buscadores Web**

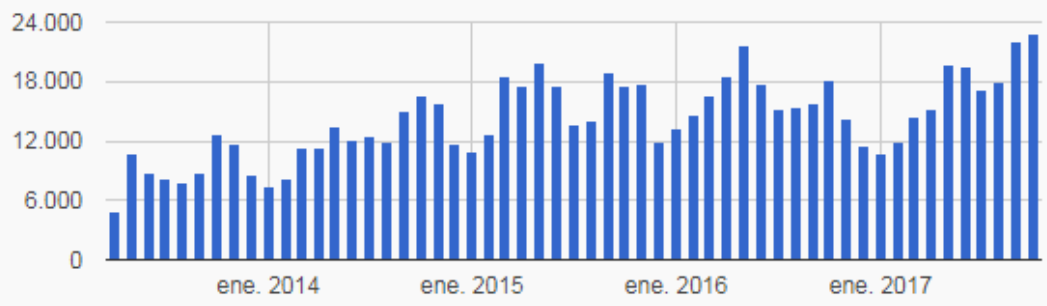
Tesis Doctorales y de Maestría FCEN-UBA

Colección de tesis doctorales y de maestría aprobadas en el ámbito de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires, con expresa conformidad de los tesistas que completaron y firmaron el Formulario de Autorización.

Datos de la Colección

Titulo	:	Tesis FCEN-UBA
Documentos	:	3020 tesis en formato pdf
Creada	:	Marzo de 2010
Actualizada	:	Agosto de 2014
Url	:	http://digital.bl.fcen.uba.ar/gsdl-282/cgi-bin/library.cgi?p=about&c=tesis

Estadística de uso y crecimiento de la colección



La digitalización retrospectiva y puesta en Acceso Abierto del conocimiento producido por una institución universitaria adquiere valor científico e histórico.

En este caso, la FCEN-UBA abrió a la comunidad uno de los canales principales del conocimiento científico que ha producido en toda su historia.

Muchas gracias

Martín Williman
mwilliman@bl.fcen.uba.ar
Ana Sanllorenti
asanllorenti@bl.fcen.uba.ar