# EL SENDERO DE LAS TESIS: Procesos de captura, OCR y publicación

Callao, Sofía · Fiorotto, Florencia · Montepagano, Pablo · Valente, Nicolás
Cruz, Micaela · Grimberg, Daniela · Rozán, Eric A. · Valdez, Elizabeth
Díaz, Yamila B. · Gooderham, Alicia · Sobio, María · Vannucci, Nancy
Schneider, Gabriela · Vázquez, Cristian
Sanllorenti, A. M. (Resp. Técnico) · Willimon, J. M. (Resp. Biblioteca Digital)

## Proyecto de digitalización retrospectiva de las Tesis de post-grado de la FCEN-UBA hasta 1999

| | |
|---|---|
| Orientación | Preservación y acceso |
| Comunidad | Investigadores, docentes, estudiantes y sociedad en general |
| Marco legal | Res. CD 2053/05, 2533/09 y 0727/13<br>La Biblioteca Central es depositaria de los trabajos de post-grado FCEN-UBA, los autores pueden autorizar el depósito en el Repositorio y después de 15 años de su defensa las tesis pueden ser digitalizados y publicadas |
| Recursos Humanos | 8 becarios con 12 horas semanales durante 8 meses<br>personal de la institución afectado: 6 |
| Fuente Financiadora | SNRD-MINCYT. Proyecto RDF3 "Mejoramiento cualitativo y cuantitativo de la Biblioteca Digital de la FCEN-UBA", aprobado por Resolución SACT 008/15 |
| Costos RH | $ 200.000 (MINCYT) - $ 8.000 (FCEN-UBA) |
| Costos Insumos | $ 148.236 (MINCYT) - $ 22.897 (FCEN-UBA) |

| | |
|---|---|
| Volumen a digitalizar | 2333 Tesis (299900 pág.) con duplicados, por alimentador automático<br>292 Tesis (29753 pág.) ejemplares únicos, de forma manual |
| Parámetros de Captura | 300 dpi Color |
| Equipos de Captura | Kodak i2400 y Avision AV1860 |
| Software | Software de seguimiento desarrollado en el proyecto<br>Drivers de los equipos de captura<br>IrfanView, software libre con licencia .. |
| Formato de los ficheros | TIFF para masters<br>JPG y PDF para derivados |
| Nombramiento de archivos | Tesis_[nº secuencial para cada tesis]_[Apellido del autor]_[extensión correspondiente] |
| Almacenamiento | RAID10 (8TB) |

## Preparación de los Documentos

Limpieza con aspiradora y pinceles para reducir el polvo de los ejemplares mas antiguos

Desencuadernado y guillotinado de los duplicados para su captura por alimentador automático

Colocado en sobres para mantener la integridad de la unidad documental

Rotulado y ordenamiento

## Captura

Acceso al Software de Seguimiento* e identificación del documento a capturar

Estimación de la cantidad de imágenes a capturar

Asignación de los Perfiles para el tratamiento post-captura de las imágenes según las particularidades del material

Captura de las imágenes por alimentador automático para duplicados desencuadernados y de forma manual para ejemplares únicos


Controles del software:
- 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 no ser alteradas por los procesos automáticos post-captura según el perfil informado

Marcado del documento en sus secciones

## Control

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

## Pre-OCR Modificación de las imágenes para mejorar la eficacia del motor OCR

Aplicación de Filtros para mejorar la separación entre figura (carácter) y fondo (papel). Eliminación de ruido y manchas, uniformando el fondo y mejorando el grosor de la fuente entrecortada, muy suave o empastada.

## OCR Reconocimiento Óptico de Caracteres

Se utiliza ScanTailor para la estructuración del documento

Se utiliza el motor de OCR Tesseract

## Post-OCR Modificaciones sobre los textos orientadas a detectar y corregir los errores de reconocimiento de caracteres.

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.)

### • Software de Seguimiento. Desarrollo propio.

Para el proyecto se desarrollo en la institución un software utilizando el Framework Django que permite:

- Asignar estados para los documentos en proceso (Digitalizando, Marcando, Pendiente, Aceptada; Rechazada, Terminada, etc.)
- Asignar operadores, revisores, administradores y puestos de escaneo.
- Automatizar y controlar el nombramiento de archivos.
- Controlar la correcta asignación de los parámetros de captura.
- Marcar secciones y ubicar imágenes o gráficos en los archivos capturados.
- Controlar los trabajos realizados. Solicitar correcciones al operador correspondiente.
- Obtener datos estadísticos del avance de la tarea.
- Administrar los procesos automáticos de post-captura hasta la obtención del derivado para publicación.

**¿Te interesó? ¿Querés que te contemos mas del proyecto?**
**¡Contactanos!**
digital@bl.fcen.uba.ar

12ª Jornada de las Bibliotecas y Centros de Documentación de la Universidad de Buenos Aires
25 de agosto de 2016
SISBI