

Mejora en los textos obtenidos por digitalización y OCR

Michel Genereux, Juan Pablo Suárez

Martín Williman, Pablo Montepagano, Nicolás Rucks, Ana Sanllorenti ()*

En la construcción de bibliotecas digitales y repositorios, los procesos de digitalización son una de las fuentes principales de incorporación de documentos. Actualmente se han convertido en una tarea cotidiana que se lleva a cabo con estándares de calidad reconocidos. Sin embargo, con frecuencia en la actualidad, los repositorios digitales están experimentando dificultades para la explotación de los textos contenidos en los documentos producto de la digitalización debido a cuestiones derivadas de la mala calidad de los resultados en los procesos de Reconocimiento Óptico de Caracteres (OCR) a partir de las imágenes.

Un mal índice de reconocimiento de caracteres afecta la indexación de los trabajos por parte de los motores de búsqueda, su posicionamiento en la recuperación de información y posibles aplicaciones de minería de datos.

Con el objetivo de mejorar los resultados del OCR sobre las imágenes máster obtenidas en la digitalización de documentos, en la Biblioteca Digital de la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires se llevó a cabo una exploración del problema.

Con un criterio experimental, se encontraron dos tipos de abordajes, según el momento de intervención en el transcurso del flujo de procesamiento de imágenes con texto.

Una de ellas es la intervención en la naturaleza y calidad de las imágenes como para que los programas OCR discriminen correctamente los caracteres. Se trata de las intervenciones **Pre-OCR**, que son aquellas que se realizan sobre las imágenes obtenidas luego del proceso de captura y están dirigidas a modificar específicamente características de las imágenes que mejoran la eficacia del funcionamiento del software de OCR.

Por otro lado, las intervenciones **Post-OCR**, son aquellas que se aplican sobre los textos obtenidos luego del proceso de OCR y están orientadas a detectar y corregir los errores cometidos por el software de reconocimiento de caracteres.

Se realizaron pruebas de intervenciones **Pre-OCR** sobre las imágenes máster obtenidas de la digitalización de 400 tesis doctorales que consistieron en aplicar filtros orientados a facilitar la capacidad de discriminación y reconocimiento por parte del software OCR, tales como el aumento de contraste, eliminación de manchas, identificación de imágenes y gráficos.

Como intervenciones **Post-OCR** se utilizaron listas de frecuencia de términos específicos del área temática y listas de frecuencia de errores del software OCR.

Los resultados obtenidos se compararon con los resultados sin las intervenciones y con patrones, obteniéndose mejoras significativas.

(*)

Michel Genereux, investigador

EURAC, Bolzano, Italia

Juan Pablo Suárez, consultor en procesos de digitalización

Martín Williman, Responsable de la Biblioteca Digital FCEN-UBA

Pablo Montepagano, Redes y Sistemas

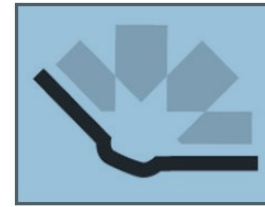
Nicolás Rucks, Bibliotecario de Sistemas

Ana Sanllorenti, Subsecretaria de Biblioteca

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

Michel Genereux, investigador

EURAC, Bolzano, Italia



JORNADA
Temas Actuales
en Bibliotecología

VI TAB, Centro Médico de Mar del Plata,
13 de noviembre, 2015

Mejora en los textos obtenidos por digitalización y OCR

Michel Généreux, EURAC Research, Bolzano, Italia
Juan Pablo Suárez, consultor en procesos de digitalización
Martín Williman, Pablo Montepagano, Nicolás Rucks, Ana
Sanllorenti, Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires

Nuestro proyecto



Proyecto de Digitalización de 3500 tesis doctorales (financiación del SNRD)

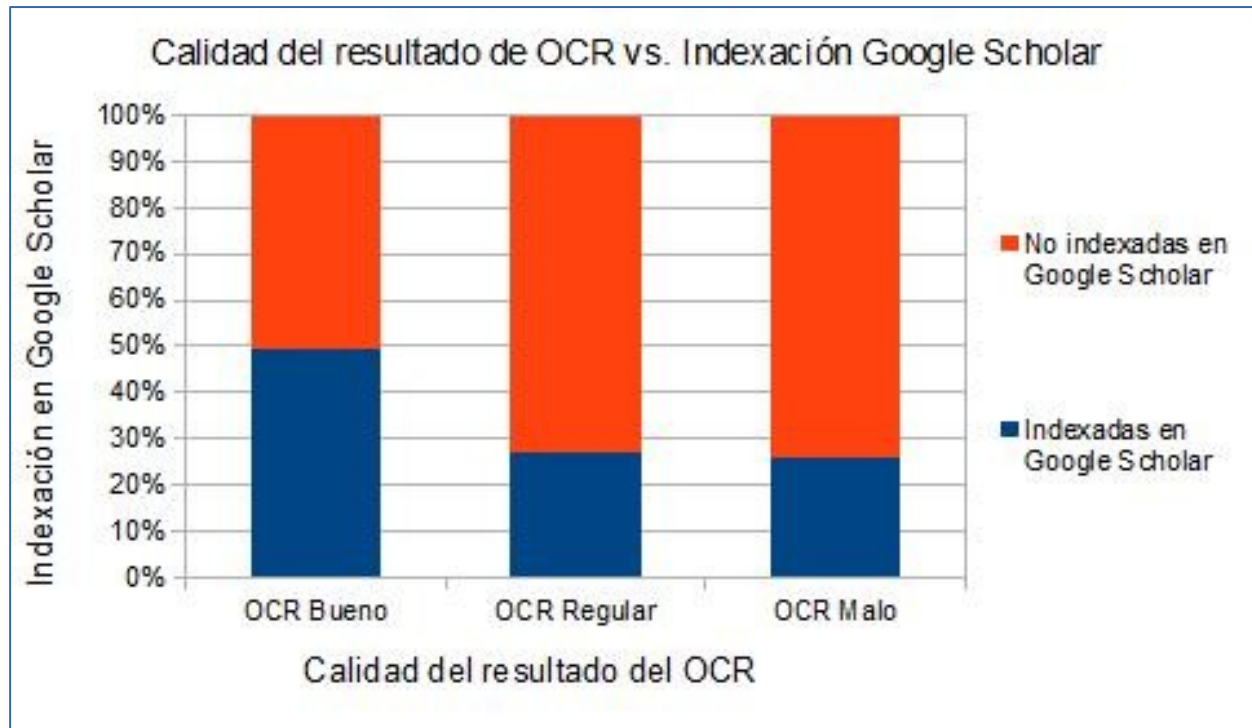
- ✓ Actualmente: 1374 tesis (400 digitalizadas)
- ✓ Al finalizar : 80 % de la colección estará constituida por tesis digitalizadas
- ✓ Cambio del perfil de la colección: documentos mayoritariamente digitalizados
- ✓ Consecuencias particulares respecto de su tratamiento y visibilidad

El texto de los documentos digitalizados

- ✓ Documentos digitalizados en PDF: más pesados
- ✓ Recuperar información, cortar y pegar, buscar palabras en el texto: dependen de la calidad de los procesos de OCR
- ✓ Alta tasa de error en OCR:
 - ✓ Obstaculiza la visibilidad porque es uno de los factores de los que depende la indexación en los buscadores (baja calidad del texto, baja indexación)
 - ✓ Reduce el rendimiento de funcionalidades de recuperación de información del software de la biblioteca digital

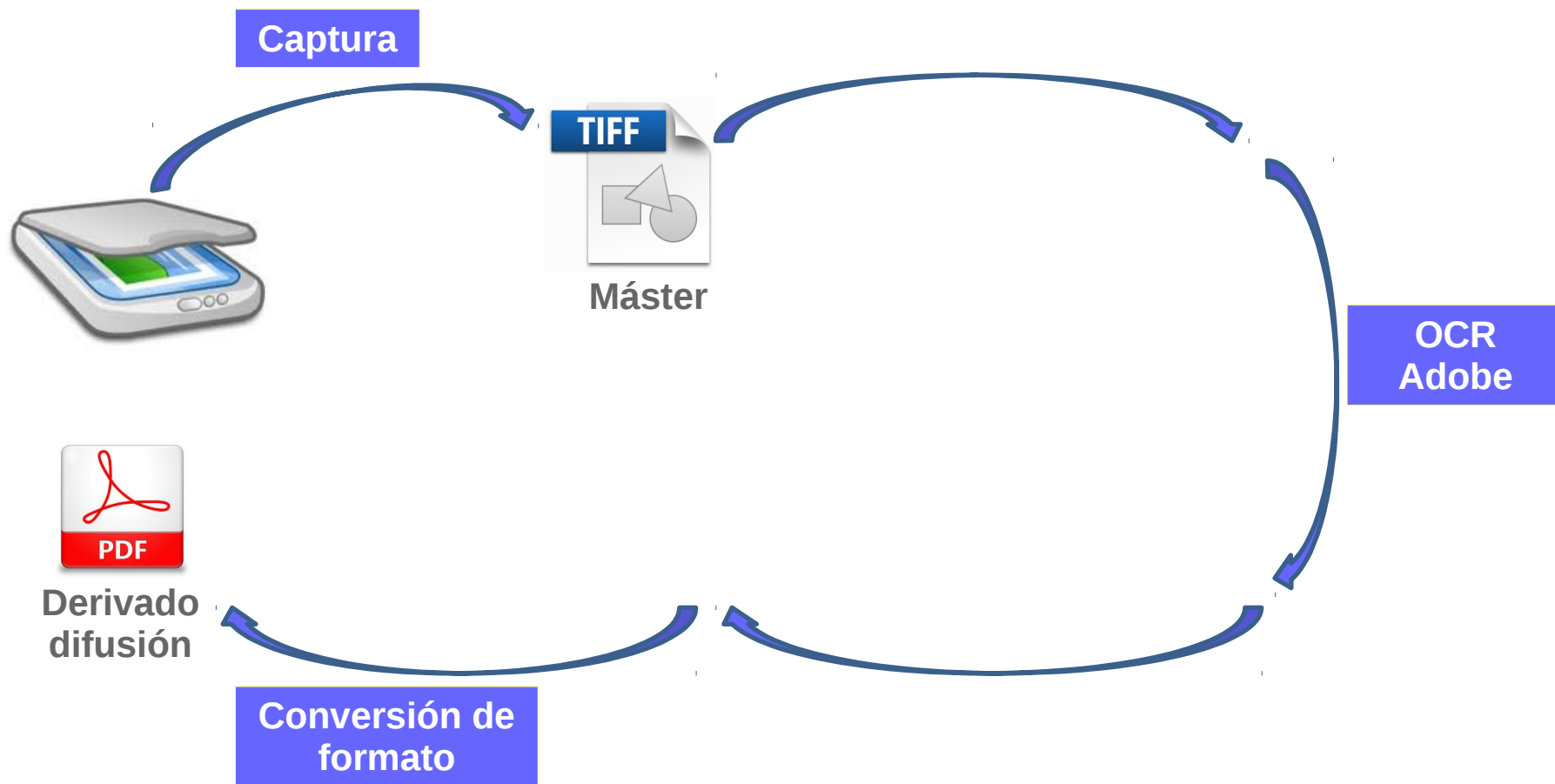
Consecuencias de un mal OCR

% de indización de nuestras tesis en Google Scholar

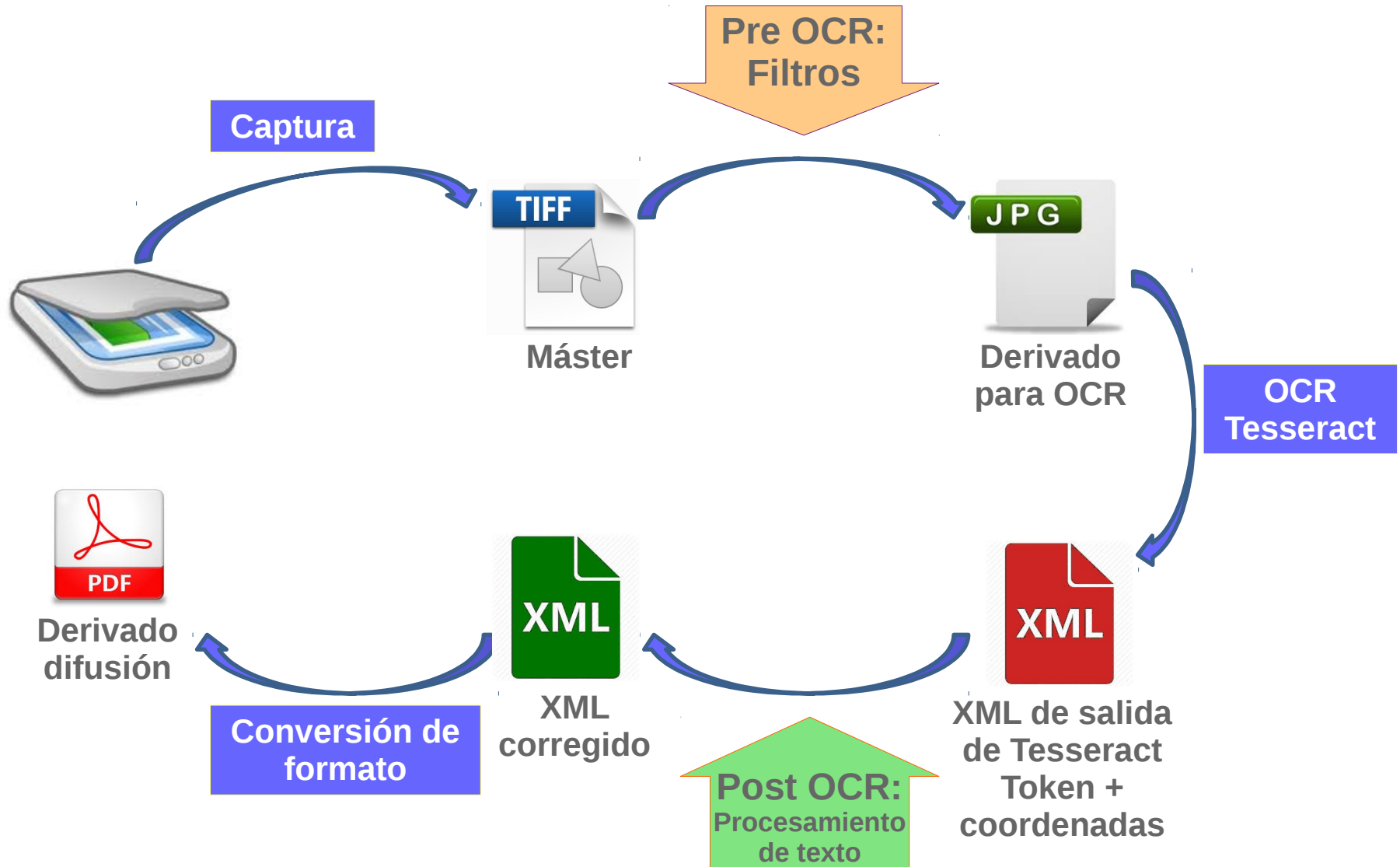


El texto indexable producto del OCR es un factor importante en la probabilidad de que un documento sea indexado por Google Scholar

Dos líneas de intervención



Dos líneas de intervención



Dos líneas de intervención

Pre-OCR: Se realizan sobre las imágenes obtenidas a través del proceso de captura y están dirigidas a modificar características de las imágenes que pueden mejorar la eficacia del posterior funcionamiento del software de OCR

Post-OCR: Se aplican sobre los textos obtenidos luego del proceso de OCR y están orientadas a detectar y corregir los errores cometidos por el software de reconocimiento de caracteres.

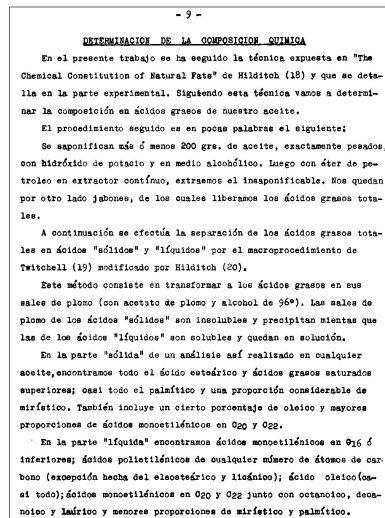
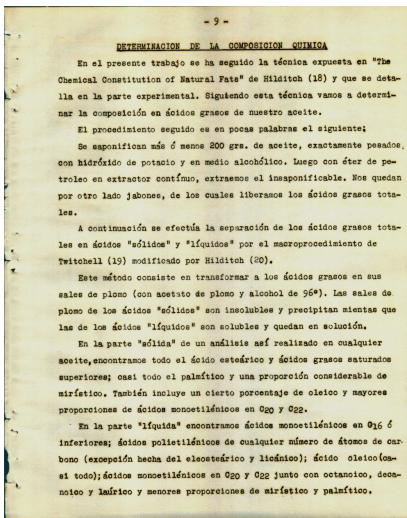
Intervenciones PreOCR:

1. Separación fondo/figura

1.1 Ajuste umbral: Conversión a bitonal

Elegir un umbral para determinar desde que valor de intensidad de luz un píxel es convertido a blanco o a negro al momento de generar el derivada bitonal para OCR.

- a) Se utilizan **Histogramas de Imagen**
- b) Aplicación de **Algoritmos de Thresholding** (umbral)



Y: Cantidad de píxeles



Histograma de Imagen

X: Intensidad de luz

Intervenciones PreOCR:

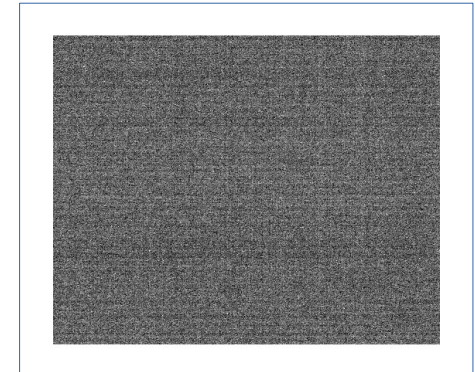
2. Factores que inciden sobre el fondo

2.1 Ruido

Aparición aleatoria de píxeles inexistentes en el original durante el proceso de captura.

Aplicación de **Filtros Denoise**, compara cada píxel con los píxeles de su entorno para corregir su valor.

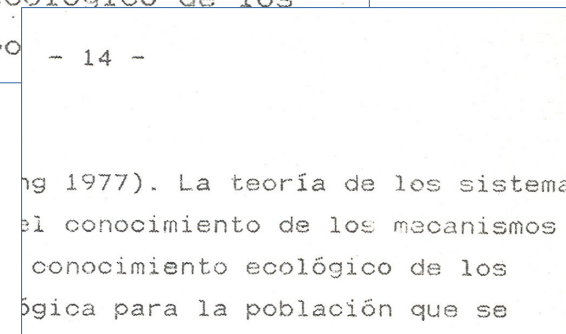
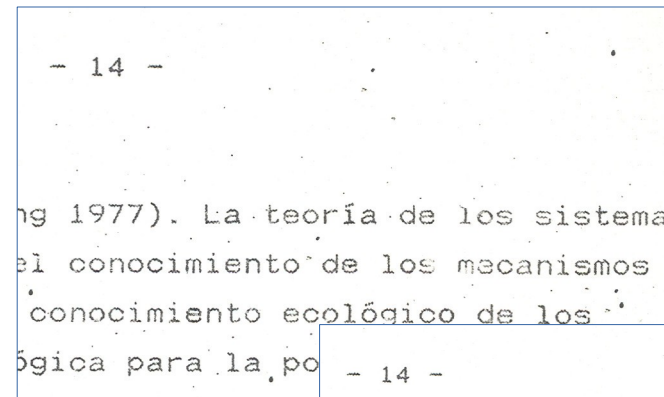
Riesgo: Efecto negativo de suavizado de la imagen (blue).



2.2 Puntos/Manchas negras

Aplicación de **Filtros Despekle** que eliminan este tipo de estructuras fijando un umbral mínimo de tamaño.

Riesgo: Eliminar elementos válidos. Para ello, el filtro despekle pondera la proximidad de los puntos negros a otras estructuras mayores (caracteres).



Intervenciones PreOCR:

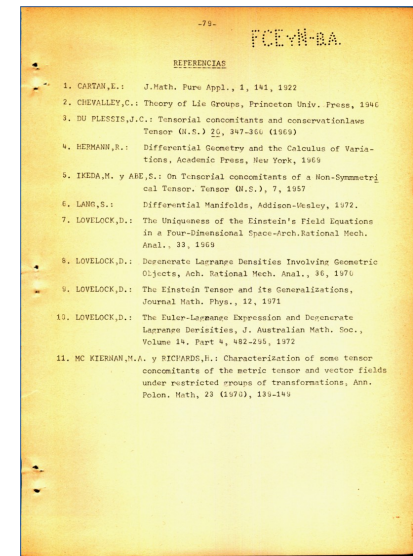
2. Factores que inciden sobre el fondo

2.3 Variaciones suaves de color en el fondo

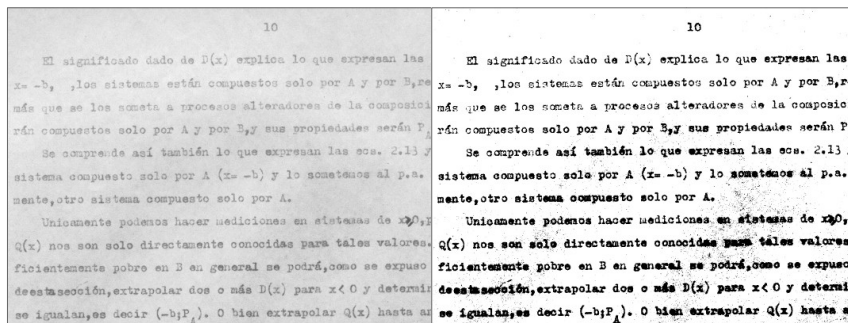
El fondo no es uniforme en los originales en papel, sino que presenta suaves alteraciones entre diferentes zonas de la hoja.

Aplicación de **Filtro High Pass** o paso alto.

Riesgo: Aumento del ruido.



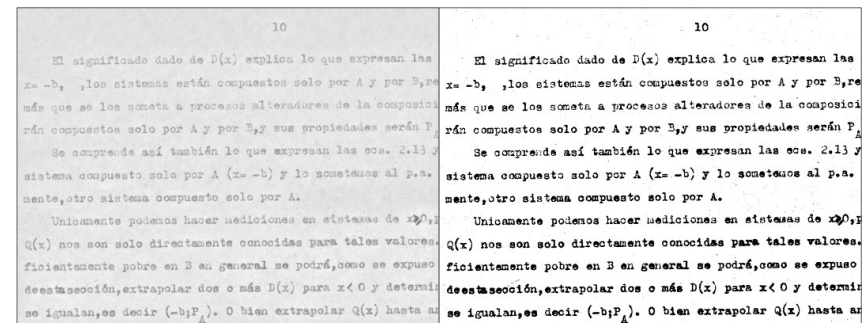
Sin Filtro High Pass



Master Tiif

Derivado b/n, texto
empastado y
entrecortado

Con Filtro High Pass



Derivado con High Pass

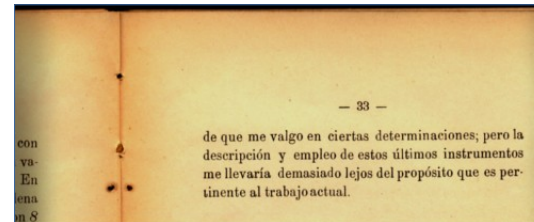
Derivado b/n, se
reducen el
empastado y el
entrecortado

Intervenciones PreOCR:

3. Factores que inciden sobre la figura

3.1 Estructuras sobrantes

Elementos en la imagen que no necesitan preservarse (marcas de anillados, los bordes de la página, o fragmentos de la página contigua).



3.2 Orientación

La ortogonalidad de los bloques de texto.

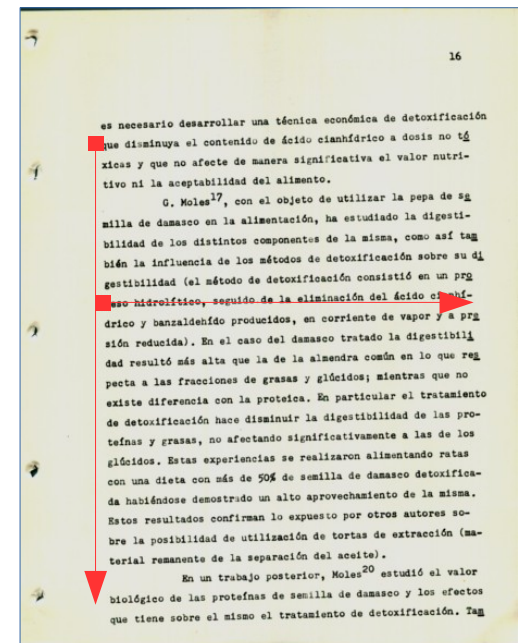
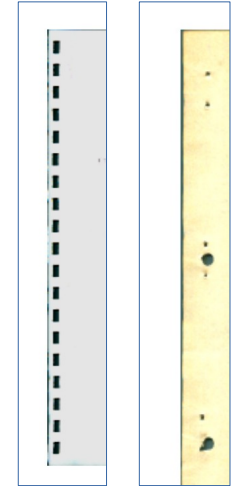
3.3 Grosor

El grosor de los caracteres. Originales en papel con tipografía empastada o de astas delgadas.

Aplicación de **Filtros de Erosión/Dilatación** que agregan o quitan píxeles alrededor de los caracteres
Riesgo: Afectar a partes del texto normales.

3.4 Bordes difusos

Aplicación de Filtros Deblurring



Intervenciones PreOCR: Aplicación de Perfiles

Perfiles de escaneo: Bordes y Perforaciones

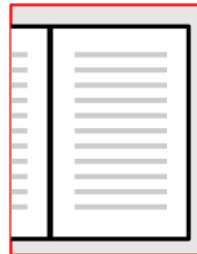
A1: Sin Borde



A2: Borde



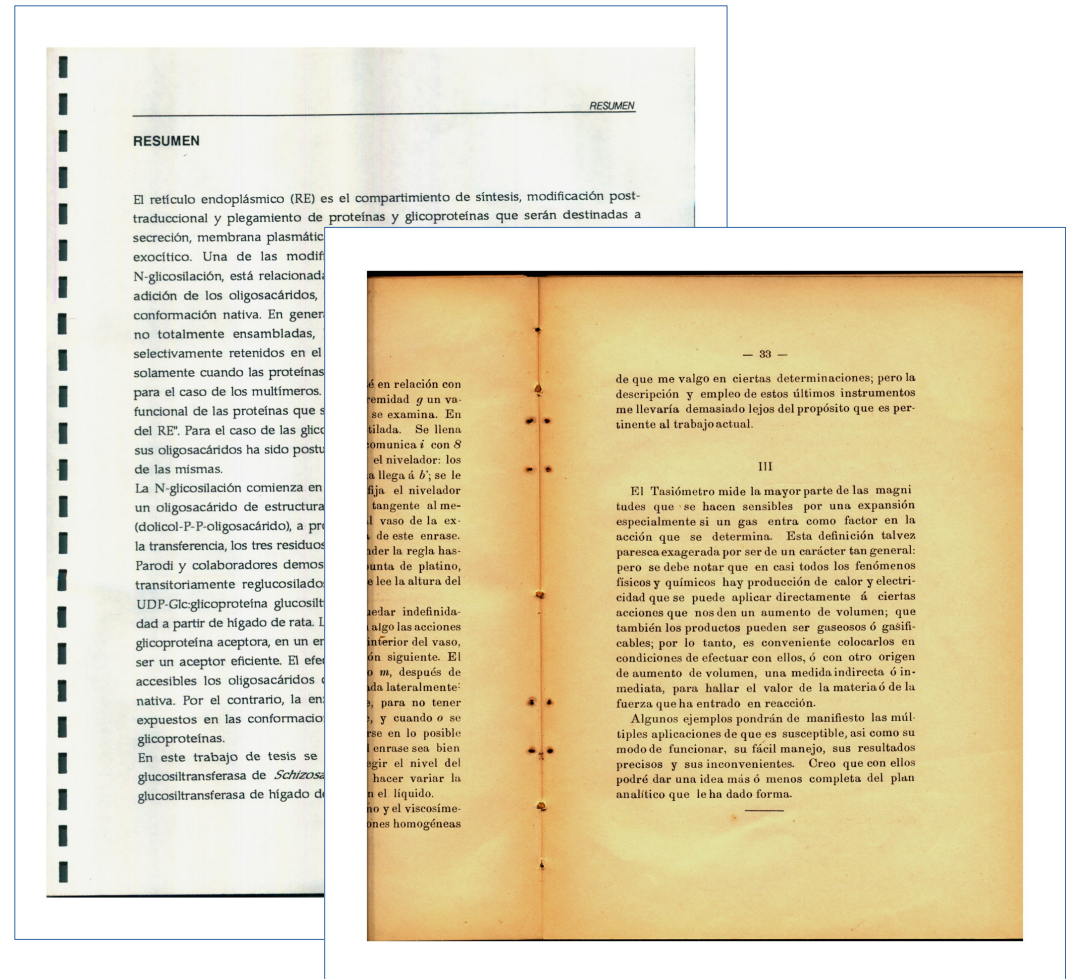
A3: texto
acompañante



A3: Borde extra



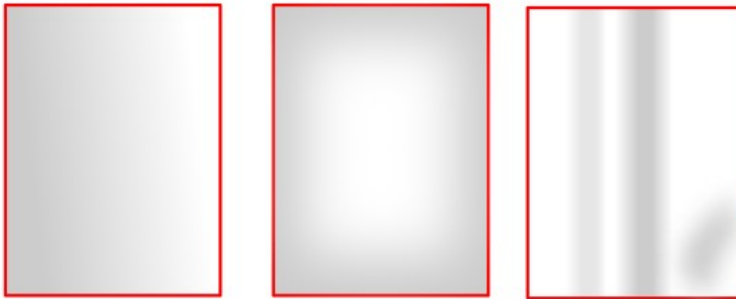
D1: Perforaciones



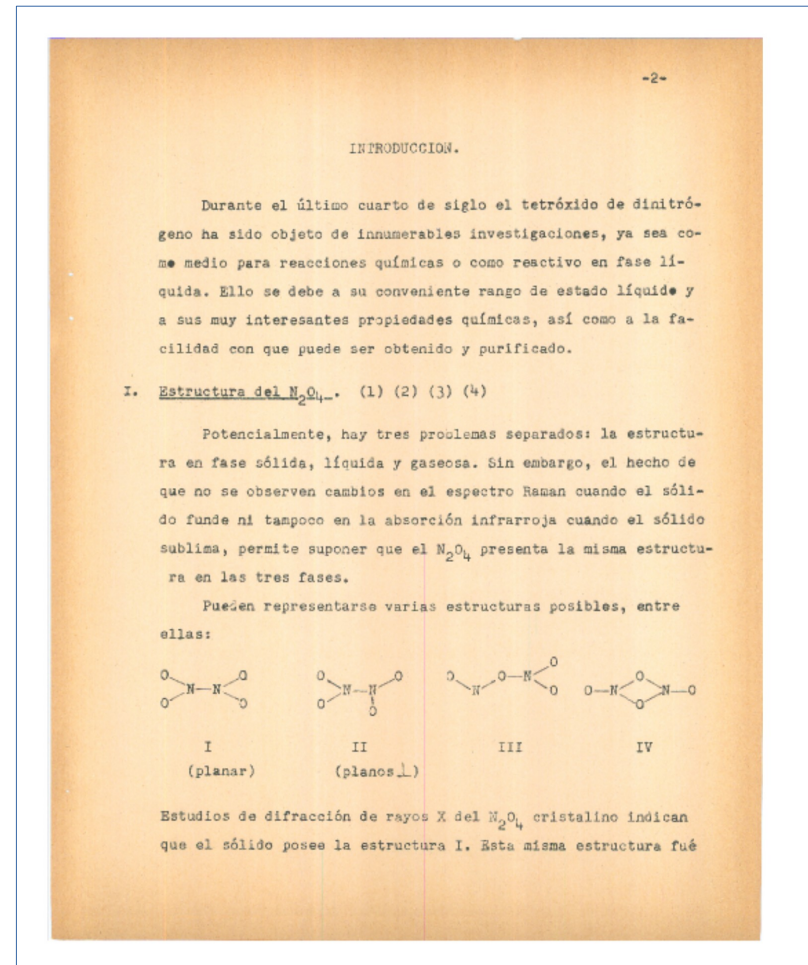
Intervenciones PreOCR: Aplicación de Perfiles

Perfiles de escaneo: Fondo

B1: Variaciones suaves de color



B2: Manchas



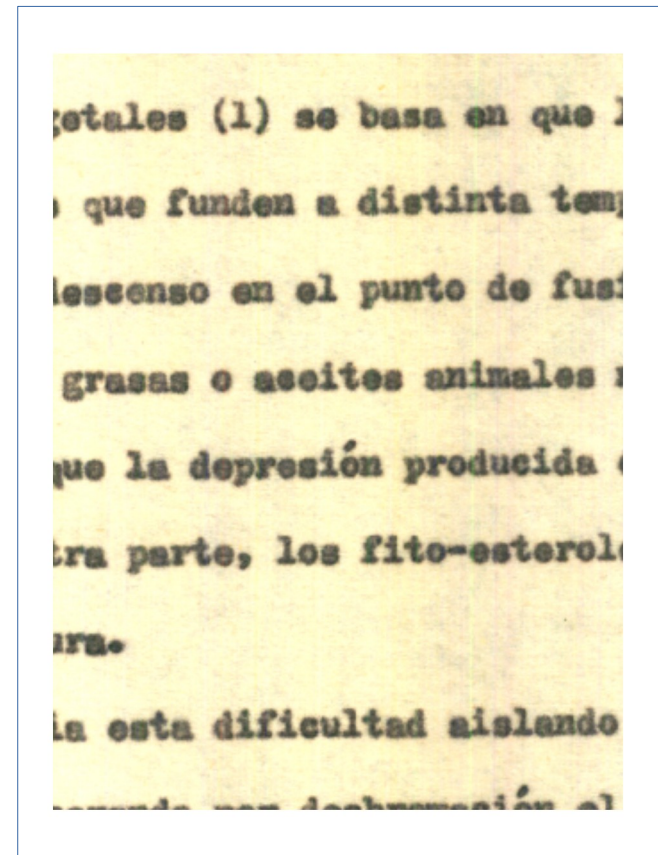
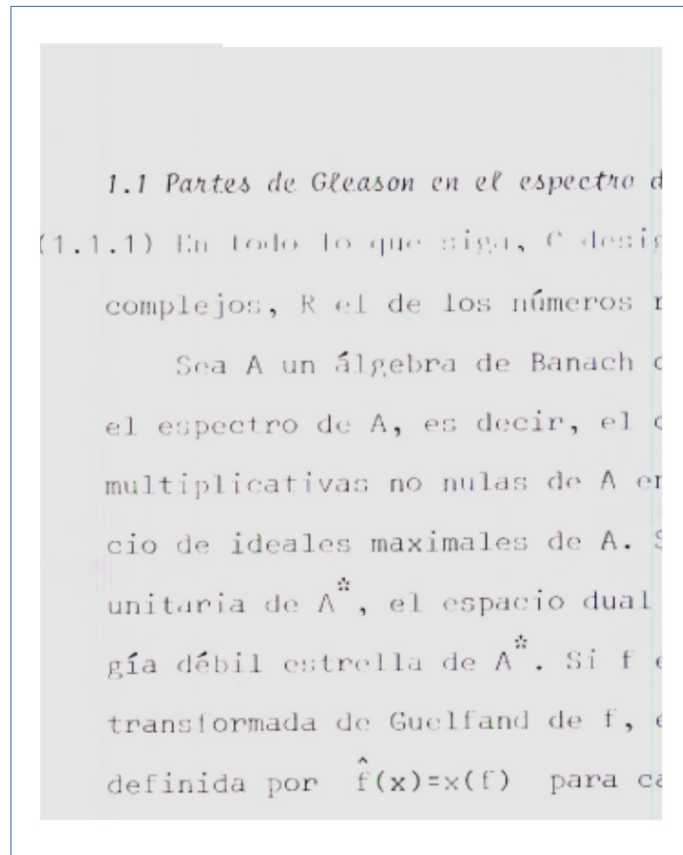
Intervenciones PreOCR: Aplicación de Perfiles

Perfiles de escaneo: Texto

C1: Texto fino



C2: Texto grueso



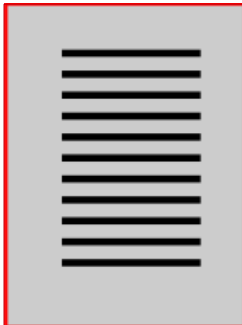
Intervenciones PreOCR: Aplicación de Perfiles

Perfiles de escaneo: Contraste y Color

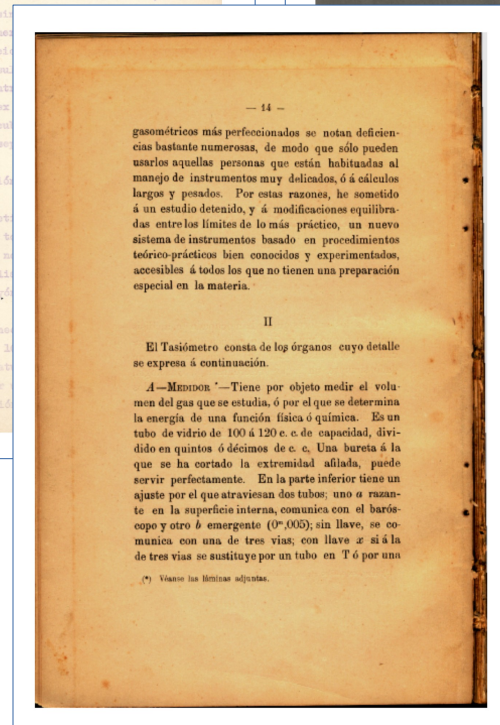
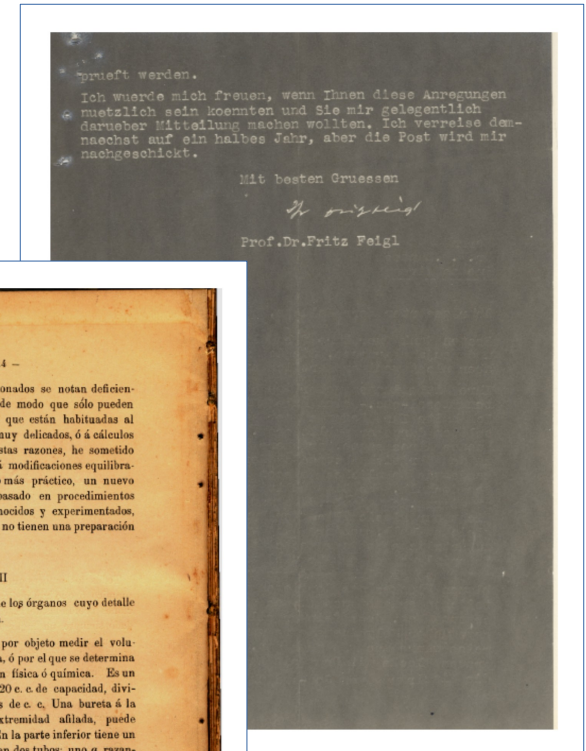
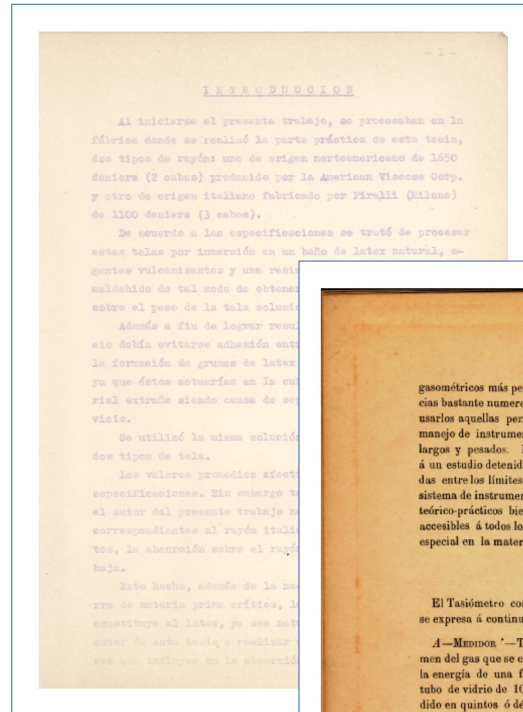
F1: Bajo contraste



G1: Oscuro



G2: Inverso



Intervenciones PostOCR: Corrección automática de errores

“El reemplazo de un término identificado por un proceso automático como **error**, por otro término que se supone como su correspondiente **correcto**”

Esto implica dos acciones:

- ✓ Identificar los errores
- ✓ Corregirlos con el mejor término candidato posible

Recursos lingüísticos:

1. Listas de términos correctos
2. Listas de errores frecuentes
3. Listas de correspondencia: error/corrección

Intervenciones PostOCR: **Recursos lingüísticos**

1. Listas de términos correctos: **14.252.173** de *bigramas*

- ✓ Diccionarios español e inglés. Listas de abreviaturas.
- ✓ Documentos del dominio disciplinar
- ✓ Bases de datos bibliográficos de bibliotecas científicas
- ✓ Listas de términos científicos muy específicos en español

2. Listas de errores frecuentes

- ✓ Identificación automática de términos no aparecen en el diccionario
- ✓ Corrección manual de documentos. Lista de los errores de reconocimiento más comunes

Ejemplo: *Reemplazo de una “l” por un “1” ó una “i”*

3. Listas de correspondencia: error/corrección

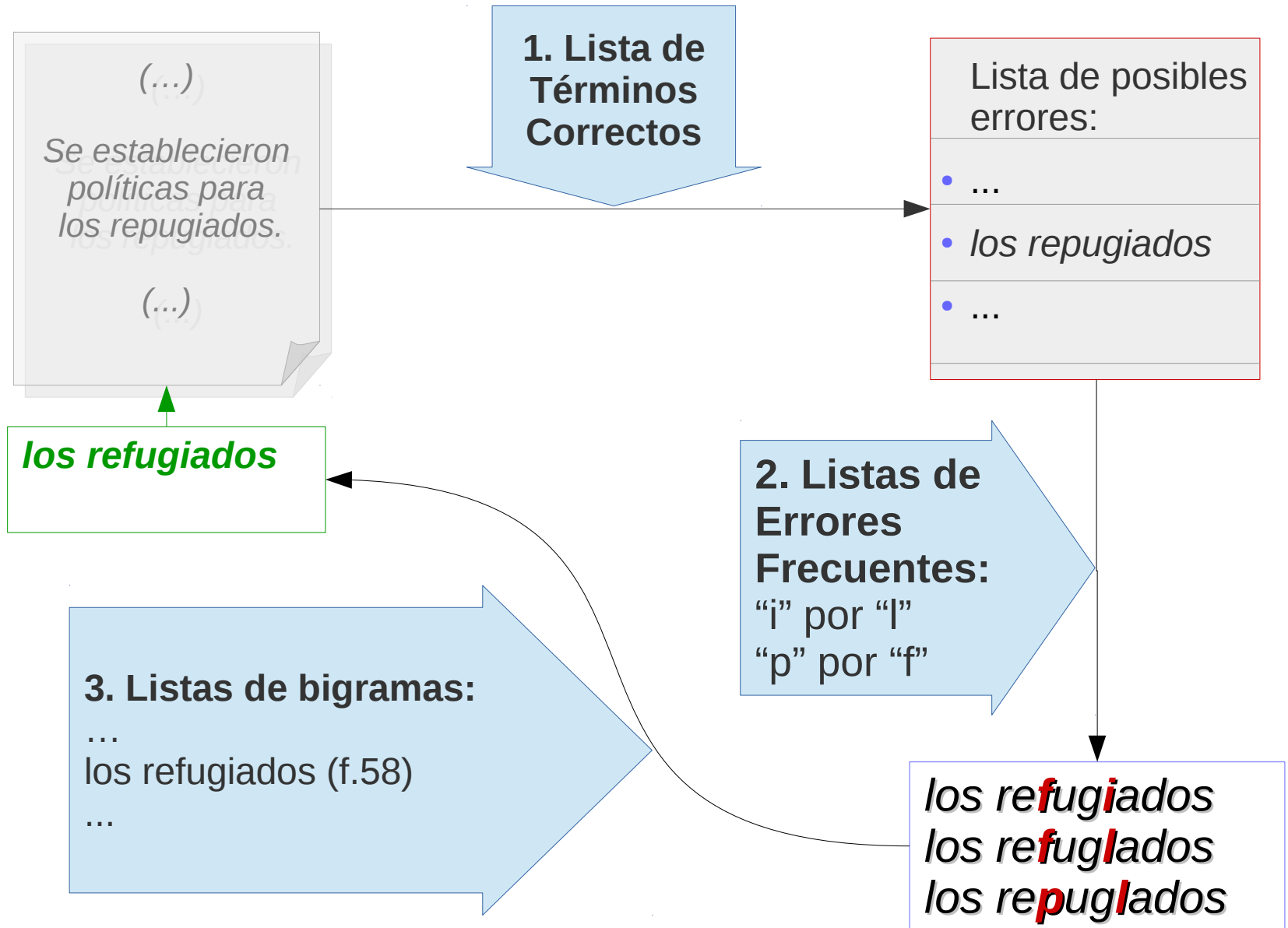
- ✓ A la lista de bigramas de errores frecuentes se les agregaro manualmente la palabra correspondiente correcta

Ejemplo: *rocas paleomagnéticas - rocas paleomagnéticas*

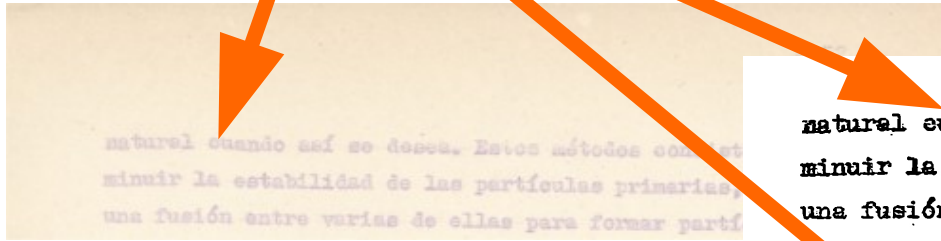
Intervenciones PostOCR: **Lista de correspondencia**

	A	B	C	D	E	F	G
1	frecuencia	contexto	error	no es un error	corrección	no lo sé	Longer context
2	2250	de	1a	si	la		infeccioso de 1a septicemia, po
3	986	en	1a	si	la		do iso-oléico en 1a fracción de
4	558	a	1a	si	la		perteneciente a 1a familia de las
5	389	que	1a	si	la		lasma, se nota que 1a materia
6	333	de	HPLY	no			fosforilación de HPLY con NetPhosK "
7	224	por	1a	si	la		(ll usas dadq por 1a
8	221	y	1a	si	la		llada y 1a comparamos con
9	192	con	1a	si	la		1a comparamos con 1a de aceites nac
10	156	sobre	1a	si	la		;on evidencias sobre 1a estabilidad de
11	148	es	1a	si	la		) con ¿res veces 1a cantidad teór
12	135	para	1a	si	la		u m y coro. mo para 1a- rminnn a. me

Intervenciones PostOCR: Corrección



OCR (Optical Character Recognition)



natural cuando así se desea. Estos métodos consisten en disminuir la estabilidad de las partículas primarias, causando una fusión entre varias de ellas para formar partículas ma-

```
Tesis_0890_Lantos_page058_edited.html
1<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4.01/loose.dtd">
2<html>
3<head>
4<title></title>
5<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
6<meta name="ocr-system" content="tesseract" />
7</head>
8<body>
9<div class="ocr_page" id="page_1" title="Image 'Tesis_0890_Lantos_page058_edited.tif'; bbox 0 0 2104 2856">
10<div class="ocr_carea" id="block_1_1" title="bbox 117 115 1977 800">
11<p class="ocr_par" dir="ltr" id="par_1" title="bbox 118 118 1975 612"><span class="ocr_line" id="line_1" title="bbox 122 118 1975 176"><span class="ocr_word" id="word_1" title="bbox 122 118 1975 176"><strong>natural/></strong></span> <span class="ocr_word" id="word_3" title="bbox 373 122 556 160"><strong>e cuando/></strong></span> <span class="ocr_word" id="word_5" title="bbox 589 118 676 160"><strong>asi/></strong></span> <span class="ocr_word" id="word_7" title="bbox 708 118 774 160"><strong>se/></strong></span> <span class="ocr_word" id="word_9" title="bbox 804 122 881 162"><strong>de sea.</strong></span> <span class="ocr_word" id="word_11" title="bbox 1201 118 1416 176"><strong>métodos/></strong></span> <span class="ocr_word" id="word_13" title="bbox 1449 122 1728 161"><strong>consistent/></strong></span> <span class="ocr_word" id="word_15" title="bbox 1756 124 1975 163"><strong>en día.</strong></span> </span>
12<span class="ocr_line" id="line_2" title="bbox 118 185 1975 265"><span class="ocr_word" id="word_17" title="bbox 118 207 1975 248"><strong>minuir/></strong></span> <span class="ocr_word" id="word_19" title="bbox 336 206 401 245"><strong>las/></strong></span> <span class="ocr_word" id="word_21" title="bbox 434 187 771 248"><strong>estabilidad/></strong></span> <span class="ocr_word" id="word_23" title="bbox 885 206 922 245"><strong>de/></strong></span> <span class="ocr_word" id="word_25" title="bbox 896 206 922 245"><strong>las/></strong></span> <span class="ocr_word" id="word_27" title="bbox 1016 206 1975 265"><strong>partículas/></strong></span> <span class="ocr_word" id="word_29" title="bbox 1355 206 1975 265"><strong>primarias,</strong></span> </span>
13<span class="ocr_line" id="line_3" title="bbox 118 269 1975 339"><span class="ocr_word" id="word_31" title="bbox 118 269 1975 339"><strong>causando/></strong></span> </span>
14<span class="ocr_line" id="line_4" title="bbox 118 348 1975 428"><span class="ocr_word" id="word_33" title="bbox 118 348 1975 428"><strong>una/></strong></span> <span class="ocr_word" id="word_35" title="bbox 244 348 358 387"><strong>fusión/></strong></span> <span class="ocr_word" id="word_37" title="bbox 647 348 922 387"><strong>varias/></strong></span> <span class="ocr_word" id="word_39" title="bbox 866 348 922 387"><strong>de/></strong></span> <span class="ocr_word" id="word_41" title="bbox 1141 348 1975 387"><strong>ellas/></strong></span> <span class="ocr_word" id="word_43" title="bbox 1296 348 1975 387"><strong>para/></strong></span> <span class="ocr_word" id="word_45" title="bbox 1516 348 1975 387"><strong>formar/></strong></span> </span>
15<span class="ocr_line" id="line_5" title="bbox 118 437 1975 517"><span class="ocr_word" id="word_47" title="bbox 118 437 1975 517"><strong>mayor/></strong></span> <span class="ocr_word" id="word_49" title="bbox 122 437 1975 517"><strong>concentración/></strong></span> <span class="ocr_word" id="word_51" title="bbox 677 437 922 476"><strong>velocidad/></strong></span> <span class="ocr_word" id="word_53" title="bbox 431 437 922 476"><strong>de/></strong></span> <span class="ocr_word" id="word_55" title="bbox 867 437 922 476"><strong>secado/></strong></span> <span class="ocr_word" id="word_57" title="bbox 867 437 922 476"><strong>de/></strong></span> <span class="ocr_word" id="word_59" title="bbox 1045 437 1975 476"><strong>últimas/></strong></span> <span class="ocr_word" id="word_61" title="bbox 1296 437 1975 476"><strong>de/></strong></span> <span class="ocr_word" id="word_63" title="bbox 1573 437 1975 476"><strong>para/></strong></span> <span class="ocr_word" id="word_65" title="bbox 1726 437 1975 476"><strong>igual/></strong></span> </span>
16<span class="ocr_line" id="line_6" title="bbox 118 526 1975 606"><span class="ocr_word" id="word_67" title="bbox 118 526 1975 606"><strong>de/></strong></span> <span class="ocr_word" id="word_69" title="bbox 122 526 1975 606"><strong>concentración/></strong></span> <span class="ocr_word" id="word_71" title="bbox 588 526 922 565"><strong>de/></strong></span> <span class="ocr_word" id="word_73" title="bbox 677 526 922 565"><strong>velocidad/></strong></span> <span class="ocr_word" id="word_75" title="bbox 989 526 1975 565"><strong>de/></strong></span> <span class="ocr_word" id="word_77" title="bbox 1082 526 1975 565"><strong>secado/></strong></span> <span class="ocr_word" id="word_79" title="bbox 1296 526 1975 565"><strong>de/></strong></span> <span class="ocr_word" id="word_81" title="bbox 1384 526 1975 565"><strong>de/></strong></span> <span class="ocr_word" id="word_83" title="bbox 1562 526 1975 565"><strong>de/></strong></span> <span class="ocr_word" id="word_85" title="bbox 1665 526 1975 565"><strong>de/></strong></span> <span class="ocr_word" id="word_87" title="bbox 1699 526 1975 565"><strong>de/></strong></span> <span class="ocr_word" id="word_89" title="bbox 1880 526 1975 565"><strong>de/></strong></span> <span class="ocr_word" id="word_91" title="bbox 1880 526 1975 565"><strong>de/></strong></span> <span class="ocr_word" id="word_93" title="bbox 1574 526 1975 565"><strong>de/></strong></span> <span class="ocr_word" id="word_95" title="bbox 1665 526 1975 565"><strong>de/></strong></span> </span>
17</p>
```


Resultados

- ✓ **OCR Adobe : aprox. 40% de aciertos, irregular muy sensible a la calidad del texto impreso**
- ✓ **Pre OCR + Teserract/ScanTailor: 90 % de aciertos**
- ✓ **Post OCR: Corrige el 50 % de los errores**

¿Consultas?

Ana Sanllorenti

asanllorenti@bl.fcen.uba.ar

Martín Williman

mwilliman@bl.fcen.uba.ar