



Protocolo de digitalización de periódicos

Guillermina Laitano, Nicolás Quiroga y Silvana Ferreyra

(Nodo Argentina, Proyecto PortADa)

1. Contexto y Objetivo

El presente protocolo se enmarca en el proyecto PortADa - Port Arrivals Data: Automatic data collection for a large-scale comparative history of 19th century shipping: a Digital Humanities approach to maritime heritage.

“En el siglo XIX, durante la transición de la vela al vapor, las ciudades portuarias fueron núcleos del colonialismo imperial y de las configuraciones capitalistas, vinculadas por el comercio y el tráfico marítimos a los procesos de globalización y a la división internacional del trabajo. En estos puertos, la llegada de los barcos representaba un acontecimiento económico, cultural y político. Al mismo tiempo, el abastecimiento de estos barcos y la distribución de la carga creaban redes de circulaciones internas basadas en el comercio y el tráfico local, regional y fluvial. Así, las noticias sobre la llegada de barcos procedentes de puertos lejanos y la concatenación de movimientos portuarios que ello implicaba, recibían un lugar de privilegio en la prensa local de estas ciudades portuarias” (Ibarz Gelabert, 2023: 3).

Uno de los objetivos principales del proyecto es construir una base de datos abierta que registre el arribo de buques durante la segunda mitad del siglo XIX. La base se centra en cuatro áreas geográficas que abarcan los puertos de Barcelona, Marsella, La Habana y Buenos Aires. Toma como fuente historiográfica las noticias publicadas por la prensa local de cada uno de los puertos sobre el arribo de buques y aplica una metodología computacional para todo el proceso: recopilación de fuentes, transformación de información en datos y construcción y publicación de la base de datos.

Este texto es un protocolo para el proceso de digitalización de las fuentes afectadas por el proyecto. En el caso de la prensa local del puerto de Buenos Aires, solo un 4% se encuentra digitalizado. Propuesto inicialmente teniendo en cuenta las fuentes relacionadas con la ciudad puerto mencionada, las indicaciones aquí sugeridas son extensivas a las tareas de capturas de imágenes del proyecto.

Incluso en los casos en que la mayor parte de las fuentes se encuentran digitalizadas, el protocolo puede utilizarse como guía para la obtención de nuevas imágenes cuando las ya obtenidas presentan problemas de calidad o fallas.

2. Definición de digitalización de documentos históricos y particularidades del presente protocolo

La finalidad principal de la digitalización es la copia en formato digital de documentos analógicos.

El formato digital permite no solo el acceso remoto, sino también simultáneo a la documentación histórica. Por lo tanto, amplía la difusión de los archivos, incrementa la accesibilidad a los mismos y habilita las colaboraciones a escala global entre investigadores.

La digitalización de los documentos históricos tiene como una de sus ventajas la rapidez a la hora de recuperar la información de interés y la agilidad para realizar las consultas. También contribuye a la preservación de los documentos originales, al minimizar su manipulación.

En el marco del proyecto PortADa, lo relevante en el proceso de digitalización consiste en obtener una imagen de una calidad tal que garantice el mayor porcentaje de éxito en la aplicación de software de reconocimiento óptico de caracteres (OCR). Se trata de algoritmos que reconocen el texto presente en una imagen y lo convierten en caracteres.

En este sentido, la digitalización debe garantizar:

- que la toma fotográfica no registre movimiento alguno
- que las tomas se realicen en un espacio con adecuada iluminación
- que se minimice lo más posible en la imagen la curvatura que genera la encuadernación
- que la imagen presente un contraste adecuado
- que la información de interés quede comprendida en el encuadre de la fotografía
- que el encuadre de la fotografía incluya márgenes (es decir, que entre las palabras de las noticias y el encuadre de la fotografía haya un espacio)
- que las líneas de texto queden horizontales en la toma (sin inclinación)

3. Descripción de los documentos a digitalizar

Los periódicos (y sus periodos) que abarcará el proyecto para el puerto de Buenos Aires son: *La Gaceta Mercantil* (enero 1850 - enero 1852); *The British Packet and Argentine News* (febrero 1852 - abril 1852); *El Nacional* (mayo 1852 - diciembre 1869); y *La Prensa* (enero 1870 - diciembre 1900).

Las ediciones de *La Gaceta Mercantil* y *The British Packet and Argentine News* se encuentran digitalizadas en la Biblioteca Nacional Mariano Moreno de la República Argentina.

Por tanto, los periódicos y periodos a digitalizar a partir del presente protocolo son: *El Nacional* (periodo mayo 1852 - diciembre 1869) y *La Prensa* (enero 1870 - diciembre 1900).

Características de los periódicos a digitalizar

El Nacional presenta sus ediciones en formato sábana (68 cm.). Se encuentra archivado en formato de encuadernación en la Biblioteca Nacional Mariano Moreno de la República Argentina (Ciudad de Buenos Aires). Dispone de la fecha de edición solo en su portada, y sus páginas no están numeradas. La información se presenta en columnas.

La Prensa varía el tamaño de sus ediciones con los años. Se encuentra archivado en formato de encuadernación en la Biblioteca Tornquist del Banco Central de la República Argentina (Ciudad de Buenos Aires). Presenta en cada una de sus páginas la fecha de edición y la numeración. La información se presenta en columnas.

4. Procedimiento para la digitalización

Fragmentos a ser digitalizados

En atención a las características arriba señaladas, se deberá obtener: una fotografía de la portada y las fotografías de la noticia en sí.

La captura de la noticia de arribos en sí puede implicar más de una fotografía. Esta situación puede derivar o bien de la extensión de la noticia (una extensión que no quepa en el encuadre de una toma) o bien de la distribución de la noticia en más de una columna. Ambos casos requerirán más de una toma por noticia.

Las tomas fotográficas deben realizarse **siempre** respetando el orden cronológico y el orden de edición. El orden de las unidades digitales (fotografías) es central para reconstruir adecuadamente la unidad de información. Se debe comenzar en el día 1 del mes enero y así sucesivamente, y se debe respetar el orden de portada y noticia.

Se ha confeccionado una grilla de control para completar con datos relevantes (cantidad de fotos, páginas) durante la toma de fotografías. La grilla de control y las instrucciones para completarlo se detallan a continuación.

Grilla de control

La/s persona/s a cargo de la digitalización completarán una grilla de control por jornada de trabajo.

En la fila inicial deberá indicar su nombre, la fecha de la jornada de trabajo y la denominación de la primera fotografía que toma.

La grilla de control tendrá impreso el nombre de la prensa y la fecha del ejemplar.

Durante la captura se indicará la cantidad de fotos que se toman por cada ejemplar del periódico, considerando tanto la foto de la portada como la o las fotos vinculadas a la/las noticias. En la otra columna se consignará el/los números de la página donde figura la noticia y la cantidad de fotos por página, respetando el siguiente formato:

- número de página (cantidad), número de página (cantidad). Ejemplo: 5 (2), 6(1)

Si el ejemplar no se encuentra en el archivo se completará con una 'x'.

Nombre: Jornada de trabajo: Denominación de la primera foto de la jornada de trabajo en el dispositivo: (ej. IMG_82837)				
Periódico	Fecha	Cantidad de fotos tomadas	Número de página (cantidad de fotos)	Observaciones
El Nacional	01/05/1852	5	5(1), 6(3)	
El Nacional	02/05/1852	2	2(1)	
El Nacional	03/05/1852	5	8(4)	
El Nacional	04/05/1852	x	x	

El campo observaciones se utilizará para informar las incidencias.

Se consignarán allí datos del contexto del proceso de digitalización. Si hay dos unidades de información en un mismo ejemplar (arribos y manifiestos) se consigna en ese campo.

También deben registrarse por ejemplo si se localizan hojas muy dañadas (rotas, con pliegues).

Finalmente, deben registrarse las decisiones que se han ido tomando ante la emergencia de situaciones no estipuladas en el presente protocolo, así como las decisiones tomadas ante contratiempos emergentes.

Memoria del dispositivo

Debe tenerse en cuenta la cantidad de memoria que requerirán las tomas de una jornada de trabajo. Por ejemplo, una tarjeta de memoria de 128 GB guarda 1600 fotos en formato RAW para una cámara de 24 Megapíxeles.

Se prevé que se tomarán alrededor de 18000 fotografías, lo que podría guardarse en 1,4 TB.

5. Dispositivo de captura

Hasta no hace muchos años, la tecnología de cámara digital no se consideraba suficientemente desarrollada como para aplicarse a los trabajos de digitalización de patrimonio bibliográfico o documental. En este sentido, el dispositivo preferido era siempre el escáner. Sin embargo, hace ya años que esta tecnología ofrece prestaciones técnicas suficientes para encarar este tipo de trabajos. Es más, ahora mismo la mayor calidad de digitalización la ofrecen las cámaras fotográficas (Consejo de Cooperación Bibliotecaria, 2021: 151).

Por su carácter flexible y versátil en este proyecto optamos por la digitalización con cámara fotográfica o celular.

Por las condiciones en que se realizará la digitalización ambos dispositivos son los más adecuados, ya que son fácilmente transportables y no requieren solicitudes de permisos especiales para su uso en las bibliotecas seleccionadas. Asimismo, garantizan la mejor manipulación y el menor desgaste de la prensa.

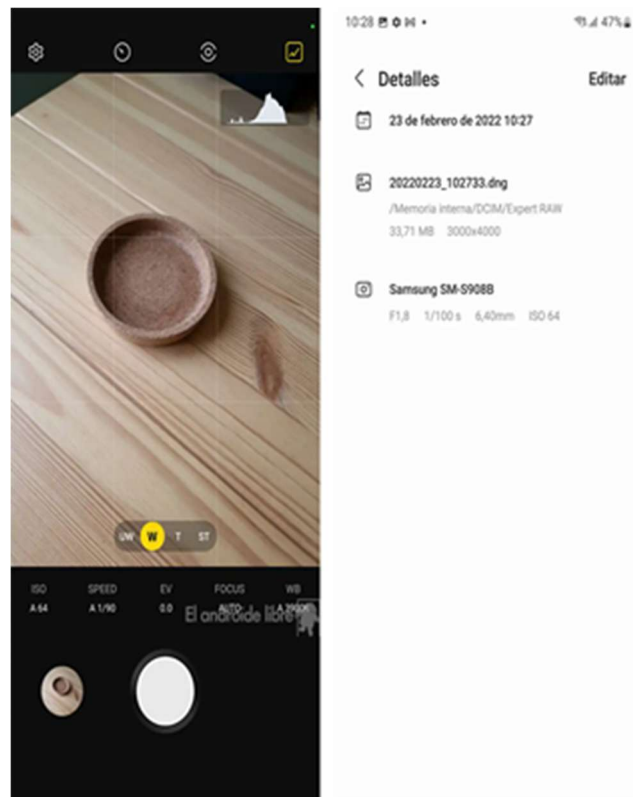
No obstante, siguiendo nuevamente el manual de buenas prácticas para la digitalización del Ministerio de Cultura de España, conviene tener en cuenta las principales desventajas del uso de una cámara o celular:

- Es necesaria mayor experticia en digitalización y fotografía digital para manejar estos equipos.
- Requiere mayor preparación: hay que montar el puesto, controlar la luz, etc.
- La captura es más lenta.
- Es más difícil mantener homogéneas las condiciones de digitalización (ajustes automáticos de la cámara, movilidad de los elementos que conforman el puesto...).
- Se pueden presentar problemas de distorsión geométrica. (Consejo de Cooperación Bibliotecaria, 2021: 153)

Es condición indispensable que la cámara o el celular posean formato RAW / DNG (negativo digital generado por el dispositivo de captura). Ese será el formato de las capturas.

Un archivo RAW también se conoce como «negativo digital» porque los datos se pueden procesar e imprimir de diferentes maneras para producir diferentes resultados, como el negativo de una cámara de cine. Al igual que un negativo, el archivo RAW nunca cambia. Cuando abres un archivo RAW en una aplicación de software, puedes procesarlo, editarlo y luego guardarlo para crear un nuevo archivo en tu ordenador (por lo general, puedes elegir entre JPEG o TIFF). El archivo RAW original no presenta ningún cambio y se puede volver a abrir y utilizar en cualquier momento para conseguir un resultado totalmente diferente (Nicholson, s.f.).

Por ejemplo, si utilizamos un celular Samsung, es posible descargar la aplicación Galaxy Expert Raw. Para acceder se abre la app de cámara del celular, se pulsa en la opción Más, se busca el modo Expert Raw y se abre su interfaz con un clic.



Fuente: extraído de Ramírez (2022, 23 de febrero).

Proponemos alcanzar capturas cercanas a los 24 bits de profundidad en los formatos indicados. Asimismo, es recomendable que la resolución de la cámara no sea inferior a los 12 Megapíxeles. Con dispositivos de última generación ese objetivo está garantizado. Si bien existen diferentes guías de digitalización en cuanto a los requerimientos de la toma y las especificidades de la imagen digital obtenida (que utilizan diferentes medidas de difícil conversión), seguimos aquí lo planteado por un proyecto archivístico de conservación digital e investigación histórica integrado por archivistas e historiadores con fuentes históricas de siglo XIX y XX (Ceva, Araujo y Torres, 2022).

Teniendo en cuenta que el aumento de resolución implica el aumento del peso de las imágenes digitales, hemos de tratar de priorizar en la ecuación la optimización de la resolución de las imágenes.

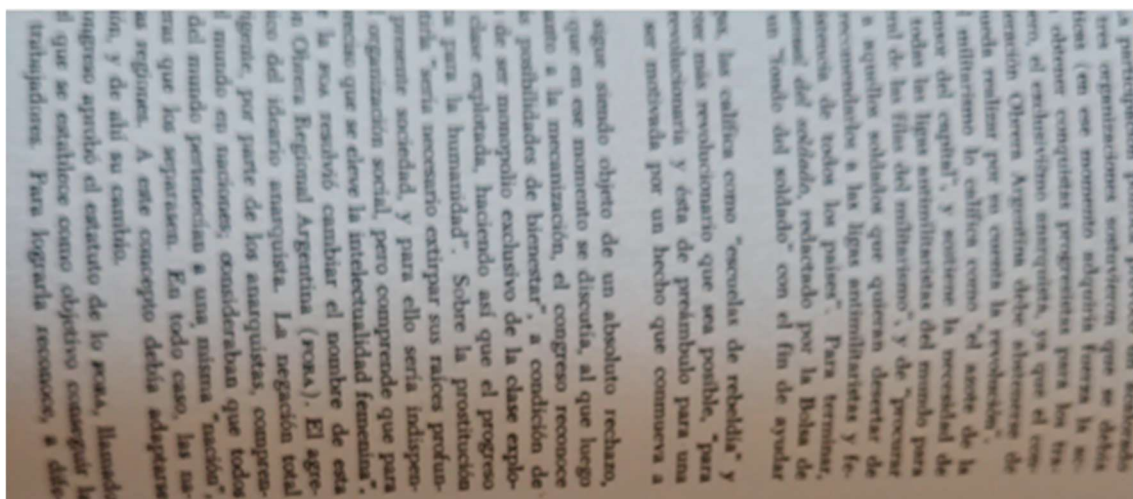


Foto tomada con cámara de 12 Megapíxeles en JPEG (3,91 MB)

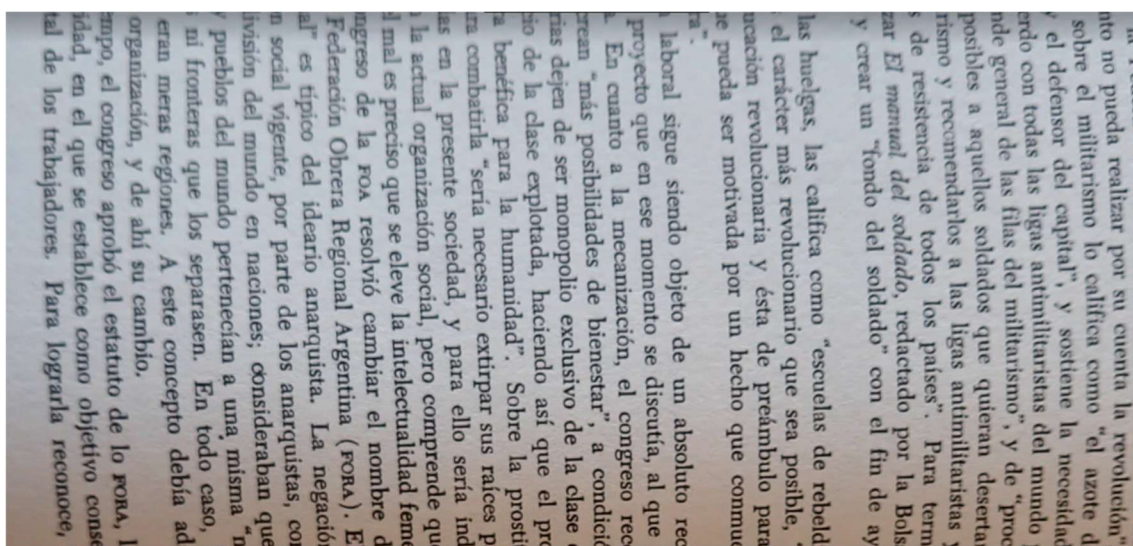


Foto tomada con cámara de 12 Megapíxeles en RAW (32 MB)

6. Recomendaciones para la captura

Uniformidad en la iluminación del documento

La luz debe mantenerse homogénea durante los procesos de captura; para ello es recomendable evitar las variaciones de la luz natural y usar luz ambiente artificial.

Es importante evitar las sombras sobre el documento. Para este fin no se podrá utilizar flash, tanto por cuestiones de conservación como de homogeneidad de iluminación.

Podemos eliminar las sombras controlando la iluminación que incide sobre el documento durante la captura (iluminando desde arriba)

Si se están digitalizando diarios encuadernados, la iluminación lateral puede producir sombras en la zona del lomo, para estos casos es preferible iluminar desde arriba o alinear la hendidura de la costura de encuadernación con las luces.

Si se utilizan varias lámparas habrá que colocarlas normalmente equidistantes al documento.

Una herramienta que parece obtener buenos resultados atentos a una ecuación precio- calidad es el aro de luz led. Si se realizan las tomas con celular, algunos vienen con soporte incorporado.

Paralelismo entre superficie del documento y plano focal

Tiene que existir un paralelismo completo entre la superficie del documento y el plano focal de la cámara antes de la captura. En su defecto, se producirán distorsiones de tipo geométrico que podrán alterar las formas y relaciones dimensionales de los contenidos. Este problema es especialmente grave cuando se va a realizar OCR de las imágenes.

Cuando se trate de dispositivos aéreos (cámara o celular con trípode), una opción es comprobar este parámetro mediante un nivel de burbuja, que se colocará sobre la cámara digital y sobre la superficie de los originales.

También puede ser útil una captura de una cuadrícula colocada sobre el área de captura. Las líneas tienen que quedar perpendiculares o paralelas, según corresponda, en la imagen generada. Habitualmente, la app de la cámara tiene incorporada una opción para visualizar la grilla en la pantalla mientras se realiza el disparo.

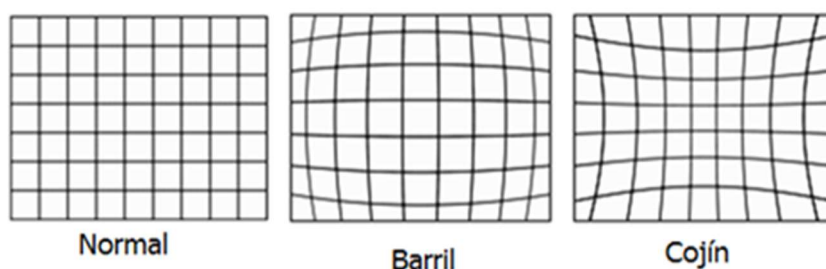


Figura 36. Distorsiones geométricas típicas

Fuente: Extraído de Consejo de Cooperación Bibliotecaria (2021: 122).

Se pueden utilizar las funciones de autoenfoco de las cámaras comprobando previamente que funcionan bien.

Montaje

Si bien hemos decidido no tomar fotografías de página completa para la prensa de Buenos Aires, lo que nos permite descartar este paso, dejamos una serie de consideraciones que pueden ser útiles para la digitalización de las fuentes en otros puertos que se incorporen a futuro en el proyecto.

Para asegurarse que el puesto de digitalización esté bien montado, y la cámara, firmemente sujeta, exploramos distintas alternativas.

Por un lado, la posibilidad de un trípode aéreo. Hay posibilidades muy interesantes, aunque un poco más caras, que brindan también iluminación.

Por otro, la posibilidad de una carpa de escaneo. En Argentina podría construirse un dispositivo similar a partir de la compra de una carpa fotográfica caja de luz.

Es recomendable utilizar tecnologías que nos permitan disparar la cámara de modo remoto (cables, dispositivos inalámbricos, por ordenador, etc.). Es conveniente utilizar cables lo más cortos que sea posible.

7. Manejo y preservación del material original

Las formas de manipulación del material original deberán respetar los criterios y protocolos del archivo donde se encuentre la documentación histórica.

Debe realizarse un manejo cuidadoso del material para evitar cualquier daño, rotura o deterioro.

Si el papel se encuentra en un estado de conservación frágil o quebradizo, se debe extremar el cuidado en su manipulación.

Al finalizar una jornada de trabajo, la documentación debe devolverse en el mismo orden en que fue entregada.

8. Control de calidad

Los controles de calidad en este proyecto forman parte de la etapa de procesamiento de las imágenes digitales obtenidas. Para la evaluación y correcciones (nuevas capturas) se procederá a analizar la calidad de las fotografías durante la etapa de “positivado” de los archivos RAW y por otro lado la verificación de la correspondencia entre la secuencia de las capturas y las grillas de control.

9. Unidades digitales. Formato y resguardo

Formato de archivo

En atención a lo establecido por la bibliografía especializada, se producirán dos tipos de copias:

- Un archivo Máster (copia de conservación). Reproducción digital de alta calidad en formato sin compresión o con compresión sin pérdida. Es el formato en que se tomará la fotografía. Los formatos óptimos para el archivo Máster son RAW o TIFF sin compresión. En este protocolo se utilizará el formato RAW.

- Un archivo Derivado (copia de consulta). Archivo creado a partir del archivo Máster, luego del procesamiento de las capturas, que sea apto para alojarse en un servidor y para su consulta (.JPEG o .PDF).

Protocolo de resguardo

Los dispositivos de almacenamiento de las unidades digitales son diversos en sus soportes y todos contienen ventajas y desventajas. Por lo tanto es recomendable la utilización de más de uno al momento de crear copias de seguridad. Además,

mientras más dispositivos se utilicen para almacenar las unidades digitales, menos riesgos se correrán en cuanto a pérdida de material.

Seguiremos la recomendación canónica de un tipo de backup: 3-2-1, es decir: al menos tres copias, al menos en dos dispositivos diferentes (por ejemplo, disco NAS, nube, cinta magnética), uno al menos fuera del ámbito de trabajo.

10. Recaptura de fotografías

Ante errores en una digitalización, que no puedan subsanarse durante el proceso de edición de la imagen, se deberá volver a realizar la fotografía.

11. Bibliografía consultada

- Biblioteca Nacional Mariano Moreno, Política de digitalización de la Biblioteca Nacional Mariano Moreno, CABA, 2020. [<https://www.bn.gov.ar/resources/bibliotecarios/politica-digit-bnmm.pdf>]
- Ceva, Mariela, Araujo, Facundo y Eugenio Torres (2022). The importance of building up a business archive: the experience in Bunge & Born Argentina. En Curra, Edoardo, Docci Marina, Menichelli Claudio, Russo Martina y Laura Severi *Stati Generali del Patrimonio Industriale 2022*. Venecia: Marsilio Editori, pp. 1988-2004
[https://www.academia.edu/90394810/The_importance_of_building_up_a_business_archive_the_experience_in_Bunge_and_Born_Argentina]
- Consejo de Cooperación Bibliotecaria Grupo de Trabajo de Buenas Prácticas en Digitalización (2021). Recomendaciones para proyectos de digitalización de patrimonio bibliográfico y fotografía histórica. Ministerio de Cultura y Deporte, España [<https://hdl.handle.net/10421/9200>]
- Ibarz Gelabert, Jordi (2023). Proyecto PortADa - Port Arrivals Data: Automatic data collection for a large-scale comparative history of 19th century shipping: a Digital Humanities approach to maritime heritage. [https://www.proyectoportada.eu/wp-content/uploads/2025/05/PortADa_PCDE_final.pdf]
- Nicholson, Angela (s.f.). Formatos (RAW, JPEG, HEIF) y compresión de imagen. Canon. [<https://www.canon.es/pro/infobank/image-file-types/>]
- Ramírez, Manuel (2022, 23 de febrero). Así es Samsung Expert RAW, la app exclusiva de dos teléfonos Galaxy que llegará a otros en días. El Español. [https://www.elspanol.com/elandroidelibre/aplicaciones/20220223/samsung-expert-raw-app-exclusiva-telefonos-galaxy/652434822_0.html]
- Sistema Nacional de Documentación Histórica (Agosto, 2019). Guía General de Digitalización de Documentos. Versión 1.0. [https://www.argentina.gob.ar/sites/default/files/guia_general_de_digitalizacion_de_documentos_vf.pdf]