UNIVERSIDAD NACIONAL DE COLOMBIA

PROTOCOLO DE ANONIMIZACIÓN (SANITIZACIÓN) DE DATOS EN CONTENIDO DE LOS DOCUMENTOS TEXTUALES PARA LA UNIVERSIDAD NACIONAL DE COLOMBIA

Tabla de Contenido

1. Información general del documento	
Objetivo	Definir el protocolo de anonimización (sanitización) de datos en contenido de los documentos textuales, teniendo en cuenta el marco normativo y técnico con el fin de facilitar la divulgación de la información, protegiendo la privacidad de los individuos y cumpliendo con la normativa vigente en materia de protección de datos personales.
Alcance	El alcance de este documento está orientado a analizar el proceso de anonimización de información (Sanitización) de documentos textuales y definir un protocolo funcional y sencillo para que pueda ser aplicado por las oficinas productoras.
Definiciones	**Anonimización**: proceso por el cual la información de identificación personal se modifica de forma irreversible de tal manera que no se pueda identificar, directa o indirectamente, ya sea por sus propios medios o en colaboración con algún tercero, a la persona asociada a dicha información de identificación personal. (Fuente: Estándar ISO / IEC 29100:2011, 2011). **Confidencialidad:** se conoce como una forma de prevenir la divulgación de la información a personas o sistemas que no se encuentran autorizados. (Fuente: Guía de Anonimización de Datos Estructurados - Archivo General de la Nación). **Dato personal**: cualquier información vinculada o que pueda asociarse a una o varias personas naturales determinadas o determinables. Por datos personales se entiende cualquier información relativa a una persona identificada o identificable. Esta persona también se conoce como "sujeto de los datos". Una persona identificable es aquella que puede ser identificada, directa o indirectamente, en particular por referencia a un identificador como un nombre, un número de identificación, datos de localización, un identificador en línea o a uno o más factores específicos de la identidad física, fisiológica, genética, mental, económica, cultural o social de esa persona. (Fuente: Superintendencia de Industria y Comercio). **Dato público**: es el dato que no sea semiprivado, privado o sensible. Son considerados datos públicos, entre otros, los datos relativos al estado civil de las personas, a su profesión u oficio y a su calidad de comerciante o de servidor público. Por su naturaleza, los datos públicos pueden estar contenidos, entre otros, en registros públicos, documentos públicos, gacetas y boletines oficiales y sentencias judiciales debidamente ejecutoriadas que no estén sometidas a reserva. (Fuente: Decreto 1377, 2013. Artículo 3, numeral 2). **Datos sensibles**: se entiende por datos sensibles aquellos que afectan la intimidad del titular o cuyo uso indebido puede generar su discriminación, tales como aquellos que revelen el origen racial o étnico, la orientación política, las convicciones religiosas o filosóficas, la pertenencia a sindicatos, organizaciones sociales, de derechos humanos

	o que promueva intereses de cualquier partido político o que garanticen los derechos y garantías de partidos políticos de oposición, así como los datos relativos a la salud, a la vida sexual, y los datos biométricos. (Fuente: Ley 1581, 2012. Artículo 5, Título III). **Dato estructurado**: están organizados conforme a un modelo o esquema. Se almacenan en forma tabular y algunas veces su estructura también incluye la definición de las relaciones entre ellos. Típicamente, están representados en bases de datos que hacen parte del funcionamiento de sistemas de información. (Fuente: CONPES 3920). **Dato no estructurado**: su organización y presentación no está guiada por ningún modelo o esquema. En esta categoría se incluyen, por ejemplo, las imágenes, texto, audios, contenidos de redes sociales, videos. (Fuente: CONPES 3920 de 2018). **Dato semiestructurado**: su organización y presentación tiene una estructura básica (etiquetas o marcadores), pero no tiene establecida una definición de relaciones en su contenido. En esta categoría se incluyen contenidos de emails, tweets, archivos XML. (Fuente: CONPES 3920 de 2018). **Disponibilidad**: es un pilar fundamental de la seguridad de la información, que debe estar alineada con la integridad, garantizando que los usuarios de los sistemas cuenten con la información en el momento de realizar una consulta. Para cumplir con la última condición se debe tener claro cuál será el flujo de datos que debemos manejar, para conocer donde se debe almacenar dicha información, qué tipo de servicio debemos contratar, etc. (Fuente: Guía de Anonimización de Datos Estructurados - Archivo General de la Nación). **Información:** se refiere a un conjunto organizado de datos contenido en cualquier documento que los sujetos obligados generen, obtengan, adquieran, transformen o controlen. (Fuente: Ley 1712, 2014. Artículo 6). **Información pública:** es toda información que un sujeto obligado genere, obtenga, adquiera, o controle en su calidad de tal. (Fuente: Ley 1712, 2014. Artículo 6). **Información pública clasificada:** es aquella información que, estando en poder o custodia de un sujeto obligado en su calidad de tal, pertenece al ámbito propio, particular y privado o semiprivado de una persona natural o jurídica por lo que su acceso podrá ser negado o exceptuado, siempre que se trate de las circunstancias legítimas y necesarias y los derechos particulares o privados consagrados en el artículo 18 de la ley 1712 de 2014. (Fuente: Ley 1712, 2014. Artículo 6).

	Información pública reservada: es aquella información que estando en poder o custodia de un sujeto obligado en su calidad de tal, es exceptuada de acceso a la ciudadanía por daño a intereses públicos y bajo cumplimiento de la totalidad de los requisitos de ley. (Fuente: Ley 1712, 2014. Artículo 6). **Integridad:** cuando hablamos de integridad en seguridad de la información nos referimos a cómo los datos se mantienen intactos libre de modificaciones o alteraciones por terceros, cuando una violación modifica algo en la base de datos, sea por accidente o intencionado, se pierde la integridad y falla el proceso. Por lo cual se debe proteger la información para que solo sea modificada por la misma persona, evitando así que se pierda la integridad. Una manera de proteger los datos es cifrando la información mediante un método de autenticidad como una contraseña o mediante huella digital. (Fuente: Guía de Anonimización de Datos Estructurados - Archivo General de la Nación). **Identificadores directos:** son todas aquellas características que por sí mismas permiten la identificación de una persona o entidad de manera inequívoca dentro de un conjunto de datos. (Fuente: Guía para la anonimización de datos e información no estructurada: estándares y lineamientos técnicos. Centro Nacional de Memoria Histórica). **Identificadores indirectos o cuasi - identificadores:** son aquellas características que por sí solas no permiten la identificación de una persona o entidad, pero que relacionados o en combinación con otros identificadores indirectos podrían permitir la identificación dentro de un conjunto de datos. (Fuente: Guía para la anonimización de datos e información no estructurada: estándares y lineamientos técnicos. Centro Nacional de Memoria Histórica). **Seudonimización:** se entiende la sustitución de cualquier característica identificativa de los datos por un seudónimo o, lo que es lo mismo, un valor que no permita identificar directamente al interesado. Se establece la seudonimización como el tratamiento de datos personales de tal manera que los datos personales ya no pueden atribuirse a un sujeto de datos específico sin el uso de información adicional, siempre que (a) dicha información adicional se mantenga por separado, y (b) esté sujeta a medidas técnicas y organizativas para garantizar que los datos personales no se atribuyan a una persona identificada o identificable. (Fuente: Guía de Anonimización de Datos Estructurados - Archivo General de la Nación).

Documentos de Referencia (Opcional)	**Constitución Política de Colombia 1991**: Artículo 15. *"Todas las personas tienen derecho a su intimidad personal y familiar y a su buen nombre, y el Estado debe respetarlos y hacerlos respetar"*. De igual modo, tienen derecho a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bancos de datos y en archivos de entidades públicas y privadas. **Ley 1266 de 2008**. Por la cual se dictan las disposiciones generales del hábeas data y se regula el manejo de la información contenida en bases de datos personales, en especial la financiera, crediticia, comercial, de servicios y la proveniente de terceros países y se dictan otras disposiciones. **Ley estatutaria 1581 de 2012**. Reglamentada parcialmente por el Decreto Nacional 1377 de 2013, reglamentada parcialmente por el Decreto 1081 de 2015. Ver sentencia C-748 de 2011. Ver Decreto 255 de 2022. Por la cual se dictan disposiciones generales para la protección de datos personales. **Ley 1712 de 2014**. Por medio de la cual se crea la Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional y se dictan otras disposiciones. **Ley 1755 de 2015**. Por medio de la cual se regula el derecho fundamental de petición y se sustituye un título del Código de Procedimiento Administrativo y de lo Contencioso Administrativo. **Ley 1273 de 2009.** Por medio de la cual se modifica el Código Penal, se crea un nuevo bien jurídico tutelado - denominado *"de la protección de la información y de los datos"* y se preservan integralmente los sistemas que utilicen las tecnologías de la información y las comunicaciones, entre otras disposiciones. **Ley 594 de 2000.** Por medio de la cual se dicta la Ley General de Archivos y se dictan otras disposiciones **Decreto 1377 de 2013**. Por el cual se reglamenta parcialmente la Ley 1581 de 2012, Derogado Parcialmente por el Decreto 1081 de 2015.
Condiciones Generales	Protocolo de anonimización de datos en contenido de los documentos de la UNIVERSIDAD NACIONAL DE COLOMBIA, puede presentar modificaciones de conformidad con la actualización de las normas en la materia.

2. Introducción

La Ley General de Protección de Datos Personales (Ley 1581 de 2012), establece una serie de principios para regular el manejo de la información personal y asegurar el derecho fundamental de hábeas data de las personas, uno de estos principios, es conocido como circulación restringida, el cual indica que, excepto que la información sea de dominio público, los datos personales no deben estar disponibles en internet u otros medios de comunicación masiva, a menos que se pueda controlar técnicamente el acceso para limitar la divulgación. Esta regulación requiere la implementación de medidas técnicas, humanas y administrativas para salvaguardar la información sensible contra la adulteración, pérdida, consulta, uso o acceso no autorizado.

En la Universidad Nacional de Colombia, los documentos electrónicos de archivo pueden contener información personal sensible, cuya exposición podría comprometer la privacidad de los individuos. El uso indebido de estos datos puede resultar en discriminación y afectación a la intimidad del titular de la información, toda vez que, al revelar información como el origen racial o étnico, orientación política, convicciones religiosas o filosóficas, pertenencia a organizaciones sociales de derechos humanos o partidos políticos, datos relativos a la salud, la vida sexual, datos biométricos entre otros, se pone en riesgo el cumplimiento constitucional y normativo de los derechos fundamentales de los individuos.

Para garantizar el derecho de acceso a la información pública y las excepciones a la divulgación de información, sin contravenir la regulación nacional en materia de protección de datos personales, es necesario utilizar las técnicas adecuadas para salvaguardar la información sensible contenida en los documentos antes de su publicación y de este modo, proteger los derechos fundamentales de sus titulares y evitar posibles consecuencias legales.

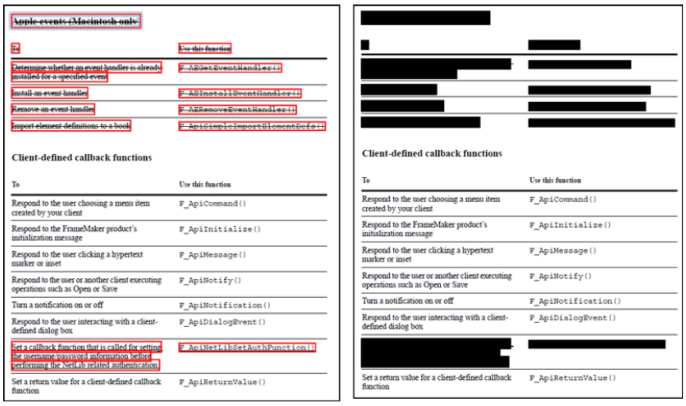
La anonimización de datos se puede llevar a cabo en documentos físicos, en formatos audiovisuales o de multimedia; sin embargo el presente documento está orientado al proceso de desinfección de información (Sanitización) que debe utilizarse en la divulgación parcial de los contenidos de los documentos electrónicos de archivo textuales que no estén protegidos por las excepciones de la Ley 1712 de 2014, y presenta en su primera parte las generalidades del proceso de anonimización, seguido de las técnicas útiles para el proceso y, finalmente el documento expondrá los pasos para llevar a cabo el proceso.

3. Generalidades

En aras de asegurar el derecho de acceso a la información pública, las garantías y las excepciones a la publicidad de información sin afectar lo dispuesto en la regulación nacional en materia de protección de datos personales; muchas veces se requiere eliminar el contenido de un documento antes de su publicación debido a las posibles consecuencias legales. En tal sentido la anonimización de datos es un proceso esencial para proteger la identidad de las personas u organizaciones, eliminando cualquier posibilidad de identificación; este proceso se realiza para facilitar la divulgación, la publicación y el intercambio de datos, sin vulnerar los derechos a la protección de datos, de manera que no se puedan identificar directa o indirectamente las personas asociadas a dicha información de identificación personal.

Cuando los documentos estaban en formato de papel, la anonimización consistía en tachar términos, frases, secciones o bloques enteros de texto. Otros contextos en los que se podría aplicar la anonimización de

documentos son, por ejemplo, los documentos de resumen legal de un caso judicial que deben ponerse a disposición de la prensa, o los documentos que contienen datos médicos o personales.



Client-defined callback functions

To	Use this function
Respond to the user choosing a menu item created by your client	F_ApiCommand()
Respond to the FrameMaker product's initialization message	F_ApiInitialize()
Respond to the user clicking a hypertext marker or inset	F_ApiMessage()
Respond to the user or another client executing operations such as Open or Save	F_ApiNotify()
Turn a notification on or off	F_ApiNotification()
Respond to the user interacting with a client-defined dialog box	F_ApiDialogEvent()
Set a return value for a client-defined callback function	F_ApiReturnValue()

En conclusión, la "*anonimización de datos*" es un proceso que consiste en enmascarar, modificar, ocultar o eliminar datos privados o confidenciales de un documento a la vez que se conserva su formato original, con el fin de proteger datos confidenciales en su lectura.

3.1. ¿Para qué anonimizar datos?

1. Para proteger los derechos de los titulares de los datos e información y reducir o eliminar definitivamente el riesgo de reidentificación.

2. Para evitar no solo la identificación directa, sino también la identificación indirecta: esta se obtiene del cruce de datos y otras fuentes de información.

3. Para permitir el uso, la divulgación, el intercambio y la publicación de la información de manera adecuada y para los fines establecidos, garantizando los derechos a la intimidad y privacidad de las personas.

Ejemplo de la información confidencial que un documento podría incluir:

- Código estudiantil
- Nombres de menores de edad
- Fechas de nacimiento
- Dirección de domicilio
- Números de teléfono
- Orientación política
- Origen racial o étnico
- Notas académicas

4. Técnicas para el proceso de anonimización datos en documentos textuales de la UNAL

En la Universidad Nacional de Colombia la divulgación parcial de los contenidos de los documentos electrónicos de archivo que no estén protegidos por las excepciones de la Ley 1712 de 2014, puede hacerse mediante la producción de una versión pública proveniente y certificada por la misma entidad productora teniendo en cuenta las técnicas de anonimización o sanitización pertinentes para tal fin.

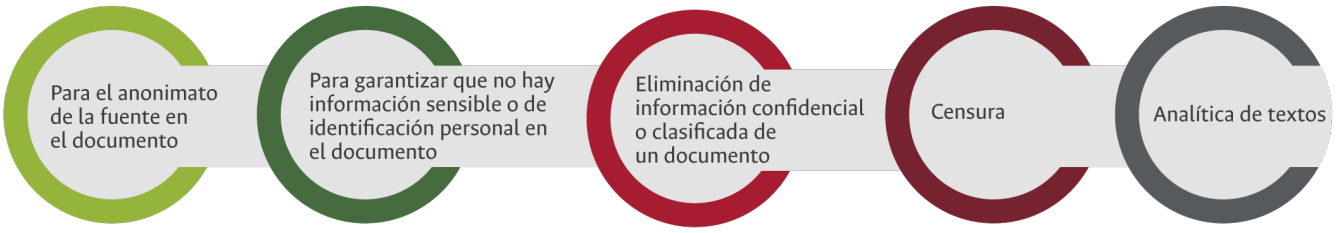
Es importante mencionar que la autenticidad, integridad, e inalterabilidad de los documentos originales y que son considerados documentos de archivo NO podrán verse afectados por la aplicación de estas técnicas, por lo tanto, es necesario que estos procesos se realicen sobre una copia del documento original.

4.1. Limpieza de datos en documentos electrónicos de archivo:

Es el proceso de eliminar o modificar información sensible en documentos o bases de datos para evitar la exposición de datos personales o confidenciales, de tal manera que se pueda dejar a disposición de un público más amplio manteniendo su utilidad y funcionalidad sin revelar la información sensible allí contenida, asegurando que los datos sensibles no puedan ser recuperados o identificados, protegiendo así la privacidad y seguridad de la información. Un aspecto clave del proceso de sanitización de documentos es que el documento saneado debe seguir manteniendo la utilidad de la información sin revelar la información sensible.

Una forma de sanitización podría ser borrar la información de los documentos por completo o mostrar la información enmascarada a los usuarios no autorizados.

4.1.1. Propósitos de la desinfección (Sanitización)



Las técnicas de desinfección por lo general utilizan un proceso encomendado que se ocupa de editar contenidos o camuflar la información confidencial del documento original, esta técnica es conocida como redacción de datos. Este método hace uso de técnicas inteligentes que hacen que la audiencia en general desconozca la información, pero con un examen cuidadoso se podría descubrir la pieza secreta. Dentro de la desinfección (Sanitización) se deriva el término *"redacción"* el cual se explica a continuación:

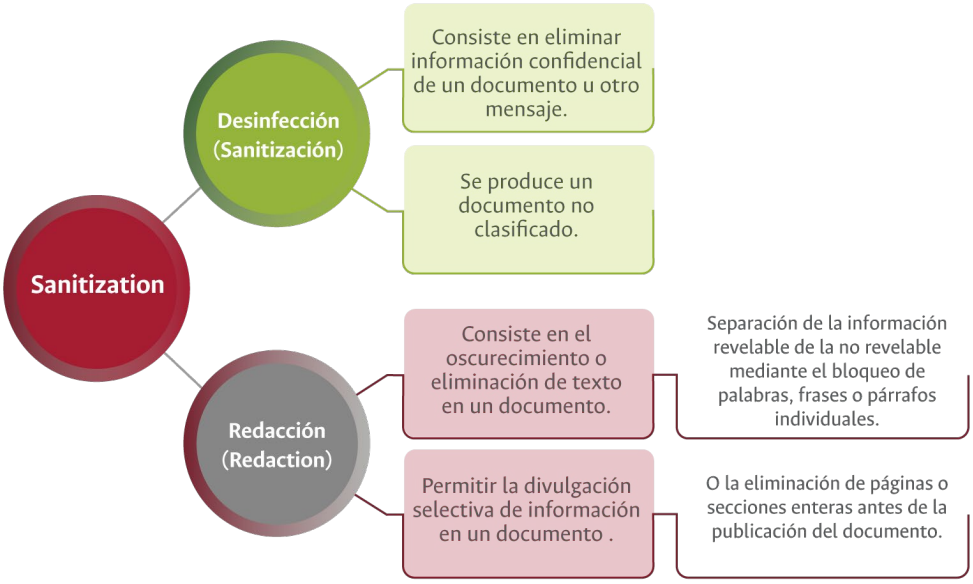
4.1.2. Redacción

La redacción consiste en el oscurecimiento o eliminación de texto en un documento, el bloqueo de palabras, frases o párrafos individuales o la eliminación de páginas o secciones enteras antes de la publicación del documento. Su objetivo es permitir la divulgación selectiva de información en un documento mientras se mantienen en secreto otras partes del documento. Por lo general, el resultado es un documento que es adecuado para su publicación o para su difusión a otras personas que no sean la audiencia prevista del documento original.

La redacción es la eliminación permanente de datos confidenciales, el equivalente digital de "tapar" texto en material impreso. La redacción se puede lograr simplemente eliminando caracteres de un archivo o registro de base de datos, o reemplazando caracteres con asteriscos u otros marcadores de posición.

La redacción automatizada de datos es un método efectivo para eliminar datos confidenciales de documentos, hojas de cálculo y otros archivos, sin alterar el contenido restante del archivo.

Ejemplo de redacción de datos:



En un conjunto de datos una persona aparece como "Hombre, 28 años, código postal 01522". Usando la redacción de datos, el punto de datos puede verse así:

"Hombre, edad —, código postal ——"

4.1.3. Enmascaramiento

El enmascaramiento es una técnica de anonimización que implica la alteración de datos sensibles de manera que se preserve la estructura de los datos, pero se asegure que la información original no pueda ser identificada ni recuperada. Esta técnica no solo elimina los datos originales, sino que los reemplaza con valores generados aleatoriamente o creados utilizando un conjunto de parámetros específicos. Estos parámetros están diseñados para garantizar el anonimato y mantener la integridad estructural de los datos. El enmascaramiento puede también generalizar los datos hasta un punto en que no puedan ser utilizados para identificar a una persona específica.

Ejemplo de enmascaramiento de datos:

En un conjunto de datos una persona aparece como "Hombre, 28 años, código postal 01522". Usando el enmascaramiento de datos, el punto de datos puede verse así:

"Hombre, edad 20-30, región Andina"

5. Pasos para aplicar la Sanitización en un documento electrónico

La desinfección (Sanitización) de documentos consiste en dos actividades principales:

Actividad 1
Detección de datos sensibles dentro del texto.

Actividad 2
Ocultar la información previamente detectada, con el objetivo de minimizar el riesgo de divulgación, causando la menor distorsión posible en el contenido del documento.

¿Quién lo puede realizar?

El proceso lo puede realizar la misma oficina productora, teniendo en cuenta las técnicas de anonimización o sanitización pertinentes para tal fin y descritas en el presente documento.

¿Sobre qué documentos textuales puedo realizar el proceso?

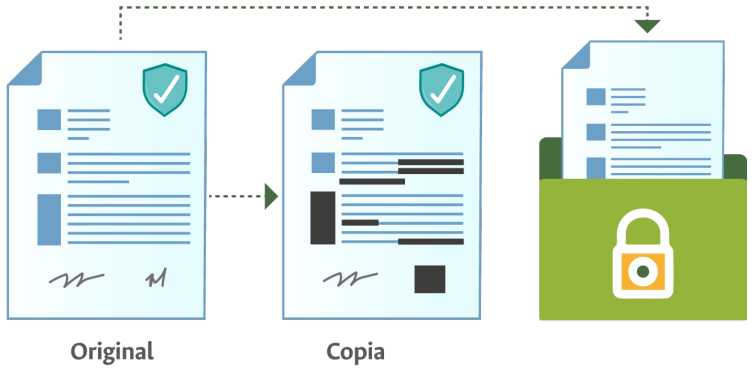
El proceso que tiene como finalidad la divulgación parcial de los contenidos y se realiza sobre los documentos electrónicos de archivo textuales que NO estén protegidos por las excepciones de la Ley 1712 de 2014.

¿Cómo se realiza el proceso?

El procedimiento descrito se puede realizar de manera manual o utilizando herramientas ofimáticas para tal fin. Para tener seguridad en el proceso se recomienda que se realice con el uso de la herramienta de PDF.

1

Tenga en cuenta que la divulgación parcial de los contenidos se realiza sobre los documentos electrónicos de archivo que NO estén protegidos por las excepciones de la Ley 1712 de 2014, es así como lo primero que debe hacer es validar el índice de información clasificada y reservada de la institución, para garantizar que el documento textual sobre el cual se realizará el proceso no hace parte de las series y subseries documentales objeto de reserva.



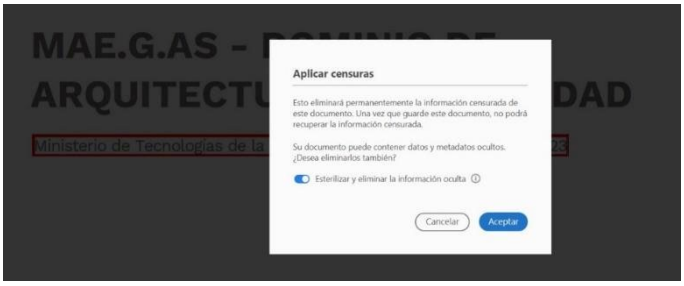
2 Una vez garantizado el cumplimiento del primer paso, proceda a realizar una copia del documento textual para realizar el procedimiento sobre la copia, tenga en cuenta que no se puede alterar el documento original.

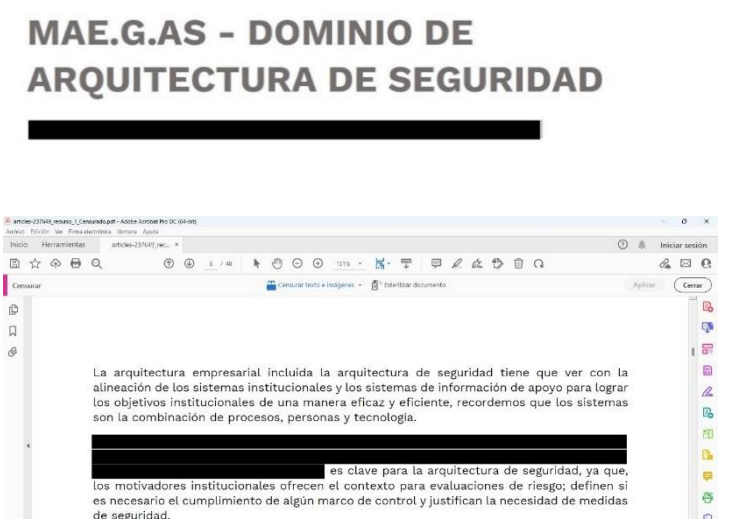
3 Identifique y liste los datos sensibles, confidenciales, o de identificación personal existente en él texto y defina cuál de las técnicas propuestas (redacción o enmascaramiento) que va a aplicar para la sanitización del contenido en aras de ocultar la información previamente detectada, con el objetivo de disponer de una versión pública del documento textual en donde se minimice el riesgo de divulgación, causando la menor distorsión posible en el contenido del documento.

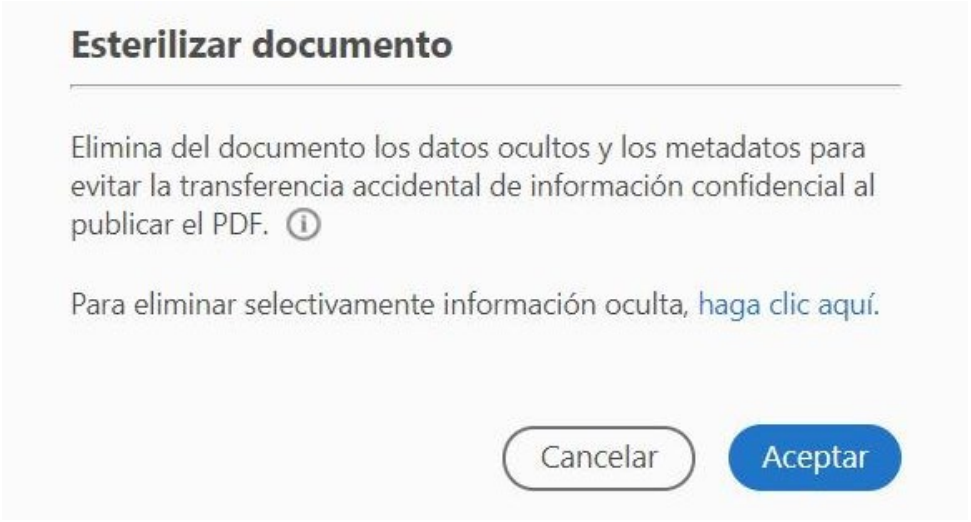
4 **Si la técnica a utilizar es la redacción,** tenga en cuenta que esto Implicará la eliminación o modificación permanente de la información personal, sensible, o confidencial identificable de un documento textual, y que esta técnica asegura que la información confidencial no pueda ser recuperada o reconocida.

5 **Si la técnica a utilizar es el enmascaramiento** tenga en cuenta que esto implicará únicamente la sustitución de la información personal, sensible, o confidencial con valores predeterminados,

Ejemplo práctico para la implementación de las técnicas en herramienta PDF:







6 Verificación, revisión y liberación de la versión pública proveniente y certificada por la misma oficina productora.

Finalmente, se presentan algunas herramientas útiles para el proceso, aclarando que no son las únicas:

Herramienta	Descripción	Características
CaseGuard	CaseGuard es una herramienta de redacción poderosa y confiable que	• Redacta cualquier documento independientemente de la industria. Las aplicaciones típicas incluyen la eliminación de información confidencial, como nombres, números de tarjetas de

	elimina información confidencial de archivos PDF, imágenes, audio y video complejos.	crédito, direcciones de correo electrónico y otros datos privados de registros médicos, policiales, personales y bancarios, formularios de impuestos y más. ● Le permite detectar automáticamente y redactar rápidamente información confidencial, como caras, matrículas, personas, pantallas, papeles, vehículos y más, a partir de una amplia gama de imágenes. ● Le permite redactar, transcribir y traducir de manera fácil, rápida y automática archivos que contienen información confidencial. ● Le permite realizar una redacción masiva, consiguiendo eliminar de forma automática y rápida información confidencial de miles de páginas e imágenes y horas de grabaciones de video o audio. ● La solución integral es adecuada para varios tipos de archivos en casi cualquier industria. Sus flujos de trabajo intuitivos y fáciles de usar lo hacen adecuado para todos, incluidos aquellos con poca experiencia en redacción.
VIDIZMO	Herramienta de redacción impulsada por IA de VIDIZMO es la mejor solución. Utiliza un indexador de IA para identificar rápidamente información de identificación personal en archivos digitales.	● Redactar pruebas de audio y vídeo de múltiples fuentes, incluidas las grabaciones de vigilancia de CCTV, vídeos de cámaras corporales y de salpicadero, grabaciones de llamadas telefónicas, etc. ● Detecta y rastrea automáticamente rostros y cuerpos para una redacción rápida y sencilla. ● Pixela, recorta, difumina parcial y complentamenta, a testigos y transeúntes, matrículas, etiquetas de nombre y objetos personalizados de las pruebas de vídeo. ● Silencia o borra la información confidencial de las grabaciones de audio.
Sighthound Redactor	Sighthound oculta automáticamente personas, rostros, vehículos y matrículas y también permite la edición manual. Es compatible con Windows Desktop, Linux, servidores Windows o servicios en la nube. Ofrece redacciones automatizadas con soporte completo de API.	● Redacción automática de personas, rostros, vehículos y matrículas en videos con opciones de edición humana. ● Múltiples tipos de detección. ● Rastree objetos en cada cuadro a lo largo del video automáticamente. ● Redacción de audio. ● Opciones de implementación: escritorio, servidor cliente, basado en la nube. ● Carece de redacción de documentos. ● Sin opción de transcripción y traducción.

Wondershare PDFelement	Permite la eliminación permanente de información confidencial de documentos PDF. Esta herramienta de redacción edita documentos PDF con una variedad de funciones. Puede crear, editar, convertir, organizar y anotar sus archivos PDF cómodamente.	• Edita documentos PDF agregando texto, enlaces, imágenes o formas. • Resalta, subraye, agregue comentarios o dibuje formas sobre documentos PDF. • Convierte múltiples archivos de Word, Excel y PNG a PDF simultáneamente. • Organiza archivos PDF fusionando, dividiendo o girando páginas PDF. • Protege con contraseña archivos PDF.