La Informática Forense y sus Analogías: la Balística Digital

Bruno Constanzo*, Pablo Cistoldi, Juan Iturriaga*, Ana Haydée Di Iorio*
InFo-Lab, Laboratorio de Investigación y Desarrollo de Tecnología en Informática Forense
Universidad FASTA, Ministerio Público de la Provincia de Buenos Aires, Municipalidad de General Pueyrredon
Mar del Plata, Buenos Aires
**{bconstanzo, juan, diana}@ufasta.edu.ar ; pcistoldi@mpba.gov.ar*

Abstract

Con el progreso de las tecnologías de la información sobre múltiples aspectos de la vida diaria, se hace cada vez más común la presencia de evidencia digital en las investigaciones judiciales. La evidencia, en cuanto contenedor físico de datos con valor investigativo, presenta distintos grados de complejidad, y la evidencia digital es un exponente excepcional en este aspecto. Algunas preguntas poco frecuentes en cuanto a la evidencia digital, y muy difíciles de responder son ¿"Quién" o "Qué" creó esta evidencia? ¿La evidencia, fue modificada, editada o manipulada de alguna manera? Para responder a estas preguntas, es necesaria la aplicación de técnicas de "balística digital", una rama poco explorada de la informática forense en nuestro país.

1. Introducción

La informática forense se enfrenta constantemente a nuevos retos. Si bien los principios básicos de esta disciplina permanecen inalterables, año tras año aparecen problemáticas novedosas, que exigen renovar procedimientos, métodos y herramientas de abordaje. Dentro de esta necesaria y permanente actualización, la búsqueda y empleo de nuevas analogías y metáforas aparece como una estrategia prometedora. Un ejemplo particular de estos avances lo es el de la denominada *balística digital*, donde se cuenta tanto con técnicas clásicas, como nuevas técnicas que permiten extraer información complementaria. Éstas son aplicables a temas actuales, como podría ser, la transferencia de contenido por medio de redes sociales.

La intervención del experto puede girar alrededor de la búsqueda, obtención, análisis y/o presentación de la evidencia digital que exhibe la propia parte. Pero también puede requerirse al perito establecer si una evidencia presentada por la parte contraria es auténtica y, en caso de no serlo, determinar las circunstancias de su falsificación o adulteración.

Especialmente en el segundo caso, la labor del informático forense presentará algunas analogías con las tareas frecuentes de especialistas de otras disciplinas.

Frente a la creciente variedad de labores del informático forense, resulta útil establecer y aplicar categorías apropiadas para clasificar la evidencia digital. Estas categorías no son teóricas y abstractas, sino prácticas y contextuales, pues sirven como criterios orientadores para actuar en casos concretos:

1. Dentro de los propósitos de un caso específico, habrá evidencias digitales *atómicas* y *compuestas*. Las evidencias atómicas son aquellas cuya división carece de sentido investigativo o probatorio. Las evidencias compuestas son aquellas que poseen sectores o aspectos cuyo estudio separado reviste utilidad.
2. Del mismo modo, existen evidencias *no asociables*, *asociables* y *mixtas*. Las evidencias no asociables son aquellas que no poseen valor para esclarecer o probar los hechos. Las evidencias asociables son aquellas que aportan información relevante por sí mismas, y las evidencias mixtas lo hacen cuando los datos que aportan son vinculados con información proveniente de otras evidencias. En algunos casos una evidencia suministra determinada información por sí misma y, además, contribuye a aportar nueva información cuando se la relaciona con otras.

A los fines prácticos, la noción de la evidencia como "contenedor físico de información" es útil para la labor del especialista. Será éste quien, en función de las concretas necesidades y particularidades de cada caso, trazará los límites y relaciones pertinentes dentro del conglomerado de campos magnéticos, pulsos electrónicos y otras representaciones binarias que debe analizar.

De este modo, una evidencia podrá ser una carpeta de archivos, una partición de un disco, un conjunto de emails, o tan sólo un documento o una parte de éste, según el contexto procesal de cada caso. Lo mismo sucede con los conjuntos de evidencia relevante. Por ejemplo, en algunos supuestos puede ser factible y pertinente obtener y analizar un conjunto formado por distintas versiones de un documento, software de creación o edición, datos de memoria volátil y datos de tráfico de red, pues de este cotejo es posible extraer nueva información. Serán entonces los criterios de delimitación y de agrupación los que terminen definiendo, en cada caso, lo que constituye una evidencia y un conjunto de evidencias.

Cuando una de las partes aporta una evidencia digital o un conjunto de ellas al proceso, la contraparte puede aplicar otros criterios de delimitación y agrupación. Volviendo a los ejemplos, si la Defensa ofrece un determinado archivo de imagen como "evidencia", la Acusación tal vez necesite analizar la composición de ese archivo (que entonces será analizado como una *evidencia compuesta*). La descomposición de esta evidencia compuesta en *evidencias atómicas*, como pueden ser los metadatos propios de un software de edición específico, pueden llevar a sospechar que el archivo fue editado, o que su origen no es el invocado. Y ello, a su vez, puede tornar conveniente buscar evidencias de otras clases, como podrían ser el acceso a software de edición en el disco o en línea, y de evidencias que muestren su probable empleo para la creación o edición del archivo analizado.

A continuación, se dará el marco teórico, en primer lugar, desde un punto de vista más cercano al derecho, y luego profundizando en temas técnicos propiamente dichos. Luego se enunciará el problema, y se dará una introducción a un proyecto capaz de proveer de las herramientas necesarias para este tipo de análisis. Por último, se darán ejemplos de casos de aplicación, y se llegará a las conclusiones.

2. Marco Teórico

2.1. Nuevas analogías y metáforas

Para llevar a cabo eficazmente estas tareas, se requiere, por un lado, conocer las particularidades del caso y el rol específico asignado al experto en el contexto de la estrategia de la parte a quien se asiste. Por otro lado, también parece útil ampliar las herramientas conceptuales, una labor que puede ser potenciada mediante el uso de analogías, algo corriente en la informática. Hoy se habla con naturalidad de virus, gusanos, troyanos, todos estos términos que provienen de otras disciplinas.

Desde la perspectiva de la informática forense, quizás sea conveniente explorar la utilidad de incorporar a la especialidad conceptos provenientes de otras disciplinas más añosas en su aplicación forense:

- Una evidencia digital puede estar compuesta por varias capas o estratos "geológicos" con características específicas.
- También la física y la química ofrecen conceptos útiles: mezcla reactiva y no reactiva, aleación, fusión, etc. Estas analogías pueden servir para llevar al campo de la informática la identificación de adulteraciones e intrusiones, por ejemplo.
- En ocasiones será necesario realizar una serie de experimentos y cotejos para establecer si uno o más archivos fueron creados, editados y/o enviados empleando una misma aplicación/software de escritorio o en línea. Para facilitar esta tarea es útil (al igual que sucede con la balística) contar con una base de conocimiento actualizada que identifique las "marcas" específicas que deja cada aplicación, en un archivo, al ser utilizada.
- En algunos casos, se debe cotejar y correlacionar un conjunto de documentos (por ejemplo, imágenes) con la finalidad de saber si son parte de una misma "familia" y, eventualmente, establecer un "árbol genealógico" de las distintas versiones.
- En ciertas investigaciones, quizás convenga trazar "mapas" físicos y/o lógicos, para indicar los lugares de origen y las "rutas" de una o varias evidencias.
- Empleando el lenguaje caligráfico, el cotejo de evidencias y la identificación de rasgos característicos puede llevar a confirmar que un documento o parte del mismo fue hecho por "la misma mano" que elaboró otros de origen indubitado.
- Las nociones de sensibilidad y especificidad, originarias de la medicina, pueden ser aplicadas a los "indicadores digitales" de distintas clases de maniobras delictivas.

2.2. La balística digital

La balística digital es la rama de la informática forense que analiza los archivos de la evidencia digital para brindar información sobre el elemento (ya sea de hardware o software) que lo creó o modificó de alguna manera.

En general, la literatura sobre el tema se enfoca sobre el caso de analizar imágenes fotográficas para verificar su autenticidad, sin embargo, las consideraciones se

pueden llevar a un nivel de abstracción superior y referirse en general a cualquier tipo de archivo.

Las técnicas de balística digital, a grandes rasgos, pueden clasificarse en tres categorías definidas:

- Las técnicas **basadas en contenido** son aquellas que estudian cómo el elemento altera (de forma sutil) al objeto real que se intenta representar en el archivo.
- Las técnicas **basadas en metadatos** estudian cómo el elemento introduce o modifica campos o secciones en el archivo que llevan en sí los metadatos propios de la aplicación.
- Por último, las técnicas **basadas en la estructura** estudian como el elemento organiza los datos de una determinada manera al guardar el archivo[14, 15].

Las técnicas basadas en el contenido son extremadamente interesantes porque permiten estudiar la información de forma independiente a su contenedor digital, y se pueden realizar análisis muy detallados e incluso se puede inferir un historial de las modificaciones que sufrió la información a medida que se fue editando[1-3]. Para realizar este tipo de análisis se requiere un profundo conocimiento del dominio y también deben conocerse las partes físicas de los elementos involucrados. Un estudio posible sobre fotografía digital es la determinación de las aberraciones ópticas que introduce el lente de una cámara para poder contrastar contra una imagen de referencia un archivo que debe validarse. Otro ejemplo, sería el análisis del mapa DCT[1] de una imagen JPG, que permite visualizar si hay zonas que hayan sido modificadas[11]. Una desventaja de estas técnicas es que algunos tipos de estudio son susceptibles a la aplicación de técnicas anti-forenses[1, 3].

Las técnicas basadas en metadatos brindan información asociada a los campos EXIF, ICC, ID3, XMP, u otro estándar de metadatos dependiendo del tipo de archivo. Volviendo al ejemplo de la fotografía digital, una imagen suele ir acompañada de metadatos EXIF que indican marca y modelo de la cámara, fecha y hora de adquisición de la imagen, información de la fotografía en si (longitud focal, apertura, tiempo de exposición, etc.). El problema con estas técnicas es que los campos de metadatos suelen ser prescindibles y, con intención o accidentalmente, pueden eliminarse del archivo sin comprometer su integridad. De todos modos, en la medida que no se hayan visto afectados, la información proveniente de los metadatos puede ser gran valor investigativo.

[1] DCT se refiere a la *Discrete Cosine Transform*, la una transformación matemática que aplica el algoritmo JPEG sobre las imágenes para su compresión.

Finalmente, las técnicas de estructura permiten identificar tanto el *encoder*[2] utilizado en la creación del archivo, como el programa que realizó la última modificación sobre el mismo. Si bien pueden pensarse mecanismos que intenten ocultar esta información en la evidencia volviendo a grabar el archivo, siempre quedan los rastros del último elemento que lo modificó y ponen en evidencia estos intentos [6].

Todas estas sub-ramas de la balística digital son importantes, y no deben considerarse como excluyentes, sino como complementarias entre sí. Brindan información (y eventualmente evidencia) sutil sobre el origen de los archivos, y de esta forma permiten evaluar la autenticidad de los mismos, con fundamentos y métodos científicos sólidos.

2.3. Conceptos generales de archivos y evidencia digital

En un sistema informático, un archivo es la unidad más pequeña en la que un usuario puede almacenar información de manera persistente. Internamente, los archivos son simplemente una secuencia de *bytes*, que codifican información de acuerdo a una especificación de formato. Dependiendo del tipo de archivo, la información que puede ser representada en el mismo, habiendo distintos formatos que permiten guardar fotografías, documentos, audio, video, bases de datos, por citar algunos ejemplos. En la Figura 1 puede verse un ejemplo generalizado de formato binario en base a segmentos.

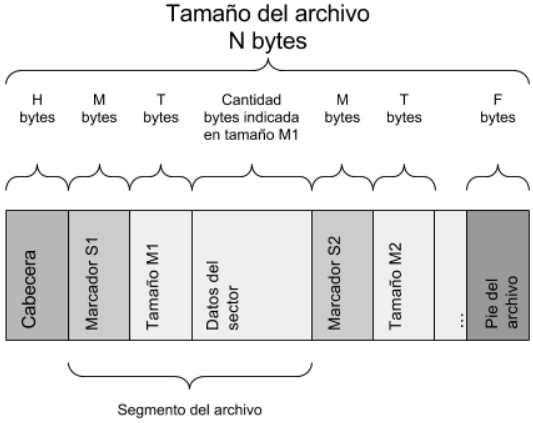

Figura 1. Estructura generalizada de formatos de archivo binarios organizados en segmentos.

Si bien ha habido un ligero resurgir de los formatos de texto basados en lenguajes de *markup*[3], la mayoría de los

[2] En los formatos donde hay un *encoder* de por medio, como JPG, PNG, formatos de audio y video.

[3] El ejemplo clásico de lenguaje de *markup* es el HTML utilizado en las páginas web, aunque la mayor cantidad de formatos heredan de XML por su diseño extendible.

formatos de archivo utilizados en la actualidad utilizan estructuras binarias para almacenar la información en forma eficiente. Estas estructuras binarias siguen una especificación de formato en la que se documenta cómo tiene que organizarse la información necesaria para representar el objeto cuya información se desea guardar, por ejemplo, una fotografía.

A modo ilustrativo, y de forma muy resumida, la estructura de un archivo de imagen JPG (ver Tabla 1) comienza con un identificador de "inicio de imagen", luego le sigue una (o varias) secciones "de aplicación" que llevan información sobre el alto, ancho y modo de compresión de la imagen, junto con otros datos y metadatos propios de las aplicaciones de edición o visualización. Luego siguen secciones que definen y almacenan tablas de cuantización y tablas de compresión (según sean necesarias), y secciones con la información de imagen propiamente dicha. Por último, siempre se encuentra un indicador de "fin de imagen" para indicar la finalización del archivo [4, 5].

Tabla 1. Estructura de un JPG ejemplo.

Marcador	Offset	Significado
SOI	0x000000	Comienzo de imagen
APP0	0x000002	Imagen JFIF
DQT	0x000014	Tabla de cuantización
DQT	0x000059	Tabla de cuantización
SOF	0x00009e	Inicio del frame (dibujado)
DHT	0x0000b1	Tabla de Huffman
DHT	0x0000d2	Tabla de Huffman
DHT	0x000189	Tabla de Huffman
DHT	0x0001aa	Tabla de Huffman
SOS	0x000261	Información de Imagen
EOI	0x22622e	Fin de Imagen

Los archivos, como elementos de un sistema digital, comparten las características de los mismos. En particular, una característica distintiva de los sistemas digitales es su tolerancia a la degradación.

En un sistema analógico, una señal se va atenuando y/o acumulando ruido durante su transmisión, o ciclo de vida. Por ejemplo, un disco de vinilo acumula desgaste sobre su superficie, alterando la calidad de sonido original. Si se amplifica una señal analógica atenuada y degradada, el ruido y las degradaciones acumuladas sobre la misma también se amplifican, y se obtiene una señal distorsionada (ver Figura 2).

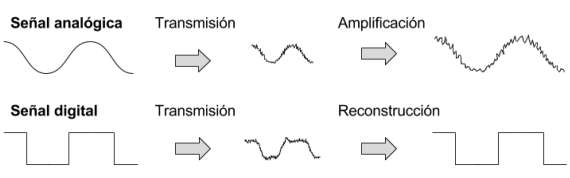

Figura 2. Señales analógica y digital y sus procesos de degradación y amplificación/reconstrucción.

Los sistemas digitales en cambio representan información discreta. Aunque estas se representen finalmente por medio de señales analógicas, lo importante es la etapa de cuantización de una señal, en la cual se asigna, de acuerdo a la intensidad de la misma, un valor discreto. En la medida que la atenuación y distorsiones sufridas por la señal no afecten los umbrales de detección, es posible reconstruir íntegra la señal digital original.

Esta característica, llevada a la evidencia digital, resulta en una de sus principales características: la evidencia completa puede reproducirse, sin alteración, tantas veces sea necesario. Los archivos también pueden copiarse de esta forma, para obtener tantas copias se necesitan, todas idénticas e indistinguibles del original.

Debe distinguirse también el proceso de copiado o duplicación, del proceso de recodificación o recompresión de un archivo. Volviendo al ejemplo de fotografía digital, no es lo mismo hacer una copia de una fotografía copiando el archivo a través del sistema operativo (*duplicación*) que abrir la imagen con un editor y volver a guardarlo (*recompresión*). En el segundo caso, se crea un nuevo archivo, que tiene nuevos datos y metadatos, producto del nuevo procesamiento que se realizó sobre la información.

3. Planteo del problema

El problema al que se enfrenta el perito informático, con relación a la balística digital, es que se le puede requerir alguna de las siguientes tareas:

- Indicar si un archivo, en cuanto evidencia digital, es auténtico.
- Indicar si un archivo, en cuanto evidencia digital, presenta evidencias de alteración o modificación.
- Detectar modificaciones en un archivo, que es o contiene evidencia digital.

4. Validación de archivos

Las técnicas de validación de archivos se desarrollaron como complemento a las técnicas de *file carving* y recuperación de archivos, un área de la informática forense en donde usualmente se manejan archivos recuperados que pueden estar contaminados por información de otros archivos, o no tener completa su estructura. Para poder filtrar los resultados útiles de aquellos que no pueden visualizarse correctamente, o para asistir a un algoritmo de recuperación [7, 8], las técnicas de validación aplican chequeos sobre los archivos recuperados y verifican que se respete la estructura de formato correcta.

En un proyecto previo, en el que participaron los autores de este trabajo, se desarrolló una implementación de validadores de archivos[9, 10], que provee un

framework de validación en lenguaje Python. Si bien el objetivo principal de este trabajo fue el soporte de herramientas de *file carving*, al entrar en contacto con las temáticas de balística digital se hizo evidente que el *framework File Validators* contaba desde sus versiones iniciales con la capacidad de realizar, de forma fácil y metódica, análisis de estructura. Con modificaciones adicionales pueden mejorarse estas capacidades e implementar los análisis de metadatos.

Actualmente el *framework* permite realizar análisis de estructura de los formatos JPG, PNG, GIF y LNK, además de proveer las funciones de validación para formatos SQLite3 y MS-OLE[4].

5. Aplicaciones prácticas

A continuación, se enuncian algunos ejemplos de situaciones que podrían resolverse aplicando técnicas de balística digital:

5.1. Fotografía adulterada

En el marco de una investigación, se desea verificar si una fotografía que fue presentada como evidencia, es original o fue adulterada con algún programa de edición de imágenes.

Aplicando análisis de contenido, pueden buscarse indicios de adulteración por medio del mapa DCT. Por otra parte, pueden verificarse los metadatos y la estructura del archivo coinciden con los que introduce un editor de imágenes o una cámara digital. Inconsistencias en al menos uno de estos elementos sería indicio de adulteración.

5.2. Análisis de video

Se presenta un video en donde se puede ver como una persona viene caminando en una dirección, y en un instante, aparece en otro lugar de la imagen y continúa su camino. El fiscal del caso pide que, por favor, se demuestre que el video ha sido editado.

En este caso, desde el punto de vista del análisis del contenido no es necesario trabajar más: resulta evidente que el video ha sido editado. Sin embargo, aplicando análisis de estructura y metadatos[12, 13] es posible demostrar que la estructura del archivo no coincide con la estructura que genera el cámara de origen, satisfaciendo así el requerimiento del fiscal.

5.3. Capturas de pantalla

En una causa se adjunta como evidencia capturas de pantalla de conversaciones de whatsapp para fundamentarla. Ante la falta de otras evidencias que soporten la denuncia, se pide verificar la autenticidad de los archivos presentados.

En primera instancia, se debe aclarar que es posible crear capturas de pantalla falsas que no presenten evidencia de modificación gráfica, ya que se altera el origen previo a la captura[5].

No obstante, haciendo un análisis sobre los metadatos y la estructura, es posible comparar las particularidades del formato original contra la evidencia provista. Si no concuerdan, es indicio que no se trata de los originales, y por lo tanto no es necesario realizar una investigación más profunda.

En este caso, se debe pedir al denunciante que aporte las capturas de pantalla originales, y si no puede producirlas, puede desestimarse la evidencia.

6. Conclusiones y Trabajo Futuro

Hay métodos científicos establecidos que permiten realizar la autenticación de documentos y determinar su origen y validez en un proceso legal. Estos métodos, con distintas complejidades de acuerdo a la dificultad de la tarea que pretenden realizar, están disponibles a las comunidades científicas para su evaluación, y a la comunidad legal para su utilización.

No solo eso, sino que se cuentan con implementaciones de software libre, realizadas por investigadores de nuestro país, que brindan un marco de trabajo para llevar a cabo algunas de estas tareas.

En cuanto al trabajo futuro que se puede realizar, es posible mejorar el *framework* de validación de archivos para mejorar sus interfaces de programación, su rendimiento e incorporar nuevos formatos de archivo.

Además, podría plantearse la creación y mantenimiento de una base de conocimiento de estructuras de formatos de archivo de acuerdo a programas y servicios para identificar los posibles orígenes. Apoyándose sobre la misma, sería posible realizar una aplicación o servicio web que realice un análisis preliminar, que permite a usuarios no expertos hacer una evaluación rápida de la evidencia, sin requerir un peritaje en casos que no lo ameriten.

7. Agradecimientos

Este trabajo es producto del Laboratorio de Investigación y Desarrollo de Tecnología en Informática Forense, InFo-Lab, impulsado por la Universidad

[4] Usualmente se reconoce este formato porque el que utilizan las versiones viejas de Microsoft Office para guardar los documentos, tanto de Word, Excel y PowerPoint, pero también es utilizado para guardar otros tipos de archivos en Windows.

[5] Por ejemplo, modificando el código HTML de una página web. Para el caso específico de conversaciones, hay aplicaciones que crean capturas de pantalla en base a configuraciones y datos provistos por el usuario.

FASTA, el Ministerio Público de la Provincia de Buenos Aires, y la Municipalidad de General Pueyrredon. Los autores agradecen a estas tres instituciones por brindar un espacio de trabajo único en el cual llevar a cabo este tipo de investigaciones.

8. Referencias

[1] Sencar, H. T., Memon, N., *Digital Image Forensics: There is More to a Picture than Meets the Eye*, Springer, 2012.

[2] Krawetz, N., "A Picture's Worth: Digital Image Analysis and Forensics", BlackHat Briefings 2007.

[3] Marrion, C. G., *Digital Image Manipulation Detection on Facebook Images*, University of Colorado, Master thesis, 2016. Disponible en http://www.ucdenver.edu/academics/colleges/CAM/Centers/ncmf/Documents/Theses/Marrion_Thesis_Spring2016.pdf.

[4] *JPEG Standard (JPEG ISO/IEC 10918-1 ITU-T Recommendation T.81)*, disponible en https://www.w3.org/Graphics/JPEG/itu-t81.pdf.

[5] *JPEG File Interchange Format version 1.02*, disponible en https://www.w3.org/Graphics/JPEG/jfif3.pdf.

[6] Farid, H., "Digital Image Ballistics from JPEG Quantization", TR2006-583, Department of Computer Science, Dartmouth College, Septiembre 2006.

[7] Garfinkel, S., "Carving Contiguous and Fragmented Files with Fast Object Validation", DFRWS 2007.

[8] Cohen, M., "Advanced Carving Techniques", Digital Investigation, Septiembre 2007.

[9] Di Iorio, A., Castellote, M., Podestá, A., Greco, F., Constanzo, B., Waimann, J., “El Framework CIRA, un aporte a las técnicas de File Carving.”. RADI Volumen II, Agosto 2013.

[10] *File Validators* disponible en https://github.com/info-lab/FileValidators

[11] He, J., Lin, Z., Wang, L., Tang, X., "Detecting Doctored JPEG Images Via DCT Coefficient Analysis", ECCV 2006.

[12] Hall, J. R., *MPEG-4 Video Authentication Using File Structure and Metadata*, University of Colorado, Master thesis, 2015. Disponible en http://www.ucdenver.edu/academics/colleges/CAM/Centers/ncmf/Documents/Theses/Hall_Thesis_Fall2015.pdf.

[13] Hall, J. R., "Video Authentication Using File Structura and Metadata", Presentación DFRWS 2015. Disponible en http://www.dfrws.org/sites/default/files/session-files/pres-video_authentication_using_file_structure_and_metadata.pdf

[14] Kornblum, J., "Using JPEG Quantization Tables to Identify Imagery Processed by Software", DFRWS 2008.

[15] Gloe, T., Fischer A., Kircher, M., "Forensic Analysis of Video File Formats", Digital Investigation 11, 2014.