OsiNET — DESARROLLO DE UNA HERRAMIENTA DE INTEGRACIÓN OSINET

OsiNET — DEVELOPMENT OF AN OSINET INTEGRATION TOOL

JOSÉ DOMINGO ROJO TORRES
Máster Universitario en Ciberseguridad. UNED Talavera

Resumen: OSINET corresponde a las siglas de Open Source Intelligence, haciendo referencia a la multitud de fuentes abiertas disponibles en internet para realizar una investigación de búsqueda de información. En este trabajo se ha desarrollado una aplicación que permite la organización y estructuración de la información obtenida durante una investigación OSINT usando las herramientas disponibles dentro del catálogo OSINT Framework y otros. Se han seleccionado aquellas herramientas sobre las que realizar extracción de datos utilizando *web scraping* para así, sintetizar la información de una forma ordenada y usable, generando un informe de inteligencia en base a los parámetros de búsqueda seleccionados.

Palabras clave: Framework, datos, inteligencia, navegador, investigación, scraping, aplicación

Abstract: OSINT stands for Open Source Intelligence, referring to the multitude of open sources available on the internet for conducting information search research. In this work, an application has been developed that allows the organization and structuring of information obtained during an OSINT investigation using the tools available within the OSINT Framework catalog and others. Those tools have been selected for data extraction using web scraping in order to synthesize the information in an organized and usable way, generating an intelligence report based on the selected search parameters.

Keywords: Framework, data, intelligence, browser, investigation, scraping, software.
.

1. INTRODUCCIÓN

En toda investigación, bien sea en el ámbito de la ciberseguridad o en cualquiera otra área, a menudo, nos lleva a la búsqueda de información a través de internet. Hoy en día, existen multitud de formas y lugares en el que es posible conseguir información de cualquier individuo, empresa, activo u otro tipo de actor dentro de un escenario de identificación de amenazas o prevención de ataques.

Normalmente, tenemos la información se encuentra diseminada por multitud de fuentes de carácter abierto, que son accesibles por cualquier entidad o sujeto que las necesite, lo cual constituye una ventaja, pero a su vez se convierte en un inconveniente puesto que, el hecho de disponer de tantas fuentes de datos hace que sea muy difícil localizar la información de un modo preciso, siendo en ocasiones muy complicado encontrar exactamente todo lo necesario para la investigación que se esté llevando a cabo.

Todas las investigaciones deben adoptar un enfoque metodológico para poder ser lo más eficaces posibles en su realización y es aquí donde aparece OSINT, *Open Source Intelligence*, término que hace referencia al conjunto de técnicas y herramientas para recopilar información pública, analizar los datos y correlacionarlos convirtiéndolos en conocimiento útil.

Como apoyo a este tipo de investigaciones y para ofrecer una solución sencilla a esta tarea de recopilación de información, se ha desarrollado OsiNET, la aplicación objeto de este proyecto, la cual se ha diseñado con el objetivo de minimizar los tiempos de búsqueda y recopilación de información, integrando, entre otros, gran parte de los enlaces disponibles en el proyecto OSINT *Framework*, https://osintframework.com, portal de referencia en este tipo de investigaciones donde se localizan agrupados por categorías una gran cantidad de páginas de búsqueda de información de distinta índole, además de enlaces a las diferentes herramientas de escritorio que pueden instalarse de forma local.

2. OBJETIVO DEL PROYECTO

Con la finalidad de disponer de un inicio sólido en una investigación OSINT, se pretende como objetivo de este proyecto, la realización de una herramienta de integración OSINT para disponer desde un único lugar, de las principales aplicaciones para la recolección y análisis de información y que se pueda realizar la recopilación de información en un informe final detallado con los datos que el usuario desee incorporar al mismo.

La aplicación está desarrolla en entorno *.Net*, en lenguaje *Visual Basic .Net*, y utiliza las herramientas disponibles en OSINT *Framework* donde, se tratará de agrupar de una forma ordenada y útil el mayor número de resultados, recopilados automáticamente desde el mayor número de herramientas posibles que sean de utilidad durante una investigación OSINT y poder facilitar el uso de una metodología de trabajo adecuada y la generación de informes, con ello, se pretende que el usuario pueda crear un punto de inicio sólido a la hora de realizar una investigación OSINT y ganar eficacia en los inicios.

El funcionamiento consistirá principalmente en la búsqueda de ciertas palabras clave que el investigador podrá introducir en los apartados correspondientes, como, por ejemplo, nombre de usuario, email, nombre de una empresa, CIF, IP, etc... y a su vez, la herramienta iniciará una búsqueda en varias de las herramientas webs de localización de información, para que el usuario determine qué información le resulta de interés y cuál quiere incorporar al informe de la investigación. También existe la posibilidad de generar el informe de forma automática.

El método usado para la recopilación de información será *web scraping*, técnica que permite leer el código HTML de una página web una vez descargado y localizar los elementos claves

junto con la información asociada para cada página, lo cual implica un desarrollo exclusivo para cada una de ellas.

Debido a que OSINT *Framework* dispone de un inmenso abanico de enlaces a aplicaciones, el primer paso ha sido seleccionar aquellos que son útiles para tal fin y que siguen activos y que sean de acceso web, puesto que el catálogo no se ha actualizado y muchos de ellos no están disponibles. Tras un exhaustivo análisis, se ha determinado utilizar aquellos sitios en los que se pueda realizar *web scraping* o que sean de un gran interés por su utilidad, que sean gratuitos y que no se precise registro, salvo en redes sociales. De este modo se consigue que la utilización de la herramienta sea más accesible.

3. CONCEPTOS TEÓRICOS

A continuación, vamos a profundizar en el concepto OSINT y su entorno, también veremos qué es *web scraping*, cómo aplicarlo correctamente y recopilar la identificación de las webs de fuentes de datos abiertos que consultaremos.

3.1 Estado del Arte

A lo largo de este apartado, exploraremos el concepto de OSINT (Inteligencia en Fuentes Abiertas) y las técnicas, pasos y herramientas que se pueden utilizar para llevar a cabo investigaciones en Internet, proporcionando diversas alternativas de obtención de información valiosa.

OSINT consiste en aprovechar cualquier fuente de información pública disponible para obtener datos sobre una empresa, organización o individuo. Como investigador, esta información puede ayudarte a responder preguntas específicas y obtener una visión más completa de los sujetos de tu investigación. Para ser más precisos, podemos decir que OSINT sintetiza el concepto de todo el conocimiento recopilado a partir de fuentes de acceso público y su proceso de recopilación. Este proceso incluye la búsqueda, selección, adquisición de la información, procesado y su posterior análisis con la finalidad de conseguir conocimiento que pueda ser utilizado en la investigación.

OSINT permiten obtener información relevante sobre personas y empresas. En la actualidad, con la expansión de Internet y las redes sociales, existe una gran cantidad de información pública compartida por los propios usuarios. Al utilizar estas fuentes, y aplicar inteligencia adecuada, podemos identificar a personas y organizaciones, así como anticiparnos a determinadas situaciones.

Además de Internet, hay una gran variedad de fuentes abiertas a partir de las cuales se puede obtener información relevante, referenciando principalmente:

- Medios de comunicación: revistas, periódicos, radio, etc.
- Información pública de fuentes gubernamentales.
- Eventos, conferencias, simposios, artículos académicos, bibliotecas en línea.
- Foros, redes sociales, blogs, wikis, etc.
- Archivos y bibliotecas.
- Boletines oficiales.
- Buzones y directorios públicos.
- Páginas blancas y amarillas.
- Obituarios y cementerios (para obtener información histórica o genealógica, por ejemplo).

En definitiva, podemos afirmar que las técnicas OSINT son en conjunto, una poderosa herramienta para recopilar información valiosa y relevante de fuentes abiertas y públicas en Internet y otras fuentes disponibles, lo que puede resultar fundamental para diversas investigaciones.

3.2. Historia de OSINT

Los inicios de OSINT [2] se remontan a 1941 cuando, en Estados Unidos surge la necesidad por parte de la *Foreing Broadcast Information Service* (FBIS), de recopilar información y generar inteligencia utilizando como fuente las transmisiones extranjeras que promocionaban la guerra, tras su análisis.

Se cree que, gracias a este trabajo, la FBIS pudo anticipar la intención de Japón de participar en la segunda guerra mundial.

Más recientemente, en 2013, la Agencia de Seguridad Nacional de Estados Unidos (NSA), hizo público un documento de formación para sus agentes, en el que instruía sobre el uso de Google para buscar información relevante.

A día de hoy, OSINT es ampliamente utilizado en todo el mundo por todo tipo de actores, desde fuerzas de seguridad, hasta el ámbito político, civil y empresarial.

3.3. Características

Las principales características respecto a la utilización de fuentes abiertas son las siguientes: [9]

- Eficiencia: Se tratan acciones donde la inversión realizada en tiempo y recursos es muy inferior a los beneficios que se obtienen.

- Rapidez: Debido a la disponibilidad de la información, podemos realizar todas las acciones de investigación en un corto espacio de tiempo.

- Intermediado: Puesto que la información ha sido generada por terceros, todos los datos han debido pasar por un intermediario, de forma que las fuentes pueden llegar a ser conocidas.

- Dependencia: Debido a que existe una publicación de información mediante una fuente y una obtención de información por un receptor, inevitablemente se crea una relación de dependencia entre ambos.

- Accesibilidad: El coste económico de conseguir información es muy bajo, lo cual hace que sea posible por un gran público.

- Voluminosidad: Debido al alto número de actores que comparten información y que hoy día es posible gracias al despliegue generalizado y global de internet, la cantidad de información publicada es de un alto volumen.

3.3.1. *Ventajas*

Una vez enumeradas las características principales de OSINT podemos identificar cuáles son sus principales ventajas [1]:

- Accesibilidad a bajo coste.

• Sencillez de acceso a la información.
• Información actualizada.

Estos tres puntos, nos llevan al cuarto y el objetivo buscado:

• Permite llevar a cabo una investigación de manera fácil y eficiente.

3.3.2. *Inconvenientes*

Como principales inconvenientes o problemas podemos identificar los siguientes:

• Exceso de información: No obstante, uno de los puntos fuertes anteriormente mencionados, también es una de las mayores desventajas, puesto que la gran cantidad de información, hace que en muchas ocasiones sea difícil discriminar la información válida y veraz de la que no lo es. Debido a ello, es necesario realizar un claro análisis de los datos concretos que precisamos para nuestra investigación.

• Fiabilidad de las fuentes: Al existir una gran cantidad de puntos donde obtener información, es muy importante realizar un correcto análisis de los lugares que vamos a consultar, así como de utilizar las técnicas de verificación adecuadas para asegurarnos que la fuente usada es la correcta.

3.4. Casos de uso

La metodología de inteligencia de fuentes abiertas (OSINT) tiene usos muy diversos, desde prevenir o parar amenazas de seguridad, investigación de mercado hasta inteligencia competitiva, por lo que es utilizado por distintos organismos, como, por ejemplo, gobiernos, empresas y organizaciones no gubernamentales. [2][9]

Los métodos más comunes en los que OSINT puede ser utilizado son: [3]

• Seguridad e Inteligencia: recopilar información de amenazas de seguridad, como puede ser actividad terrorista o ciberataques, o para recopilar información sobre gobiernos, organizaciones o individuos extranjeros.

• Estudio de mercado: recopilación de información para planificar estrategias o decisiones empresariales, analizando así información sobre competidores, tendencias industriales o el comportamiento de los consumidores.

• Investigación académica: utilizada por investigadores para recolectar información y datos.

• Periodismo de investigación: uso de modo que recolecta información sobre temas variados, que pueden variar desde políticas, negocios y crimen, que ayudará a descubrir nuevas historias o proporcionar evidencias.

• Procedimientos legales: compendiar información sobre evidencias o conductas debido a las diligencias de potenciales testigos o acusados.

• Recursos humanos.

• Investigadores privados.

• Cuerpos y fuerzas de seguridad.

• Ámbito militar.

• Ingenieros sociales.

3.5. Técnicas de investigación

Para la obtención de información mediante la inteligencia de fuentes abiertas (OSINT), existen tres métodos que pueden utilizarse [2] [9]:

• Mediante la obtención activa: a través de este método, el analista interactúa directamente con su objetivo, es decir, llevará a cabo consultas periódicas al servidor que quiere analizar, realizando la acción sin pasar desapercibido. Se aprovecha de vulnerabilidades, correos electrónicos, sistemas operativos, etc., lo que puede dejar rastro y el administrador, si tiene los conocimientos suficientes, puede conseguir ver los registros e intentos de los dispositivos que se han utilizado. Gracias a este método, el analista logra más información que con los métodos siguientes.

• Mediante la obtención pasiva: Consta de obtener información sobre el objetivo siendo discretos, que el administrador no tenga conocimientos sobre el analista. Es por ello, que se deben tener en cuenta una serie de precauciones técnicas a la hora de recopilar esa información, y que no quede rastro de la presencia del profesional en la red. Debido a que solamente va a poner acceder a la información que se pone en evidencia, la función del investigador/analista va a ser reducida a obtener los datos que van circulando por la red que el objetivo va generando.

• Mediante la obtención semi-pasiva: Trata de que el profesional debe generar tráfico en la red a través de consultas al dispositivo, que es donde se encuentra la información que se quiere obtener y analizar. Igual que en la obtención pasiva, es muy importante mantener un perfil discreto, pero a este método le acompaña el riesgo del desconocimiento de los riesgos, por lo que será más fácil que el objetivo detecte al investigador/analista.

3.6. Ciclo de Inteligencia vs Ciclo OSINT

OSINT es mucho más que obtener información. Para que una investigación se consiga llevar a buen puerto y que cumpla con la rigurosidad esperada hemos de utilizar un proceso bien definido de tal forma que nos permita hablar de un método por el cual se transforma información en inteligencia. [1][2][8][12]

Este método es el que podemos denominar Ciclo de Inteligencia o Ciclo OSINT, el cual está definido por más de una institución a nivel mundial. Por ejemplo, El CNI y la CIA disponen de proceso similares, pero diferenciando el número de pasos con la misma finalidad: la obtención de inteligencia a partir de información.

Derivado del ciclo de Inteligencia, el cual es de un amplio espectro de aplicación, podemos realizar una especificación denominada Ciclo OSINT [8] que así es como lo menciona el INCIBE.

A continuación, se detallan cada uno de ellos:

3.6.1. *Ciclo de Inteligencia - CNI*

Ilustración 1. *Ciclo de Inteligencia CNI*

El ciclo de inteligencia [12] que maneja el CNI para sus investigaciones consta de los siguientes pasos, los cuales podemos revisar en el siguiente gráfico y descripción:

• DIRECCIÓN Y PLANIFICACIÓN: El objetivo de esta fase es establecer los requisitos de inteligencia de la organización consumidora, así como la planificación de acciones tendentes a su obtención. Se prepara el escenario para el desarrollo del Ciclo de Inteligencia. La dirección nos responderá a la pregunta ¿Qué vamos a conseguir? Y la planificación a ¿Cómo vamos a conseguir?

• RECOLECCIÓN: El objetivo de esta fase es recopilar los datos en bruto necesarios (información) para producir el producto final deseado (inteligencia). La recolección de la información en bruto puede tener lugar a través de cinco fuentes básicas:

o Inteligencia Geoespacial, Geospatial Intelligence [GEOINT],
o Inteligencia Humana, Human Intelligence [HUMINT],
o Inteligencia de Fuentes Abiertas, Open-Source Intelligence [OSINT], y
o Inteligencia de Señales, Signals Intelligence [SIGINT].

• TRANSFORMACIÓN: Esta fase tiene por objeto convertir el formato de los datos en bruto recabados de las distintas fuentes, en aquellos otros formatos que posibiliten su posterior tratamiento y análisis.

• ANÁLISIS Y PRODUCCIÓN: El objeto de esta fase es integrar, evaluar, analizar y preparar la información previamente procesada, de cara a obtener el producto final deseado (inteligencia). Al igual que en la anterior, la fase de Análisis y Producción requiere de personal altamente especializado y entrenado (los analistas).

•DIFUSIÓN: El objeto de esta fase es entregar el producto final al consumidor de inteligencia que lo ha solicitado, así como a otros actores, cuando ello sea necesario y jurídicamente admisible. En esta fase, el consumidor de inteligencia recibe el producto final.

• EVALUACIÓN: Se trata de una fase transversal a todos los pasos del proceso. Durante la ejecución de las distintas fases que componen el Ciclo de Inteligencia siempre es posible realimentar cada una de ellas a partir de los resultados obtenidos, utilizando tal realimentación para ajustar y refinar las actividades realizadas, individualmente consideradas, como el Ciclo, en su conjunto.

3.6.2. *Ciclo OSINT - INCIBE*

Ilustración 2 - *Ciclo OSINT INCIBE*

Por otro lado, el ciclo OSINT se define mediante el siguiente proceso [2][8]:

• REQUISITOS: En esta fase inicial del proceso debemos definir las correctas preguntas a responder.

• FUENTES DE INFORMACIÓN: Una vez definidas las preguntas, evaluamos cuáles serán los puntos de información y herramientas a utilizar.

• ADQUISICIÓN: Suele ser la etapa que más se dilata en el tiempo, puesto que es el punto donde se consigue la información.

• PROCESAMIENTO: En esta parte se estructura la información para que pueda ser analizada acorde a los requisitos especificados.

• ANÁLISIS: Una vez procesados los datos, llegamos a la fase donde se genera la inteligencia.

• INTELIGENCIA: En esta parte final, se presenta la información obtenida de forma que pueda ser útil y comprensible con la finalidad de facilitar la posterior toma de decisiones.

Una vez analizado el ciclo, este se va a apoyar en una serie de herramientas. Una opción para saber qué tipo de herramienta se ajusta a las necesidades del usuario es OSINT *Framework*.

3.7. Análisis de la información

Merece una especial mención la fase de Análisis de información, puesto que supone el punto más importante de la investigación, puesto que su extracción, estructuración y jerarquización es crucial para realizar el informe de inteligencia que veremos en el siguiente apartado. [6]

La principal actuación a llevar a cabo en un análisis de información es la de construir una taxonomía, es decir, la creación de una clasificación de todos los elementos que componen el universo de estudio. Este ámbito se encuentra definido a través de la identificación, denominación y catalogación de los objetos que lo componen, los cuales se organizan en grupos basándose en algún factor o característica común a todos ellos.

Esta categorización se puede llevar a cabo utilizando diferentes técnicas y se pueden agrupar en cuatro grandes categorías de métodos analíticos, las cuales vemos reflejadas en la siguiente figura:

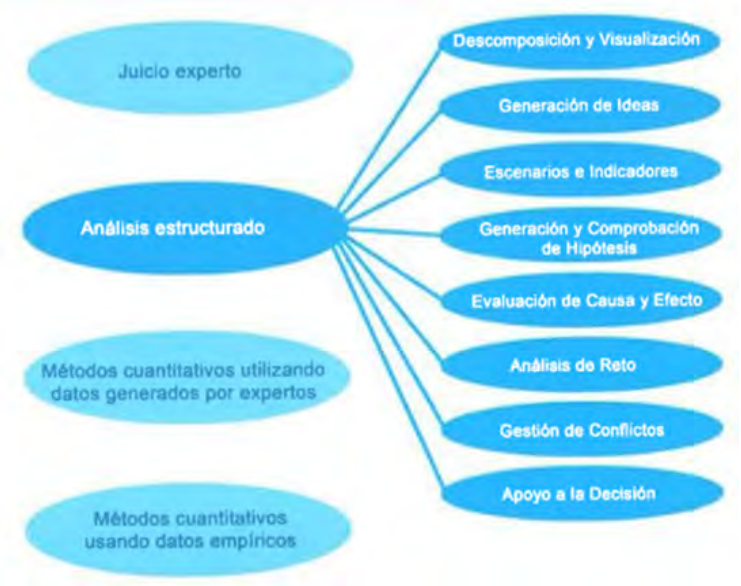

Ilustración 3 - *Categorías de Métodos analíticos*

Las características principales de cada categoría son las siguientes:

3.7.1 *Juicio experto*

Se trata de la forma tradicional en la que se llevan a cabo la mayoría de los análisis. Este método combina la experiencia en un campo del analista junto con el pensamiento crítico. Esta categoría incluye razonamiento probatorio basado en evidencias, método histórico, método de estudios de caso y razonamiento por analogía.

3.7.2 *Análisis estructurado*

Este método supone un proceso paso a paso de forma que expone las ideas del analista de manera fácilmente visibles para otros intervinientes, de forma que se posibilita la revisión del trabajo realizado, pudiendo ser discutido pieza a pieza, lo que convierte al análisis estructurado en un proceso colaborativo. La gran ventaja es que limita el impacto en la investigación de las carencias y trampas cognitivas conocidas.

Las técnicas utilizadas en este método incluyen *Brainstorming* estructurado, Escenarios, Indicadores, Análisis de Hipótesis Competidores y Comprobación de Asunciones Clave.

3.7.3 *Métodos cuantitativos utilizando datos generados por expertos*

Debido a la carencia de datos empíricos que muchos investigadores tienen para el análisis de una situación, se han realizado métodos para que datos cuantitativos generados por opiniones de expertos puedan ser utilizados y extrapolados a la investigación en cuestión.

Los métodos más usados son la Inferencia Bayesiana, modelado dinámico y simulación.

3.7.4 *Métodos cuantitativos usando datos empíricos*

Estos métodos utilizan datos recogidos mediante experimentación. Un ejemplo de este tipo de métodos es el Modelo Econométrico.

Hemos de señalar que, ninguno de estos métodos por sí mismo es más efectivo que otro. La utilización de cada uno de ellos será necesario según las circunstancias del problema a analizar y lo más probable es que se deban utilizar múltiples métodos durante el curso de un proyecto para obtener una respuesta satisfactoria.

3.8. Flujos de Trabajo

La materialización de los conceptos teóricos anteriormente descritos, debe llevarse a cabo mediante un procedimiento eficaz. Algunos autores como Michael Bazzel, nos trasladan una metodología de trabajo adecuada para realizar una investigación de forma ordenada y con una alta probabilidad de éxito.

A continuación, veremos los flujos de trabajo recomendados, según el autor, dependiendo de los datos de inicio de los que dispongamos junto con recomendaciones [24], [25], [26]:

En primer lugar, es muy importante contar con unos selectores bien definidos y precisos para poder disponer de una base sobre la que iniciar la investigación.

Los selectores son aquellos valores u objetos que asociamos a un concepto sobre el que queremos investigar.

Podemos diferenciar entre dos tipos de selectores:

3.8.1 *Selectores Comunes:*

Se trata de datos básicos para iniciar la búsqueda, normalmente se encuentran en formato texto y pueden investigarse directamente, sin ningún tratamiento previo, puesto que el elemento buscado se identifica únicamente con el valor del dato por el que es representado.

- Nombre
- Correo
- Alias/Nick
- Número de teléfono
- Empresa
- Documento de Identidad

3.8.2. *Selectores Complejos*

Este tipo de selectores son más difíciles de tratar, puesto que en sí mismos son un conjunto de información que tiene sentido individualmente y por la cual se pueden realizar búsquedas adicionales.

Además del contenido que representan, este tipo de selectores también entrañan información adicional en forma de metadatos, como puede ser información de coordenadas GPS, usuarios de creación, modificación, fechas y un largo etcétera que puede ayudar a la investigación.

- Foto
- Video
- Evento

Bazzel nos muestra diferentes flujos de trabajo según el selector desde el que partamos nuestra investigación.

3.8.3. *Nombre*

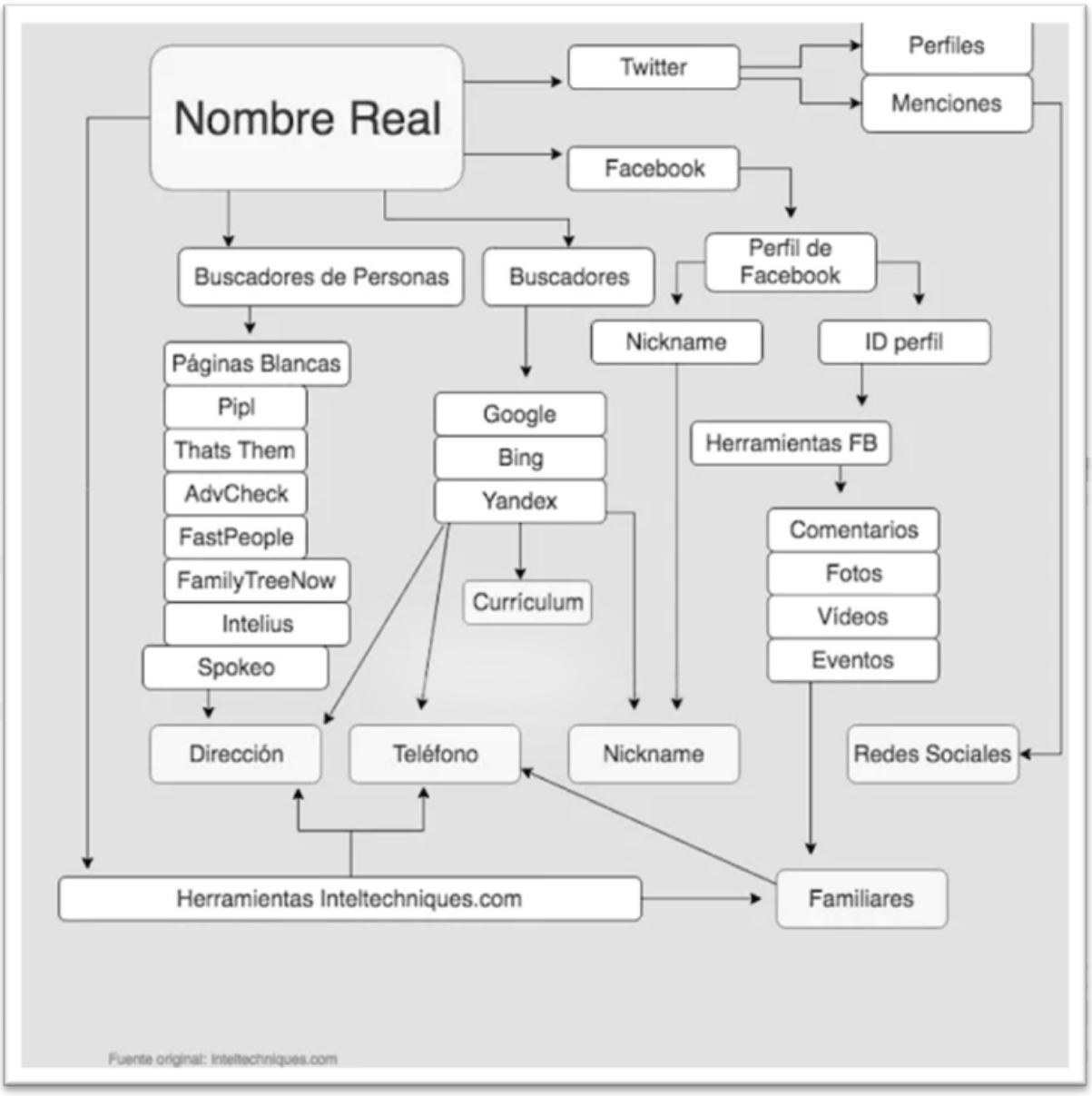

Ilustración 4. *Flujo Nombre*

3.8.4. *Usuario*

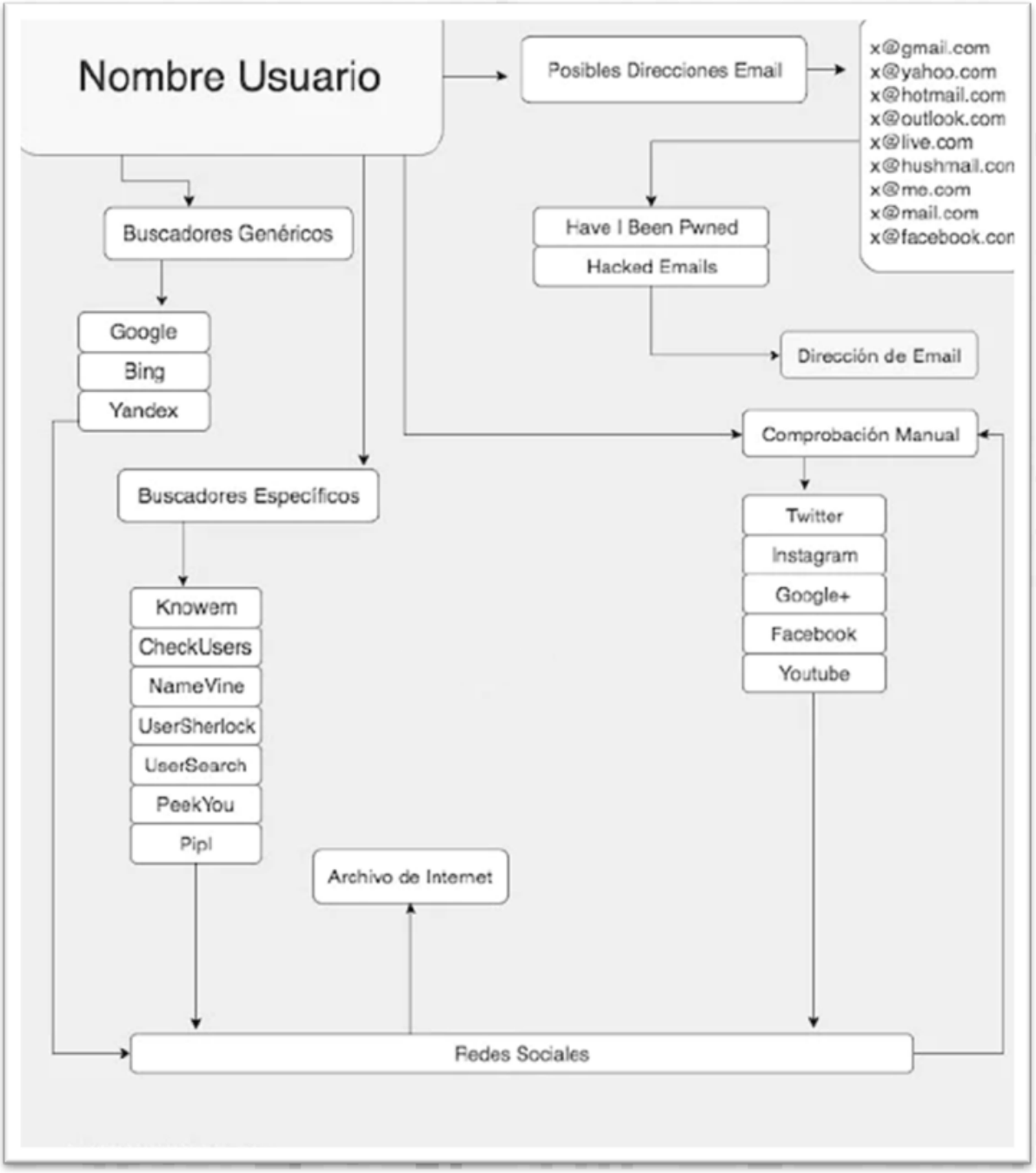

Ilustración 5. *Flujo Usuario*

3.8.5. *Ubicación*

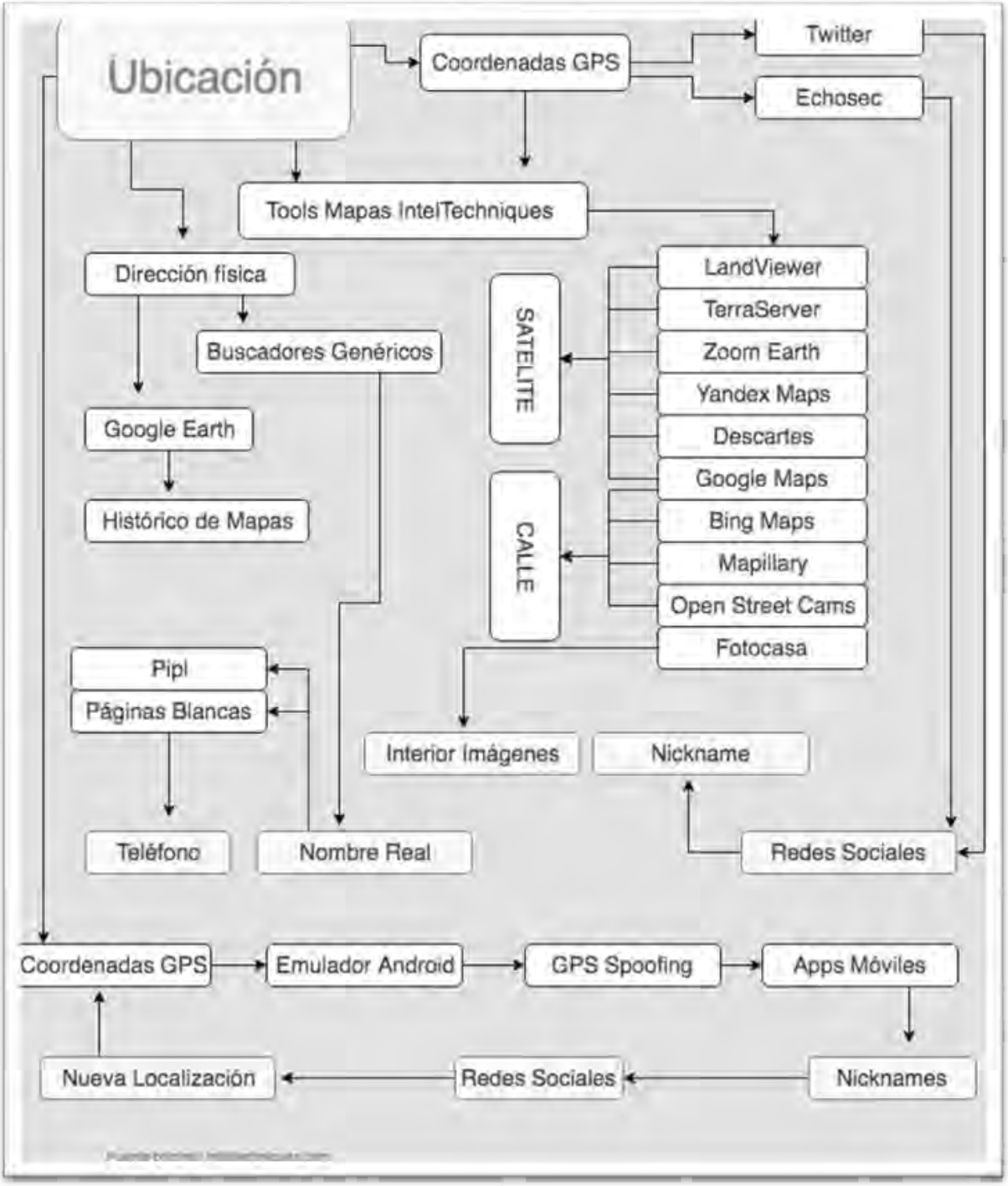

Ilustración 6. *Flujo Ubicación*

3.8.6. *Recomendaciones*

El autor aboga por el cuidado de la privacidad de cada investigador, hasta tal punto que es complicado encontrar información sobre él mismo en la red, y en su publicación ofrece algunos consejos para poder mantener nuestra privacidad a salvo durante una investigación. Dichas sugerencias, se muestran catalogadas según las características de cada tipo de dato o información que queramos mantener oculta y merecen ser comentadas:

• Redes Sociales: Existen dos premisas básicas a la hora de exponer información en una red social y que nos ayudarán a mantener nuestra privacidad y reputación asegurada. Estas premisas son:

o Nunca escribir algo en una red social que no diría ante un micrófono en presencia de miles de personas que no conozca.

o Nunca subir una foto o video a una red social que no expondría en una pantalla gigante dentro de un estadio de fútbol.

• Metadatos: Los ficheros que compartimos hoy en día en nuestro entorno digital, llevan asociados una serie datos e información muy interesante y que puede hacer que desvelemos más información privada de la que deseásemos. Se trata de los metadados, información contenida en la mayoría de los documentos que tratamos diariamente (.doc, .xls, .ppt, .pdf, .jpg, .png, etc.) y pueden desvelar el nombre de usuario de Windows, el nombre de nuestro equipo, elementos de nuestra red, fechas y horas de modificación y creación, información de coordenadas GPS y un largo etcétera. Para evitar que esta información sea trasladada junto con los ficheros que compartimos es muy interesante utilizar una herramienta de borrado de metadatos y hacer que nuestros documentos sean tratados por la misma antes de ser enviados.

• Geolocalización: Tanto gran cantidad de páginas web y redes sociales, como documentos, pueden llevar asociada información sobre nuestra geolocalización. En un primer momento puede parecer inofensivo, pero podemos revelar la información de nuestro domicilio o del lugar en el que nos encontramos si no tratamos de usar esta característica con precaución.

• Privacidad: El control de la privacidad en las herramientas que utilizamos es fundamental para evitar una exposición no deseada de los datos que comparten dichas aplicaciones, así como de información relevante derivada de nuestra actividad o falta de ella. Es muy común utilizar herramientas en las que se comparte el estado de nuestra conexión y si no verificamos con quién compartimos esos datos, es probable que estemos proporcionando información no deseada a usuarios que no conocemos.

• Servicios Online: Debido a que muchos de los servicios online que existen solicitan un email y contraseña para registrarnos y poder utilizarlos, nos exponemos a que, si ese sitio web sufre un ataque, nuestros datos sean revelados, por lo que es aconsejable disponer de una cuenta de correo que no tenga relación con nuestro nombre o datos reales, para una mayor seguridad.

• Eventos públicos: Del mismo modo, lo anterior se puede aplicar a muchos de los eventos, sorteos y promociones que nos solicitan nuestra información para promocionar sus servicios. En muchas ocasiones, al ser servicios puntuales, estos sitios descuidan la seguridad y privacidad de los datos que se recogen y pueden verse expuestos en la red con mayor facilidad que en otros servicios online.

• Foros: Se ha comprobado en diversas investigaciones que el uso de foros por parte del personal técnico de una compañía, deja mucha información sobre Sistemas Operativos, servicios, versiones, usuarios, políticas de seguridad y muchos detalles más sobre una compañía o a nivel personal, debido a que acuden a estos servicios buscando ayuda sobre un problema concreto. Es muy importante concienciar al personal para que, a la hora de buscar una solución, procuren no dejar más información que la necesaria.

• Auto búsqueda: Es una muy buena práctica realizar búsquedas periódicas de nosotros mismos, para así disponer de una medida sobre cuan expuestos estamos en la red y qué información estamos dejando en nuestras interacciones diarias.

• Alertas: Del mismo modo, es muy interesante incluir una alerta en alguno de los servicios que los buscadores o servicios adicionales proporcionan, para que se nos notifique en el momento en el que se detecten las palabras claves que hemos configurado en la alerta y tener controlada cualquier exposición de información.

3.9. El Informe de Inteligencia

El informe de inteligencia es considerado el artefacto final resultado de la investigación, donde expondremos los hechos, así como la información obtenida de forma analítica y usable, además de una colección de las fuentes consultadas. [1]

Esta última fase del ciclo de inteligencia es la más importante, ya que se trata de la herramienta definitiva para diseñar un plan de acción y una correcta toma de decisiones.

3.9.1. *Redacción del Informe de Inteligencia*

De forma habitual estos informes se componen de dos áreas diferenciadas:

• La parte ejecutiva: En esta parte se exponen los hechos, la información recopilada, después de jerarquizarla, analizarla y sintetizarla, la conclusión y, además, toda aquella información de relevancia que tenga que estar accesible de forma rápida.

• La parte técnica: Es donde se refleja el detalle de la investigación, anexo con las URLs de relevancia para la investigación (bibliografía), herramientas utilizadas para recabar la información, y todos y cada uno de los pasos que se han realizado para obtener y preservar la información.

Dependiendo de la complejidad de cada caso el informe requerirá más o menos nivel de detalle, pudiendo precisar datos o información adicional como mapas mentales, infografías, gráficos, fotografías, presentaciones o cualquier otro elemento que haga más comprensible la información aportada.

3.9.2 *Estructura del informe de Inteligencia*

Aunque cada informe puede ser realizado según las características y necesidades del objeto de la investigación, se puede definir la siguiente estructura como plantilla para un informe base:

• Identificación de documento: Se compone de los atributos identificativos del documento: Título, Fecha e identificador.

• Resumen: Recoge una síntesis de los hechos que promueven la investigación.

• Objetivos: Se identifica la finalidad de la investigación.

• Información recopilada: Se debe recoger una vez analizada y en función de su relevancia.

• Conclusión: Valoración final después del análisis de la información obtenida y consideraciones a tener en cuenta.

• Bibliografía: URLs visitadas más relevantes para la investigación, junto con todos los detalles y pasos que se han realizado para la obtención y preservación de la información.

3.10. Herramientas OSINT

Hay multitud de herramientas y servicios útiles a la hora de implementar un sistema OSINT. A continuación, se mencionan algunos de ellos: [4] [8]

3.10.1. *Buscadores habituales*

Google, Bing, Yahoo, Ask. Permiten consultar toda la información que indexan. Así mismo, permiten especificar parámetros concretos (Hacking con buscadores: por ejemplo «Google Hacking» o «Bing Hacking») de manera que se pueden realizar búsquedas con mucha mayor precisión que la que utilizan los usuarios habitualmente.

Dependiendo del buscador empleado se utilizan distintos parámetros, si bien algunos de ellos son comunes. Algunos ejemplos de búsquedas parametrizadas son las siguientes:

• Ficheros con extensión pdf de un sitio web concreto: site:cert.inteco.es + ext:pdf

• Algunos sitios hackeados: intitle:”hacked by SultanHaikal”

Mediante estos parámetros se puede obtener, entre otras cosas, información sensible como nombres de usuarios y contraseñas procedentes de volcados de bases de datos, localización de servidores vulnerables, acceso a dispositivos hardware online como webcams, cámaras de vigilancia o impresoras, o datos personales como DNI, cuentas bancarias, etc.

3.10.2. *Buscadores especializados*

• Shodan: Permite entre otras cosas localizar ordenadores, webcams, impresoras, etc. basándose en el software, la dirección IP, la ubicación geográfica, etc. Mediante este servicio es posible localizar información de interés y, en ocasiones, curiosa e incluso inquietante como, por ejemplo: acceder al sistema de control de una pista de patinaje sobre hielo en Dinamarca y descongelarla, poner en modo de prueba todo el sistema de control de tráfico de una ciudad o acceder al sistema de control de una planta hidroeléctrica en Francia.

• NameCHK: es una herramienta que permite comprobar si un nombre de usuario está disponible en más de 150 servicios online. De este modo, se puede saber los servicios que utiliza un usuario en concreto, ya que habitualmente la gente mantiene dicho nombre para todos los servicios que utiliza. Además, disponen de una API que permite automatizar las consultas.

• Knowem: es una herramienta de similares características que MameCHK pero comprueba el nombre en más de 550 servicios, incluyendo dominios disponibles.

• Tineye: es un servicio que, partiendo de una imagen, indica en qué sitios web aparece. Es similar a la búsqueda por imagen que incorpora Google Imágenes. Buscadores de información de personas: permiten realizar búsquedas a través de diferentes parámetros como nombres, direcciones de correo o teléfonos. A partir de datos concretos localizan a usuarios en servicios como redes sociales, e incluyen posibles datos relacionados con ellos como números de teléfono o fotos. Algunos de los portales que incorporan este servicio son: Spokeo, Pipl, 123people o Wink.

3.10.3. *Herramientas de recolección de metadatos*

• Metagoofil: permite la extracción de metadatos de documentos públicos (pdf, doc, xls, ppt, docx, pptx, xlsx). A partir de la información extraída se pueden obtener direcciones de correo electrónico del personal de una empresa, el software utilizado para la creación de los documentos y por tanto poder buscar vulnerabilidades para dicho software, nombres de empleados, etc.

• Libextractor: es una aplicación similar a Metagoofil que soporta muchos más formatos, si bien la información obtenida no es de tanta utilidad.

3.10.4. *Servicios para obtener información a partir de un dominio*

• Domaintools: es uno de los servicios referentes en este ámbito ya que incorpora un gran número de funcionalidades. Cabe destacar que permite crear alertas a usuarios que registran dominios, monitorizar dominios e IPs, crear alertas para dominios nuevos que contengan ciertas palabras, e incluso un servicio de investigación de gran cantidad de amenazas como «spear phishing», denegación de servicio, spam, fraude o malware.

• Robtex: muestra, entre otras cosas, la fiabilidad del dominio, su posición en el ranking Alexa, el listado de subdominios, los servidores de correo o el ISP que utiliza.

• MyIPNeighbors: permite obtener el listado de dominios que comparten servidor con el dominio indicado.

3.10.5 *APIs de diferentes servicios y RRSS*

Mediante los métodos que implementan se pueden consultar de una manera automatizada los datos publicados.

3.10.6. *Otras herramientas de interés*

• GooScan: permite automatizar búsquedas en *Google* pudiendo identificar de una manera sencilla subdominios de un dominio concreto, fugas de información o posibles vulnerabilidades.

• SiteDigger: al igual que *GooScan* permite automatizar búsquedas. Busca en la caché de Google para identificar vulnerabilidades, errores, problemas de configuración, etc.

• OsintStalker (FBStalker y GeoStalker): utilizan diferentes redes sociales como Facebook, LinkedIn, Flickr, Instagram y Twitter para recolectar gran cantidad de información sobre una persona. Permiten localizar lugares y sitios web visitados con regularidad, amigos online, etc. y mostrar los datos en *Google Maps*.

• Cree.py: permite obtener datos de *Twitter*, *Flickr* e *Instagram*. A partir de la selección de una cuenta extrae fechas e información GPS, y crea una base de datos en formato csv o kmz para visualizarlos.

• Theharvester: esta herramienta obtiene emails, subdominios, host, nombres de empleados, puertos abiertos, etc. a través de diferentes servicios como *Google*, *Bing*, *LinkedIn* y *Shodan*. Merecen una mención especial Palantir y Maltego al implementar un gran número de funcionalidades y ser unos de los grandes referentes en la materia de la inteligencia de las fuentes abiertas.

3.10.7. *Palantir*

Es una empresa que tiene como cliente a diferentes servicios del Gobierno de Estados Unidos (CIA, NSA y FBI) y que se centra en el desarrollo de software contra el terrorismo y el fraude, mediante la gestión y explotación de grandes volúmenes de información.

3.10.8. *Maltego*

Permite visualizar de manera gráfica las relaciones entre personas, empresas, páginas web, documentos, etc. a partir de información pública.

3.11. Consideraciones legales

El marco jurídico vigente para la investigación de delitos cometidos por medio de los servicios de la sociedad de la información (Internet, redes sociales, correos electrónicos, etc.), viene dado por los siguientes textos legales:

• Ley 34/2002, de 11 de julio de *Servicios de la Sociedad de la Información y Comercio Electrónico* (redacción según Ley 56/2007, de 28 de diciembre, de *Medidas de Impulso de la Sociedad de la Información*, modificada posteriormente por la Ley 2/2011, de 4 de marzo, de *Economía Sostenible*) [13].

• Ley 25/2007, de 18 de octubre de *conservación de datos de comunicaciones electrónicas y redes públicas de comunicación*. [14].

• Real decreto de 14 de septiembre de 1982 por el que se aprueba la *Ley de Enjuiciamiento Criminal.* [15]

• Ley Orgánica 13/2015, de 5 de octubre, de modificación de la *Ley de Enjuiciamiento Criminal* para el fortalecimiento de las garantías procesales y la regulación de las medidas de investigación tecnológica. [16]

• Ley Orgánica 10/1995, de 23 de noviembre, del *Código Penal*. [17]

• La *Constitución española* de 1978. Título I. De los derechos y deberes fundamentales. [18]

3.12. OSINT Framework

Podemos definir OSINT Framework como un repositorio categorizado de enlaces a multitud de fuentes de datos y herramientas para realizar de una manera ágil y rápida una investigación OSINT. Puede ser definido [20] como: "Un repositorio online que contiene multitud de recursos para llevar a cabo búsquedas en fuentes de información abierta."

Las principales categorías de búsqueda son:

• Nombres de usuarios.
• Direcciones de correo electrónico.
• Direcciones IP.
• Recursos multimedia.
• Perfiles en redes sociales.
• Geolocalización.

Y podemos destacar como mejores cualidades su sencillez, un manejo intuitivo y la gran cantidad de recursos disponibles para cada categoría.

Antes de comenzar, hemos de indicar que, el acceso a este conjunto de enlaces se obtiene mediante un portal web en la dirección https://osintframework.com/ y es muy importante conocer la disposición de los enlaces y la nomenclatura asociada a cada herramienta.

• Si no aparece ninguna letra al lado del nombre de la herramienta significa que se abrirá una nueva ventana o pestaña del navegador con la web referenciada en el link.

• (T) – Indica que la herramienta a la que hace referencia nos traslada a una página para descargar e instalarla localmente.

• (D) – Indica que se precisa de *Google Dork* para su funcionamiento.

• (R) – Es necesario registrarse en la herramienta para su utilización.

• (M) - Indica que la propia URL debe contener el término de búsqueda, por lo que se debe editar manualmente para que nos muestre una página de resultados.

Su funcionamiento es sencillo, puesto que simplemente hay que desplegar categorías y subcategorías de los elementos del árbol hasta llegar al enlace que queremos y una vez allí nos abrirá una nueva ventana con la página donde podremos realizar la búsqueda o la instalación de la herramienta seleccionada.

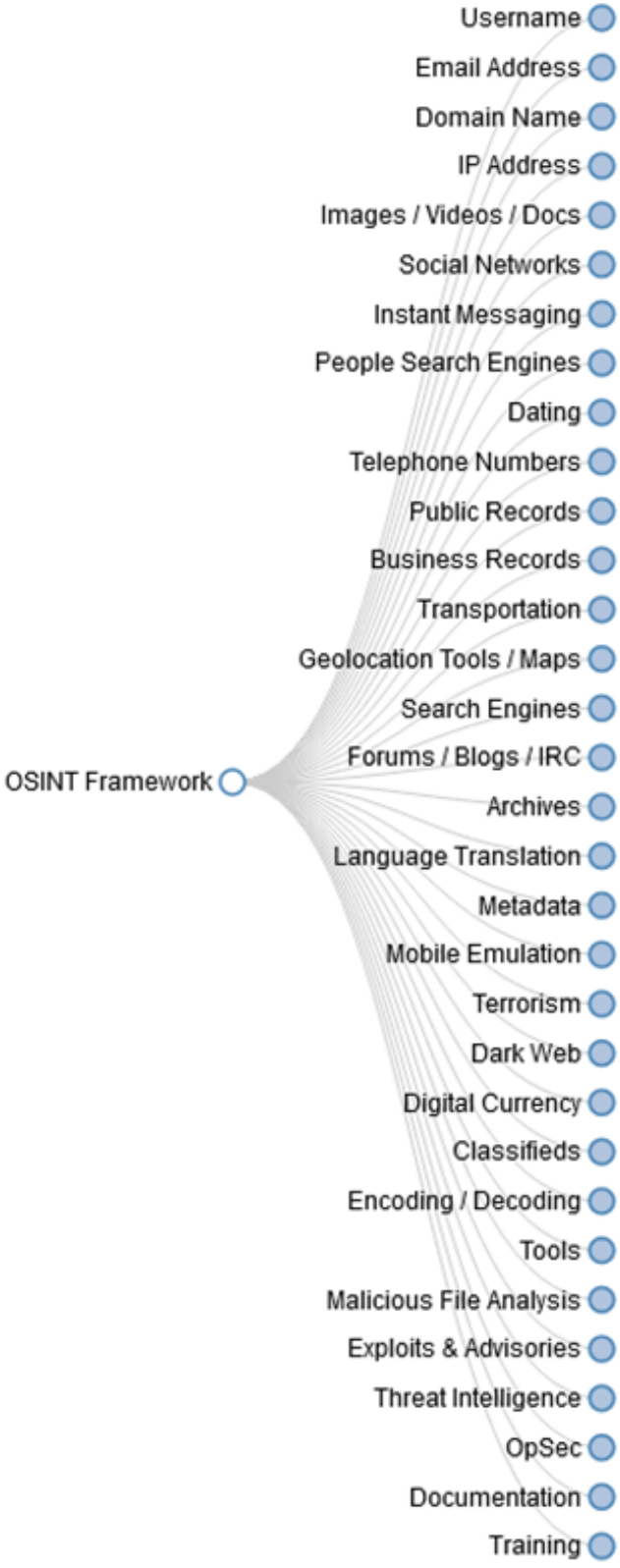

Ilustración 7. *Árbol OSINT Framework*

Imaginemos un caso de uso, donde se precise localizar las redes sociales en las que un usuario se ha registrado. Para ello podemos utilizar la herramienta "Namechk" que se encuentra dentro de la categoría "Username" y subcategoría "Username Search Engine".

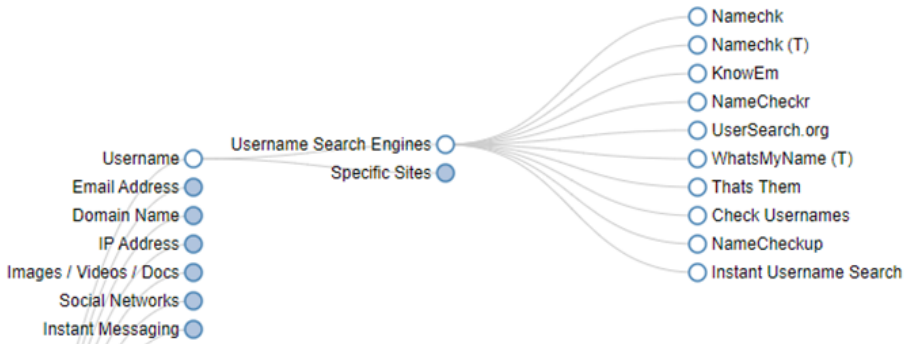

Ilustración 8. *Namechk*

Una vez llegados al enlace final, se apertura la web en cuestión para realizar la búsqueda y en ese momento la investigación queda fuera de ámbito de OSINT *Framework*.

Cabe destacar que, muchas de las herramientas y enlaces que se muestran dentro del árbol categorizado están caídos o han dejado de tener soporte por lo que es muy importante comprobar las herramientas antes de utilizarlas.

Afortunadamente, gracias al gran número de herramientas y enlaces recogidos, existe más de una para realizar la misma función y se dispone de alternativas para seguir con la investigación.

3.13. Web Scraping

Denominamos *Web scraping* [19] a la técnica de recopilar información de modo automático de sitios web mediante programas software que simulan la navegación de un humano, bien sea utilizando el protocolo HTTP o incrustando el navegador en una aplicación.

Básicamente, lo que se consigue es recopilar y estructurar información accediendo al código HTML de las páginas y localizando ciertas palabras claves donde se encontrarán los datos que se necesitan para la acción planificada.

Esta programación suele realizarse Ad hoc para cada web y para cada dato que se quiera consultar.

3.13.1. *Técnicas*

Algunas de las técnicas más importantes utilizadas para realizar este tipo de automatizaciones son las siguientes:

- Copiar y Pegar: Utilizada para trasladar datos de una aplicación o fichero a un formulario web.

- Expresiones Regulares: Consiste en localizar cierta información dentro del código HTML de una web para posteriormente localizar la parte variable del código. Es una técnica de parseo. Esta será la funcionalidad utilizada en nuestro proyecto.

- Protocolo HTTP: Se caracterizan por acceder a información de servidores remotos mediante peticiones HTTP utilizando sockets para emular un navegador y que el servidor

devuelva la información requerida.
• Parsers de HTML: Algunos lenguajes como XQuery se pueden utilizar para parsear documentos, recuperar y transformar el contenido y utilizarlo.

3.13.2. *Casos de uso*

Las aplicaciones de Web scraping suelen ser utilizadas en diferentes ámbitos de automatización. Algunos de ellos son los siguientes:

• Herramientas RPA: Desde hace algunos años, el auge de las herramientas de automatización de escritorio denominadas RPA (*Robot Process Automation*) ha impulsado un gran número de desarrollos de aplicaciones que realizan web scraping en muchas webs corporativas con la finalidad de leer y enviar información entre las diferentes aplicaciones corporativas, lo que provoca un ahorro considerable de tiempo y en última instancia de personal.

• Minería de información: Muchas empresas utilizan estas técnicas para obtener información de un ámbito concreto y generar su propia información.

• Mantenimiento de Histórico Web: Este tipo de técnicas son utilizadas para alimentar las bases de datos de archivo web, pudiendo así realizar una línea temporal de la evolución de las páginas web.

• Rastero de información en general.

3.13.3. *Ventajas*

• Reducción de tiempos de realización de tareas
• Minimización de errores.
• Reducción de costes de personal.

3.13.4. *Inconvenientes*

• Relativa complejidad de desarrollo.
• Intolerancia a cambios en el código.
• En ocasiones, se pueden violar los términos y condiciones de algunos sitios web.

4. ASPECTOS RELEVANTES DEL DESARROLLO DEL PROYECTO

4.1 Elección de fuentes

OSINT *Framework* recopila multitud de sitios y herramientas web para realizar una investigación OSINT eficiente, brindándonos la posibilidad de localizar información en una vasta colección de herramientas.

A la hora de analizar nuestro proyecto, hemos de determinar sobre qué herramientas podemos actuar de forma que podamos realizar las operaciones que necesitamos y la extracción de información sea posible de un modo eficiente.

Tras un primer análisis de los enlaces disponibles en la plataforma, observar las categorías que se presentan y siguiendo el marco de los flujos de trabajo presentados en la obra de Michael Bazzel, hemos determinado diferentes páginas web, que estaban contenidas en el portal de

estudio, según los selectores que se puedan introducir en las mismas, y los cuales determinarán la información a buscar.

Estos selectores han sido asociados a las diferentes categorías en las que OSINT *Framework* agrupa sus herramientas, para así disponer de un listado ordenado que poder trasladar a nuestra aplicación. Se han elegido selectores comunes, para poder realizar de un modo más sencillo las automatizaciones.

4.1.1. *Criterios de Selección*

La selección de las fuentes elegidas para la búsqueda de información se ha basado en los siguientes criterios:

1- Disponibilidad: Se trata de una obviedad, pero desde que se creó OSINT *Framework* muchos sitios a los que intenta enlazar han sido desconectados y existe una gran cantidad de enlaces rotos.

2- Tipología: El sitio web debe ser una herramienta o servicio web. En la plataforma existen enlaces a repositorios de *GitHub* donde descargar la herramienta para su posterior instalación, o lugares de descarga de fabricantes. Este tipo de aplicaciones quedan fuera del alcance de este proyecto en esta primera versión.

3- Diversidad: El total de sitios en los que realizar búsquedas para nuestra investigación debe poder cubrir las categorías de búsquedas necesarias según los selectores elegidos.

4- Freeware: Debe tratarse de herramientas y soluciones gratuitas.

5- Sin Registros: Con el fin de limitar la privacidad según los criterios y consejos de la obra de Michael Bazzel descritos en el apartado "Recomendaciones" del capítulo 3 de este documento, se han buscado lugares en los que no sea necesario registro para obtener información. (Salvo en redes sociales).

6- Tratamiento *anti-scraping* limitado: Para poder llevar a cabo de una forma eficaz la automatización, es necesario que las herramientas utilizadas permitan realizar *scraping* sobre las mismas, así como automatizar algunos eventos para facilitar el trabajo del analista.

7- No uso de CAPTCHA: Por el momento no es posible salvar la utilización de tecnología CAPTCHA dentro de la aplicación. Se considerará como línea de trabajo futura.

4.1.2. *Categorías y Selectores*

- Datos Personales
  - Nombre
  - Usuario
  - Email
  - NIF

- Datos Empresariales
  - Nombre Empresa
  - CIF

- Datos de Red
  - o Dominio
  - o IP

- Vehículos
  - o Matrícula
  - o V.I.N. (Número de bastidor)

- Transporte
  - o Vuelo
  - o Avión
  - o Barco

4.1.3. *Enlaces y Sitios*

Finalmente, tras un exhaustivo análisis de cada uno de los enlaces presentes en la plataforma OSINT *Framework* y evaluando los criterios anteriormente descritos, se han determinado un total de 15 sitios con 39 posibilidades de búsquedas, puesto que en una misma web se puede buscar por diferentes criterios, y de las cuales se puede extraer información de forma automática de 24 combinaciones posibles.

En la siguiente matriz se establece la correspondencia entre el servicio que se solicita, los campos que pueden usarse en dicho servicio y la categoría a la que pertenece cada selector:

		Datos Personales				Datos empresariales		Datos de Red		Vehículo		Transporte		
	URL	Nombre	Usuario	Email	NIF	Epmpresa	CIF	Dominio	IP	Matrícula	VIN	Vuelo	Avión	Barco
Usuarios y correo electrónico	https://namecheckup.com https://haveibeenpwned.com www.google.com	* *	* *	* * *	*	* *	*							
Social Media	https://www.facebook.com/search/top/?q=DireccionEmail https://twitter.com/search?q=ValorBusqueda	* *	* *	* *		* *								
Empleados	https://www.Linkedin.com http://recruitin.net/	* *	* *	* *										
Personas	https://webmii.com/people?n=NombrePersona	*	*	*		*								
Empresas	https://www.axesor.es/buscar/empresas?tabActivo=empresas&q=NombreempresaoCIF					*	*							
Dominios	https://whois.domaintools.com/nombredominio.com https://mxtoolbox.com/SuperTool.aspx?action=mx%3anombredominio.com							* *	*					
Trasnporte	https://flightaware.com/live/flight/NumeroVuelooMatriculaAvion https://www.marinetraffic.com/en/ais/details/ships/imo:IMOBarco https://www.seisenlinea.com/calcular-fecha-matriculacion/ https://www.seisenlinea.com/informe-numero-de-bastidor-coche/?vin=IdentificadorBastidor									* 	 *	*	*	 *

Tabla 10. *Matriz de enlaces*

Nombre	NameCheckUP
URL	https://namecheckup.com
Registro	No Necesario
Búsqueda	Automática
Modo	Inserción de dato y acción buscar
Campo	serachText
Acción	onSearchSubmit()
Botón	searchBtn
Descripción	Site que permite la localización de nombres de usuarios en un conjunto de redes sociales a las que el sistema accede y devuelve la información en forma de baldosas de colores, cuyo significado es el siguiente: - Rojo Claro: Usuario existente - Rojo: Error al acceder - Verde: Usuario Libre

Tabla 11. *NameCheckUP*

Nombre	Haveibeenpwned
URL	https://haveibeenpwned.com
Registro	No Necesario Búsqueda
Búsqueda	Semiautomática
Modo	Inserción de dato automática y acción buscar manual
Campo	Account
Acción	submit()
Botón	searchPwnage
Descripción	Buscador de filtraciones de datos en diferentes sites. Localiza al usuario en bases de datos de filtraciones de información conocidas y documentadas. Si el usuario ha sido filtrado, indicará la plataforma, fecha y tipo de datos que fueron explotados.

Tabla 12. *HaveIBeenPwned*

Nombre	Google URL
URL	https://www.google.com
Registro	No Necesario
Búsqueda	Automática
Modo	Parámetro de URL
Campo	search?q=
Acción	Navigate() Descripción
Descripción	Buscador de información en la web. Ampliamente utilizado. Permite la búsqueda de cualquier dato entre la información de todas las páginas que su robot mantiene indexadas. Debido a que su resultado es impredecible, no es posible automatización de rescate de información.

Tabla 13. *Google*

Nombre	Facebook
URL	https://www.facebook.com
Registro	Necesario
Búsqueda	Automática
Modo	Parámetro de URL
Campo	search/top/?q=ValorABuscar
Acción	Navigate()
Descripción	Red social donde los usuarios realizan interacciones con otros usuarios y publican información. Acceso mediante registro. La búsqueda se realiza de forma automática, pero se ha de navegar hasta la página de información del perfil elegido y desde allí se podrá realizar la extracción.

Tabla 14. *Facebook*

Nombre	Twitter - X
URL	https://twitter.com
Registro	Necesario Búsqueda
Búsqueda	Automática
Modo	Parámetro de URL Campo
Campo	search?q=ValorBusqueda
Acción	Navigate()
Descripción	Red social donde los usuarios realizan interacciones con otros usuarios y publican información. Acceso mediante registro. La búsqueda se realiza de forma automática, pero se ha de navegar hasta la página de información del perfil elegido y desde allí se podrá realizar la extracción.

Tabla 15. *Twitter*

Nombre	LinkedIn URL
URL	https://www.Linkedin.com
Registro	Necesario Búsqueda
Búsqueda	Automática
Modo	Parámetro de URL
Campo	/search/results/all/?keywords= ValorBusqueda
Acción	Navigate()
Descripción	Red social donde los usuarios realizan interacciones con otros usuarios y publican información sobre todo relevante a su currículum y perfil profesional Acceso mediante registro. La búsqueda se realiza de forma automática, pero se ha de navegar hasta la página de información del perfil elegido y desde allí se podrá realizar la extracción de los datos profesionales.

Tabla 16. *LinkedIn*

Nombre	Recruitin URL
URL	http://recruitin.net
Registro	No Necesario
Búsqueda	Manual
Modo	Inserción de Dato y búsqueda manual
Campo	-
Acción	-
Descripción	Compositor de sentencias de búsqueda de información en Google para Linkedin. En base a una serie de parámetros seleccionados compone una sentencia de búsqueda muy precisa de información en Google. Enfocada a búsqueda de perfiles profesionales.

Tabla 17. *Recruitin*

Nombre	Webmii
URL	https://www.webmii.com
Registro	No Necesario Búsqueda
Búsqueda	Automática
Modo	Parámetro de URL
Campo	people?n=NombrePersona
Acción	Navigate()
Descripción	Plataforma de búsqueda de personas. Al trasladar el nombre de un usuario, muestra información localizada en su base de datos, mostrando diferentes enlaces a donde se encuentran los datos. No es posible automatizar extracción debido a la impredecibilidad de la información mostrada.

Tabla 18. *Webmii*

Nombre	Axesor URL
URL	https://www.axesor.es
Registro	No Necesario
Busqueda	Automática
Modo	Parámetro de URL
Campo	buscar/empresas?tabActivo=empresas&q=NombreOCIF
Acción	Navigate()
Descripción	Plataforma de búsqueda de empresas capaz de localizar información a través de CIF o Nombre. Muestra información muy completa, como dirección beneficios, teléfonos de contacto, etc. Ofrece versión de pago más completa.

Tabla. 19. *Axesor*

Nombre	Domain Tools
URL	https://whois.domaintools.com
Registro	No Necesario
Búsqueda	Automática
Modo	Subdominio de URL
Campo	/nombredominio.com
Acción	Navigate()
Descripción	Este sitio web nos proporciona información muy importante de un dominio o IP, capaz de geolocalizar, mostrar su *whois*.

Tabla 20. *Domain Tools*

Nombre	Mx ToolBox URL
URL	https://mxtoolbox.com
Registro	No Necesario
Búsqueda	Automática
Modo	Parámetro de URL Campo
Campo	SuperTool.aspx?action=mx%3a
Acción	Navigate()
Descripción	Sitio web que realiza búsqueda de los MX relacionados con el correo electrónico de un dominio en concreto.

Tabla 21. *MX Toolbox*

Nombre	Seis En Línea (Matrículas) URL
URL	https://www.seisenlinea.com/calcular-fecha-matriculacion
Registro	No Necesario
Búsqueda	Semiautomática
Modo	Inserción de dato automática y acción buscar manual
Campo	matricula_text
Acción	submit()
Descripción	Esta plataforma muestra la información de matriculación de un vehículo, fecha, modelo y posible etiqueta ambiental, recibiendo como valor de búsqueda la matrícula del vehículo.

Tabla 22. *Seis en Línea (Matrículas)*

Nombre	Seis En Línea (VIN)
URL	https://www.seisenlinea.com/informe-numero-de-bastidor-coche
Registro	No Necesario
Búsqueda	Automática
Modo	Parámetro de URL
Campo	/?vin=IdentificadorBastidor
Acción	Navigate()
Descripción.	Esta plataforma muestra la información de matriculación de un vehículo, fecha, modelo, propietarios, histórico de registros y toda la información técnica del vehículo. Además, ofrece la posible letra de matrícula del mismo. El parámetro que debe trasladarse es el identificador de bastidor del automóvil.

Tabla 23. *Seis en Línea (VIN)*

Nombre	Flight Aware
URL	https://flightaware.com/live/flight
Registro	No Necesario
Búsqueda	Automática
Modo	Subdominio de URL Campo
Campo	/IdentificadorVueloOMatriculaAvion
Acción	Navigate()
Descripción	Esta plataforma muestra la información de un vuelo o avión recibiendo como parte de la url el valor a buscar. Ofrece información en tiempo real de vuelos y aviones

Tabla 24. *Flight Aware*

Nombre	Marine Traffic URL
URL	https://www.marinetraffic.com
Registro	No Necesario
Búsqueda	Automática
Modo	Parámetro de URL
Campo	/en/ais/details/ships/imo:
Acción	Navigate()
Descripción	Esta plataforma muestra la información de un barco recibiendo como parámetro el IMO del barco. Ofrece información en tiempo real de la embarcación y el viaje que esté realizando en ese momento.

Tabla 25. *Marine Traffic*

4.2. Arquitectura de la aplicación

La aplicación está realizada en una arquitectura multicapa, empleando el patrón de desarrollo modelo vista-controlador para poder separar de un modo más eficiente y mantenible la lógica de la aplicación de la interfaz. En este desarrollo es especialmente importante utilizar este patrón, puesto que cada acceso a un sitio web está representado por una clase que representa al modelo y además especifica una clase más abstracta que representa a un navegador que es el que se instancia un primer momento.

Es el controlador el encargado de especificar el tipo de clase que se especifica.

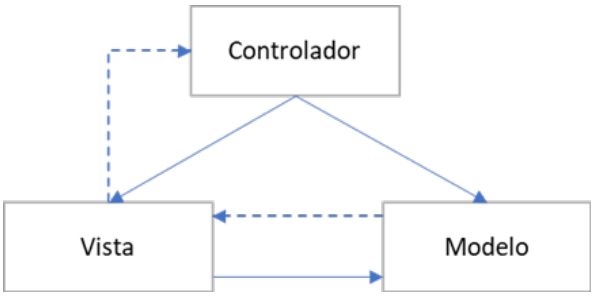

Ilustración 14. *Modelo Vista Controlador*

Hemos de destacar, que la aplicación se ha desarrollado de tal modo que la visual de cada sitio web sobre el que se realiza la búsqueda de información es totalmente dinámico, permitiendo que cada sitio web disponga de su propia clase que específica a una clase más abstracta y que es usada por el controlador para instanciarla y mostrarla en la interfaz.

Observando los diagramas de flujo y de secuencia del funcionamiento general de la aplicación, se comprende mejor la utilidad de este patrón, puesto que observamos como cada proceso asociado a una categoría de selectores, lanza n-objetos de navegación hacia las herramientas seleccionadas:

Ilustración 15. *Diagrama de Flujo*

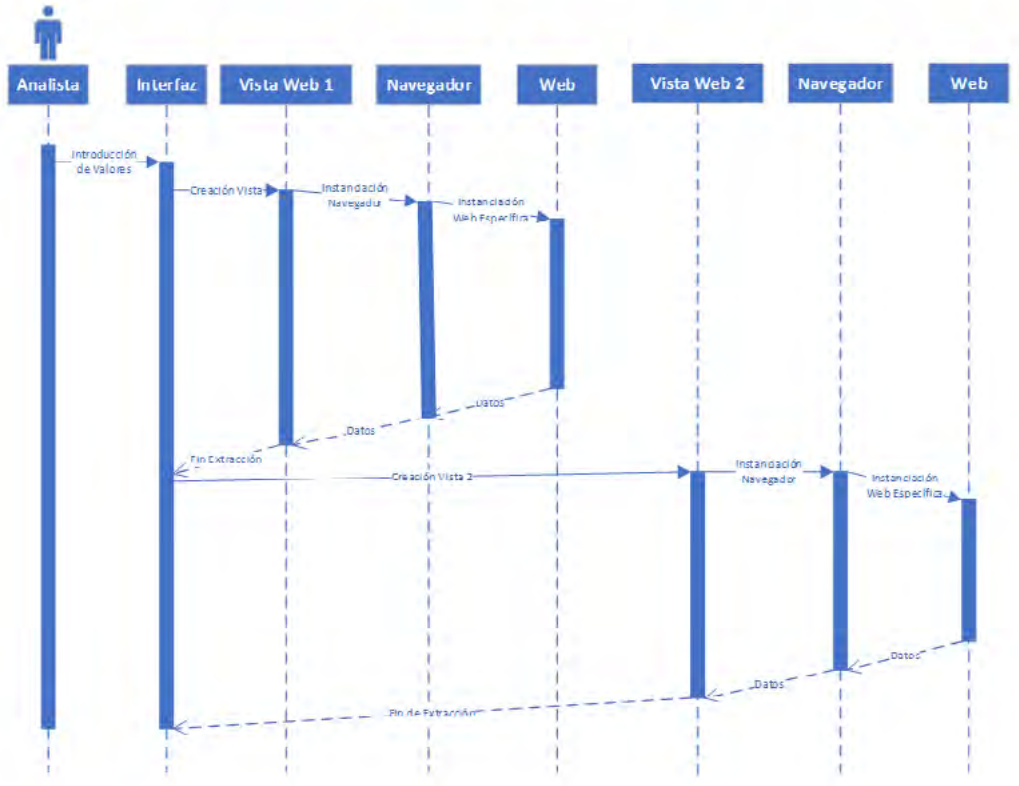

Ilustración 16. *Diagrama de secuencia*

4.3. Uso de *web scraping*

El objetivo para conseguir una automatización eficaz durante el inicio de una investigación OSINT con una herramienta de desarrollo específico, pasa por la utilización de técnicas *web scraping*, es decir, que nuestra aplicación debe ser capaz de leer el código HTML de diferentes sitios web y de interactuar con él para enviar peticiones, consultar y extraer la información que se quiere utilizar.

Esta parte afecta directamente a la capa funcional o lógica de la aplicación y es, en gran medida, el corazón de la misma.

El proceso de implementación de esta tecnología, lleva asociado un gran trabajo de análisis y diseño, puesto que para que todo funcione correctamente, se ha de evaluar cada sitio web que vamos a utilizar y posteriormente determinar si es posible automatizar el envío de información, para minimizar las interacciones del analista y si no es así, solicitar al usuario que ejecute los eventos que se precisan para realizar la búsqueda.

Para poder realizar todo el proceso correctamente, hemos de tener en cuenta tres elementos fundamentales de cada herramienta web que vayamos a utilizar, que pueden ser identificados mediante el analizador de código incorporado en los diferentes navegadores del mercado.

1. Campos de entrada: Normalmente, las páginas dedicadas a la búsqueda de información en fuentes abiertas, disponen de un campo de entrada en el que indicar el valor buscado. Hemos de identificar este valor para que pueda ser usado por el código de nuestra aplicación. Por ejemplo, buscamos identificadores con el nombre “search”. la página.

2. Funciones: Se trata de los eventos que podemos invocar desde nuestra aplicación o llamadas a funciones que podamos utilizar mediante la interacción con el navegador, para ejecutar de forma automática las búsquedas dentro las diferentes plataformas web. En este punto es donde entran en juego las técnicas de bloqueo de *scraping* y en ocasiones han de realizarse estas invocaciones de forma manual, mediante el click del ratón o la pulsación de "Enter" en el teclado. Por ejemplo, el elemento "submit" o la función **onClick()** ambas con una implementación muy extendida.

3. Campos de datos: Esta es la parte más variable del sistema, puesto que cada página muestra la información de un modo diferente y la información que queramos recoger es distinta en cada caso según nuestros objetivos. Se localiza mediante la lectura del código HTML e identificamos clases u objetos del código para poder rescatar los datos de forma precisa.

Una vez la aplicación ha conseguido descargar el resultado de la búsqueda, es cuando se realiza la validación y selección de los datos y se devuelven al controlador de la aplicación para ser mostrados en la interfaz y que el analista determine si son válidos para su investigación, otorgándole la posibilidad de añadirlos al informe.

A continuación, podemos observar en el siguiente diagrama de secuencia la ejecución de esta parte de la aplicación:

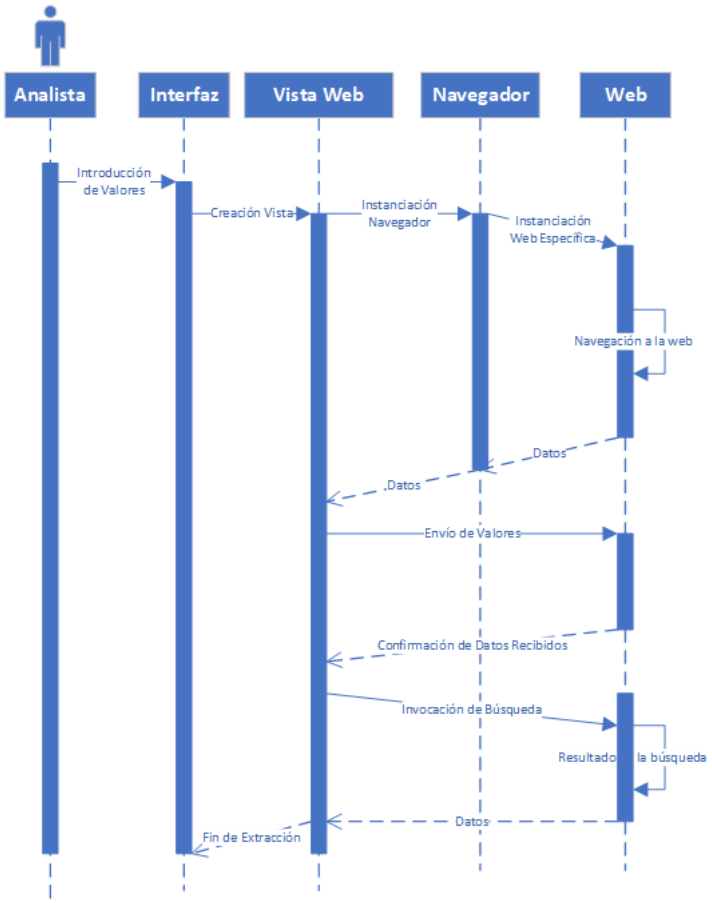

Ilustración 17. *Diagrama de secuencia de Web Scraping*

4.4. Diseño de Interfaz

El diseño de la interfaz viene condicionado por los selectores escogidos en un primer momento, así como las categorías a las que pertenecen de tal forma que se ha optado por una interfaz dinámica, la cual realiza la creación de objetos a medida que se introduce la información de búsqueda y se completan los campos donde se indican los valores de los selectores.

Se ha realizado un diseño de interfaz lo más limpio posible y que otorgue al analista la funcionalidad precisa y necesaria, para que, controlando que no se deja nada por revisar, pueda realizar la base del informe de inteligencia.

La interfaz consta de los campos que representan a los selectores, agrupados en las categorías que hemos definido anteriormente. Dependiendo de qué campos complete el usuario, el sistema realizará la carga automatizada y dinámica de todos los sitios donde se pueden utilizar los selectores y ofrecer una información interesante para el comienzo de la investigación.

Ilustración 18. *OsiNET Agrupación*

Esta primera carga es totalmente automatizada en varios sitios web, pero en otros ha de ser lanzada por el usuario mediante la interacción del mismo sobre la propia web, puesto que los sistemas *anti-scraping* no permiten el uso de *bots*.

Además de páginas donde podamos extraer la información de forma automática, se abrirán enlaces a diversos buscadores de información, cuyos resultados no son predecibles y no es posible automatizarlos, pero sin embargo pueden resultar de gran ayuda navegar por ellos y extraer la información de modo manual.

La visual se muestra de forma dinámica para cada web, en modo de pestañas que están agrupadas en las mismas categorías que los selectores, para que el usuario tenga siempre la perspectiva de qué resultados se han obtenido en base a qué campos.

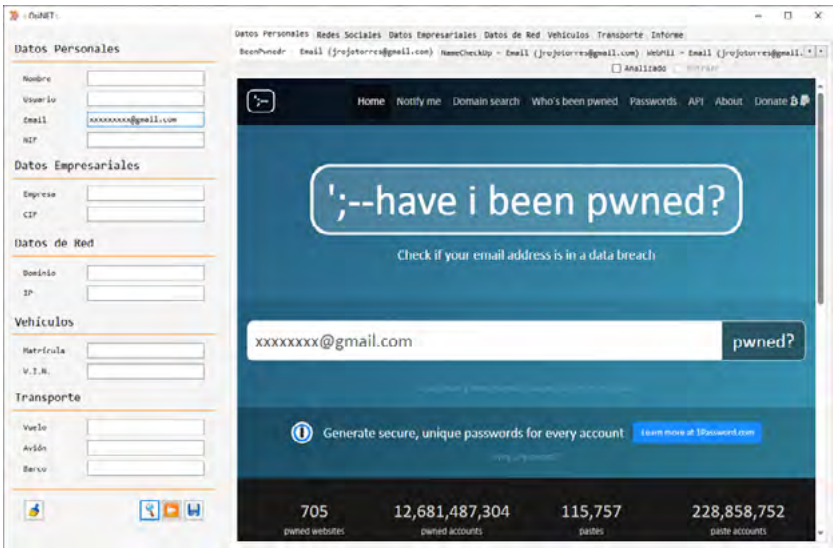

Ilustración 19. *OsiNET Vista Completa*

Una vez se han decidido que campos se van a utilizar y se han completado, la aplicación dispone de un botón para comenzar la acción de cargar la información. Tras la carga, se puede navegar por las pestañas, indicando si queremos extraer la información de la pestaña en la que nos encontramos o no mediante el *check* correspondiente y en caso afirmativo se extraerá la información de esa web tras una validación de la misma.

Siempre hemos de marcar el *check* "Analizado" para indicar al sistema que ya hemos pasado por esa herramienta, puesto que de no hacerlos nos mostrará una alerta para evitar que olvidemos analizar alguno de los sitios.

El *check* "Extraer" sólo se iluminará cuando nos encontremos en la página exacta donde se ha programado la extracción de información y si se marca, el *check* "Analizado" también se marcará automáticamente.

4.5. Presentación de Inteligencia

Nuestra herramienta dispone de la capacidad de creación de un informe de inteligencia en base a los datos que el analista ha dispuesto como seleccionables para mostrar y utilizar.

Como hemos visto anteriormente, este informe se compone de los siguientes apartados:

- Identificación de documento.
- Resumen.
- Objetivos.
- Información recopilada.
- Conclusión.
- Bibliografía.

El documento que genera la aplicación es en formato ".doc" para que sea compatible con la mayoría de procesadores de texto del mercado. Será exportado mediante las librerías que *MS Office* pone a disposición para ser utilizadas por *Visual Studio*, otorgando una gran flexibilidad de desarrollo al ser totalmente compatibles puesto que se trata del mismo fabricante de software.

El documento generado, se centrará en completar la información recopilada como los enlaces de la bibliografía que se han obtenido de forma totalmente automatizada, ahorrando una gran cantidad de tiempo al investigador en su trabajo inicial.

5. CONCLUSIONES Y LÍNEAS DE TRABAJO FUTURAS

Tras la realización de este proyecto y evaluado su potencial, podemos concluir que la creación de programas específicos para integrar herramientas OSINT pueden ser muy fructíferos y ayudar en gran medida a todos los implicados en una investigación. Este tipo de aplicaciones facilita el inicio de cualquier investigación, sirviendo como marco de trabajo y guía de ejecución de los ciclos de inteligencia y de los flujos de trabajo a seguir que garantiza el éxito del proceso.

Además, permite reducir los tiempos de localización de la información, optimizar costes y minimizar al máximo los errores que puedan producirse durante el tratamiento de los datos, al seleccionar y trasladar aquellos que precisamos. Recordemos que en muchas ocasiones la información que se extrae al terminar un proceso o selección, sirve como entrada para el siguiente proceso a medida que avanzamos y profundizamos para llegar a nuestro objetivo, es decir, la validez y exactitud de un dato nos condicionará el resultado del siguiente proceso de búsqueda y con esta aplicación minimizamos la posibilidad de error.

Hemos de resaltar que nuestra aplicación depende en gran medida de aplicaciones web de terceros, las cuales pueden estar sujetas a disponibilidad y cambios que escapan a nuestra voluntad y control, por lo tanto, el riesgo técnico más importante es la dependencia de la disponibilidad y estabilidad de las soluciones que se consultan dentro de la aplicación.

Finalmente, como líneas de trabajo futuras, podemos determinar los siguientes puntos:

• Revisión de los enlaces y mantenimiento de las búsquedas de datos dentro de las herramientas web.

• Mejoras de formato en el informe de inteligencia.

• Creación y exportación de *datasets* con la información obtenida para su exportación y posible tratamiento en otras herramientas.

• Añadir más herramientas y sitios web al sistema.

• Adaptar el sistema para que pueda comunicarse mediante API con los diferentes sitios de búsqueda en los que esta tecnología se permita.

• Almacenar información para retomar investigaciones.

• Posibilidad de ampliar sitios con protección *anti-scraping* mediante tecnología de resolución de CAPTCHA y otras técnicas capaces de eludir la protección sobre la obtención de información automatizada. Existen productos que pueden integrarse en el código de una aplicación como, por ejemplo, la solución *Selenium*.

6. BIBLIOGRAFÍA

[1] Gutiérrez, J. (2021): *Metodología OSINT para investigar en Internet*. Ediciones Ciberpatrulla.

[2] Gutiérrez, J. (2018): T*écnicas OSINT para investigación en Internet. Manual para investigadores*. Ediciones Ciberpatrulla.

[3] Perrino Navas, I. (2022): *Investigación y extracción de datos con técnicas OSINT*. Madrid: Universidad politécnica de Madrid.

[4] Cabello Gallego, A. (2021): *Herramientas y técnicas OSINT para la extracción y el análisis de la información procedente de fuentes abiertas*. Alcalá de Henares: Universidad de Alcalá.

[5] Brezo Fernández, F. y Rubio Viñuela, Y. (2019): *Manual de Ciberinvestigación en Fuentes Abiertas: OSINT para Analistas*. Independently Published.

[6] Richard J. Heuer Jr, Randolph H. Pherson (2015): *Técnicas analíticas estructuradas para el análisis de inteligencia*. Murcia: Plaza y Valdés Editores.

[7] Gutierrez, J.:"Qué es OSINT", [en línea] <https://ciberpatrulla.com/que-es-osint/>.

[8] Martinez, A.: "OSINT – La información es poder", [en línea] < https://www.incibe.es/incibe-cert/blog/osint-la-informacion-es-poder >

[9] Wikipedia. Varios Autores: "Inteligencia de fuentes abiertas" [en línea] < https://es.wikipedia.org/wiki/Inteligencia_de_fuentes_abiertas >

[10] Microsoft: "Microsoft Edge WebView2" [en línea] < https://learn.microsoft.com/es-es/microsoft-edge/webview2/ >

[11] OSINT Framework, [en línea], < https://osintframework.com/ >

[12] CNI-CERT (2015): "Guía de seguridad (CCN-STIC-425) Ciclo de inteligencias y análisis de intrusiones". Madrid, Centro Criptológico Nacional.

[13] Gobierno de España (2002), *Ley 34/2002, de 11 de julio, de servicios de la sociedad de la información y de comercio electrónico*.

[14] Gobierno de España (2007), *Ley 25/2007, de 18 de octubre, de conservación de datos relativos a las comunicaciones electrónicas y a las redes públicas de comunicaciones*.

[15] Gobierno de España (1882), Real Decreto de 14 de septiembre, *Ley de Enjuiciamiento Criminal.*

[16] Gobierno de España (2015), *Ley Orgánica 13/2015, de 5 de octubre, de modificación de la Ley de Enjuiciamiento Criminal para el fortalecimiento de las garantías procesales y la regulación de las medidas de investigación tecnológica*.

[17] Gobierno de España (1995): *Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal*

[18] Gobierno de España (1978): *Constitución Española*.

[19] Wikipedia. Varios Autores: "Web Scraping" [en línea] < https://es.wikipedia.org/wiki/Web_scraping >

[20] Gutiérrez J.: "Qué es y cómo puedes utilizar sus recursos para acelerar tus investigaciones", [en línea], < https://ciberpatrulla.com/osint-framework/ >

[21] Michael Bazzell: [en línea] < https://www.linkedin.com/in/michael-bazzell-a83572122/ >

[22] Wikipedia. Varios Autores: "Mr. Robot", [en línea], < https://es.wikipedia.org/wiki/Mr._Robot >

[23] Odint.Net (2022): "Los 7 mejores libros sobre OSINT", [en línea] < https://odint.net/libros-osint/ >

[24] Emiliano Piscitelli (2021): "OSINT para el bien y para el mal", [en línea] < https://blog.smartfense.com/2021/06/osint-para-el-bien-y-para-el-mal.html >

[25] OSINTeame (2019): "OSINT (Open Source Intelligence) aplicado a las investigaciones" , [en línea], < https://medium.com/@osinteame/osint-open-source-intelligence-aplicado-a-las-investigaciones-10d2599f4296 >

[26] Michael Bazzell (2023), *OSINT Techniques: Resources for Uncovering Online Information*, Independently published.