Universidad de Alcalá
Escuela Politécnica Superior

GRADO EN INGENIERÍA INFORMÁTICA

Trabajo Fin de Grado

Investigación y extracción de datos en fuentes abiertas

Autor: David Galindo Blanco

Tutor/es: Manuel Sánchez Rubio

2018

UNIVERSIDAD DE ALCALÁ

Escuela Politécnica Superior

Grado en Ingeniería Informática

Trabajo Fin de Grado

Investigación y extracción de datos en fuentes abiertas

Autor: David Galindo Blanco

Tutor/es: Manuel Sánchez Rubio

TRIBUNAL:

Presidente:

Vocal 1º:

Vocal 2º:

FECHA:

Agradecimientos

Quiero dar las gracias a mi familia, en especial a mis padres, hermano y abuela, por todo el apoyo que me han dado, no solo en la realización de este proyecto, si no a lo largo de estos años de estudios, y a los que espero devolverles todos sus esfuerzos con la realización de este trabajo y la conclusión de esta carrera.

Agradecer también a mis compañeros, que considero ya grandes amigos, con los que he compartido buenos y malos momentos, y que me han ayudado a llegar hasta aquí.

Por último, dar las gracias a la Universidad de Alcalá y a su profesorado, en especial a mi tutor, Manuel Sánchez Rubio, por toda la ayuda que me ha prestado y darme la oportunidad de realizar este trabajo.

Índice

Índice de figuras

1. Sumario

En el mundo actual, Internet se ha convertido en la mayor fuente de información existente. Toda la actividad que realizamos en la red genera una huella digital, con todo tipo de información sobre nosotros o nuestro entorno, y que está disponible de forma pública.

A lo largo de este documento, estudiaremos las distintas fuentes en las que podremos localizar esta información, así como algunas de las principales herramientas a la hora de realizar la extracción de estos datos, para su posterior análisis y generación de inteligencia.

2. Summary

In today's world, the Internet has become the largest source of existing information. All the activity that we do in the network generates a fingerprint, with all kinds of information about us or our environment, and that is available openly. Throughout this document, we will study the different sources in which we will be able to locate this information, as well as some of the main tools at the time of extracting these data, for their later analysis and generation of intelligence.

3. Palabras clave

Internet, información, inteligencia, redes sociales, privacidad, motores de búsqueda.

4. Resumen

El concepto de OSINT (Inteligencia en fuentes abiertas) está cobrando cada vez más importancia dentro del campo de la ciberseguridad. Se trata de un proceso que consiste en generar inteligencia (entendiendo inteligencia como el análisis y procesamiento de los datos para que puedan ser correctamente utilizados) a partir de la información extraída de distintas fuentes públicas en Internet o en otros medios tradicionales para la investigación de un objetivo.

Cuando hablamos de estas fuentes abiertas, nos referimos a aquellos puntos clave dentro de Internet donde podemos encontrar información acerca de personas, lugares o instituciones de forma pública, por lo que su acceso y recolección se encuentra dentro del marco legal.

Entre las distintas fuentes disponibles, destacan los motores de búsqueda, ya que son los encargados de indexar gran parte de la información disponible en la red; las redes sociales, que debido al exceso de confianza de la gente a la hora de publicar aspectos de su vida privada se han convertido en uno de los principales puntos de acceso a datos críticos sobre los posibles objetivos; y los distintos datos tanto de carácter personal como corporativo que podemos obtener a través de diferentes medios, y a partir de los cuales podemos realizar diferentes tipos de búsqueda para obtener aún más información.

Para cada una de estas fuentes, estudiaremos las herramientas más utilizadas y las distintas formas de aprovechar sus recursos, que nos pueden ayudar a extraer la mayor cantidad posible de datos, con el fin de realizar informes de inteligencia más completos.

5. Introducción

5.1. Planteamiento

En este Trabajo de Fin de Grado (TFG), se estudiará el proceso de investigación y extracción de toda la información que podemos obtener a través de fuentes abiertas, fruto de la actividad que realizan las personas en Internet, lo que forma parte de un proceso mayor conocido en el mundo actual de la seguridad de la información como inteligencia en fuentes abiertas u OSINT.

Empezaremos contextualizando la aparición de este tipo de investigaciones a raíz del crecimiento que ha experimentado Internet en los últimos años dentro de nuestra sociedad, y como en él vertimos todo tipo de información, generando así una huella digital con todo lo que se puede encontrar en la red sobre nosotros.

Posteriormente, comentaremos el proceso de OSINT. Este proceso, se apoya en la existencia de fuentes abiertas y en las herramientas disponibles para obtener datos de estas para ser usados en un contexto de inteligencia. Esta generación de inteligencia está formada por una serie de fases que deberemos seguir para poder presentar toda la información obtenida de la forma adecuada y que facilite su posterior uso. Hablaremos también de las principales fuentes donde podremos buscar datos de distintos tipos de objetivo que queramos investigar y nos centraremos en el análisis de algunas de las herramientas más útiles o importantes que han sido desarrolladas por distintas organizaciones o personas a lo largo de los años y que puedan ser accedidas de forma libre y gratuita.

Para realizar este análisis de las distintas herramientas OSINT, clasificaremos estas herramientas en función de las fuentes que se utilicen, como motores de búsqueda, redes sociales, datos personales y datos corporativos, comentando también los principales frameworks que facilitan la tarea de extracción, ofreciendo una amplia visión de las fuentes disponibles y las herramientas que disponemos para cada una de ellas.

5.2. Objetivos

El objetivo por alcanzar en este Trabajo de Fin de Grado es el de reunir la información disponible en libros o artículos acerca de la extracción de datos en fuentes abiertas. Se pretende así, mostrar la información más relevante del tema y presentarla de una forma adecuada para que pueda utilizarse como material de referencia para aquellas personas que estén interesadas en el estudio de este tema, el cual cada vez va cobrando más importancia en el mundo de la seguridad de la información, ya que cada vez son más las personas e instituciones interesadas en el manejo de estos datos.

El trabajo pretende no solo hablar sobre las distintas técnicas de obtención de información, si no también poder concienciar a todas aquellas personas que utilizan Internet del riesgo al que se exponen al no realizar un uso seguro y consciente del mismo.

A nivel personal, el objetivo del desarrollo de este proyecto será el de poner en práctica los conocimientos y competencias adquiridos a lo largo de la carrera, y ampliarlos en el ámbito específico del tema tratado.

6. Información en Internet

Hoy en día, la proliferación del uso de Internet con cerca de 2.500 millones de usuarios en todo el mundo y la facilidad que este ofrece para la publicación de contenido a través del uso cada vez más extendido de redes sociales, webs, blogs u otros medios, han favorecido a la existencia de una gran cantidad de información online, ya sea de carácter personal, empresarial, etc.

Esto puede ser fácilmente demostrado analizando la cantidad de información que maneja por ejemplo Google, con alrededor de 30 billones de páginas web, lo que

supone más de 1000 terabytes de información, u observando el crecimiento que han tenido las redes sociales como Facebook que pasó de tener 1 millón de usuarios en su nacimiento en 2004 a los más de 2 billones de usuarios con los que cuenta en la actualidad.

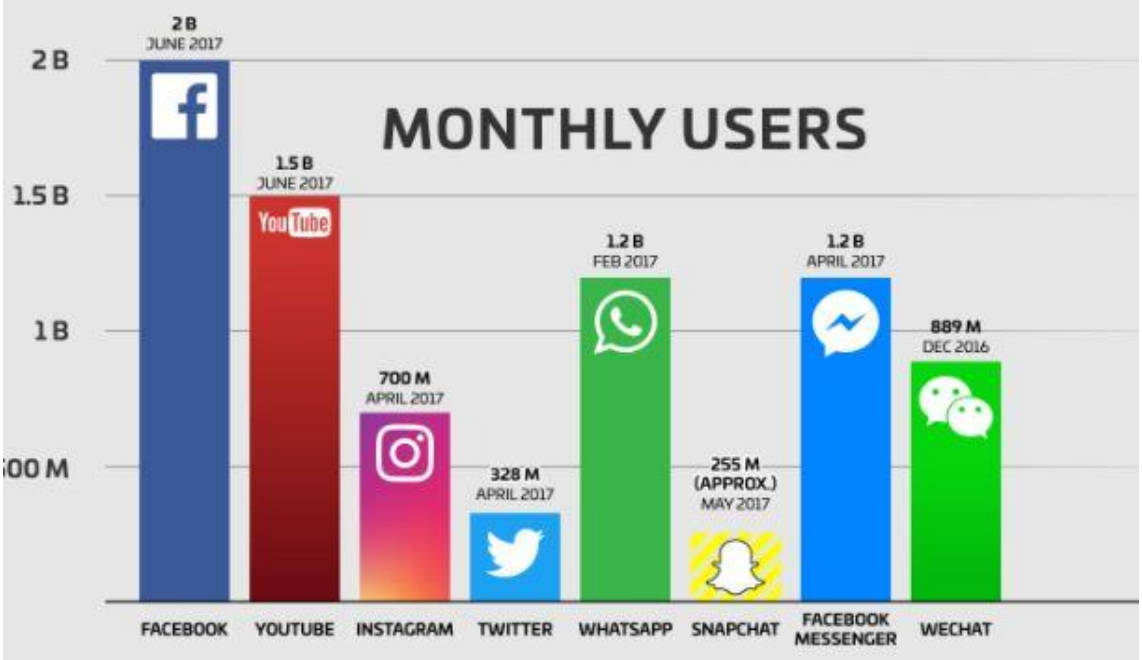

Figura 1: Usuarios mensuales activos en redes sociales (2017)

Para proveer aún más contexto, el antiguo CEO de Google, Eric Schmidt, decía estas palabras en el año 2010:

"*Cada dos días, creamos tanta información como lo que hicimos desde los inicios de la civilización hasta el año 2003 … Esto es algo así como cinco exabytes de datos*".

Sin embargo, estos datos representan la superficie de la red. No podemos dejar de mencionar también la cantidad de información que encontramos en la Deep Web, ya que, aunque no existen cifras exactas, su volumen se estima incluso más extenso. En ella encontramos toda aquella información que no ha sido indexada por los motores de búsqueda o que solo puede ser accedida mediante el uso de software especializado, metodologías o permisos especiales.

Figura 2: La parte oculta de Internet

6.1. Huella digital

Estas cifras permiten hacernos una idea de la enorme cantidad de datos disponibles en la red. Con la evolución de las tecnologías de la información y comunicación, cualquier persona puede utilizar las funciones que nos ofrece Internet para realizar casi cualquier tipo de actividad.

En la actualidad, nuestra vida esta tan adaptada a los entornos digitales que no reparamos en el torrente de información personal que dejamos en las redes sociales o al navegar por Internet. Es, por ejemplo, muy común entre muchos usuarios de las redes sociales, publicar fotos, videos o comentarios a nuestros perfiles que pueden revelar mucho de nosotros como donde vivimos o con quien nos juntamos y que pueden suponer un peligro ya que no sabemos las intenciones de las personas que pueden acceder a esta información.

Algunas de las formas en las que dejamos nuestra huella digital pueden ser:

- **Sitios web y compras en línea**: Actualmente está muy extendido en este tipo de sitios el uso de cookies que consisten en una pequeña información enviada desde estas webs a los navegadores y que les permiten consultar nuestros hábitos de navegación, para el denominado spyware, que utilizan las agencias de publicidad para la entrega de anuncios dirigidos que nos muestran productos sobre los cuales hemos estado leyendo.
- **Redes sociales**: Todos aquellos retweets o comentarios en Facebook dejan un registro en Internet. Por ello, es importante conocer cuáles son las configuraciones de privacidad por defecto de nuestras cuentas y estar pendiente de las mismas. A menudo, estos sitios introducen nuevas configuraciones y políticas de privacidad que pueden aumentar la visibilidad de nuestros datos, confiando en que el usuario aceptará todos los términos que se están introduciendo, sin pararse siquiera a leerlos, o que en muchos casos no tendrán los conocimientos necesarios para entenderlos.
- **Ordenadores, teléfonos o tablets**: Algunos sitios web pueden generar listados con los diferentes dispositivos que utilizamos para acceder a los mismos. Aunque a veces esto puede ser usado para proteger nuestras cuentas, es importante conocer que datos recogen sobre nuestros hábitos. A su vez, estos dispositivos tienen activada la opción de ubicación o geolocalización, que puede permitir conocer a estos sitios web o cualquier persona donde nos encontramos en cada momento.

Todas estas acciones que realizamos crean un rastro en la red que es lo que conocemos como nuestra huella digital, y a la que se puede acceder fácilmente a través de las herramientas que nos ofrece OSINT.

Figura 3: Huella digital

7. Inteligencia en fuentes abiertas (OSINT)

7.1. Concepto

OSINT es un acrónimo anglosajón el cual se refiere a Open Source Intelligence, que se traduce a nuestro idioma como Inteligencia en Fuentes Abiertas, y se considera un tipo de inteligencia ya que se estructura a partir del procesamiento y análisis de la información obtenida de forma legal y ética de diferentes fuentes abiertas o públicas.

A parte de OSINT, existen múltiples disciplinas de recolección de inteligencia:

- SIGINT (Signals Intelligence): Inteligencia a partir de la intercepción de señales.
- GEOINT (Geospatial Intelligence): Inteligencia obtenida por medio de la geolocalización.
- HUMINT (Human Intelligence): Inteligencia adquirida por individuos, que interactúan con otras personas por medio de redes sociales y otros canales de comunicación.

- Otras como ELINT (Electronic Intelligence), FININT (Financial Intelligence), IMINT (Imagery Intelligence), COMINT (Communication intelligence).

El concepto de OSINT, bajo un nombre u otro, ha estado a nuestro alrededor cientos de años. A pesar de que con la llegada de las nuevas tecnologías de la información y comunicación hayamos centrado en Internet las posibles fuentes de recopilación de información, la extracción de datos puede ser realizada a partir de cualquier material público disponible como:

- Medios de comunicación: periódicos, revistas, radio y televisión de cualquier región del mundo.
- Internet, publicaciones online, blogs, grupos de discusión, medios ciudadanos (por ejemplo, videos u otros contenidos creados por los usuarios), Youtube y otras redes sociales (como Facebook, Twitter, Instagram, etc.).
- Datos públicos del gobierno (como los Boletines Oficiales del Estado), informes gubernamentales, presupuestos, audiencias, directorios de teléfono, conferencias de prensa, sitios web o discursos. A pesar de que estas fuentes provengan de fuentes oficiales, son de acceso público, y pueden ser usadas abiertamente y gratuitamente.
- Publicaciones académicas y profesionales, información adquiridas de revistas de artículos, simposios, conferencias, tesis, etc.
- Datos comerciales, imágenes comerciales, evaluaciones financieras e industriales y bases de datos.
- Literatura gris, informes técnicos, preimpresiones, patentes, documentos de trabajo, documentos comerciales, trabajos no publicados y boletines informativos.

Sin embargo, lo más importante a tener en cuenta es que OSINT no consiste en un trabajo de investigación en el que obtenemos diversos fragmentos de información acerca de un objetivo, si no que consiste en un proceso de

inteligencia para crear un conocimiento personalizado para dicho objetivo, que puede ser un individuo o grupo específico.

7.2. Importancia de OSINT

En la actualidad, OSINT se encuentra en todas partes. Los distintos gobiernos que basan sus procesos de inteligencia en su habilidad para adquirir todo tipo de datos tienen su propio uso de OSINT. Sin embargo, al final del día, todos usamos OSINT: cuando realizamos búsquedas en Internet para comparar distintos tipos de productos que queremos comprar o para encontrar alguna persona a la que queremos conocer, estamos básicamente adquiriendo y seleccionando datos de fuentes abiertas.

Tal es la importancia de la inteligencia en fuentes abiertas, que cada vez más organizaciones están desarrollando sus propias estrategias de OSINT. Las correctas herramientas, combinadas con las habilidades de equipos de profesionales dedicados a la búsqueda de información, puede ayudar a estas organizaciones de forma cada vez más efectiva, proporcionando apoyo en la disminución de riesgos gracias al conocimiento estadístico y predictivo del análisis de grandes volúmenes de información.

Todo esto hace que OSINT cobre cada vez más importancia dentro del ámbito de la ciberinteligencia. La ciberinteligencia permite, a partir de la adquisición y análisis de la información, identificar, rastrear, predecir y contrarrestar las capacidades, intenciones y actividades de los atacantes, y ofrecer cursos de acción con base en el contexto particular de la organización, que mejoren la toma de decisiones. A parte de proveer de dicho apoyo a las organizaciones, permiten mantener la seguridad de todos los ciudadanos gracias a su uso en varios frentes como la lucha contra el terrorismo o la persecución de distintos tipos de delincuentes informáticos.

Para resumir, algunos de los ejemplos de la utilización de OSINT pueden ser los siguientes:

- Conocer la reputación online de un usuario o empresa
- Realizar estudios sociológicos, psicológicos, lingüísticos, etc.
- Auditorías de empresas y otros tipos de organismos con el fin de evaluar el nivel de privacidad y seguridad.
- Evaluación de las distintas tendencias de los mercados.
- Identificación y prevención de posibles amenazas en el ámbito militar o la seguridad nacional.
- Por otro lado, también puede tener usos negativos, ya que puede ser utilizado por ciberdelincuentes para lanzar distintos tipos de ataques contra organizaciones o personas.

7.3. Fases de inteligencia

Como hemos comentado anteriormente, la extracción de datos en fuentes abiertas forma parte de un proceso aún mayor que utiliza estos datos extraídos para generar inteligencia.

Para empezar, tenemos que tener en cuenta los problemas que pueden surgir en el desarrollo de este proceso.

Algunos de estos problemas pueden ser:

- **Demasiada información**: como ya se ha puesto de manifiesto, la cantidad de información pública disponible en Internet es más que notable. Es por ello, que se debe realizar un proceso exhaustivo para identificar y seleccionar las fuentes de información más importantes que van a ser recopiladas y que servirán en la generación de inteligencia.

- **Fiabilidad de las fuentes**: es importante valorar previamente las fuentes de las que se va a realizar la recopilación, ya que una selección errónea de las mismas puede provocar desinformación o resultados equivocados.

Por todo esto, se recomiendan una serie de fases o pasos a seguir para el desarrollo de este proceso y que se basa en el denominado ciclo de inteligencia.

Las fases de este ciclo son las siguientes:

Figura 4: Ciclo de inteligencia

- **Requisitos**: Establecer los requerimientos que se quieren cumplir. Esto es, saber los objetivos que se desean obtener, la información que se quiere tener y el tiempo que se va a necesitar.
- **Fuentes de información**: Encontrar las fuentes de información más relevantes que serán utilizadas para obtener la información. Además, se deberá realizar una planificación, definiendo cuál será la estrategia para la recolección de información, el tipo de información y el

contenido, definiendo y clasificando la disponibilidad y fiabilidad de las fuentes y los flujos de la comunicación.

- **Adquisición**: Consiste en conseguir la información de las diversas fuentes de información públicas que se han identificado en la fase anterior, es decir, obtener la información en bruto. Cuanto mayor cantidad de información consigamos mejor, pero siempre debemos tener en cuenta los distintos atributos relacionados con esta información como su contexto, fiabilidad de las fuentes, integridad, fecha, etc. Estos atributos serán importantes en el desarrollo de las siguientes fases. Además, la extracción de datos debe realizarse bajo un marco legal, ya que de lo contrario se podría anular la eficacia de los resultados. Para ello, nos serviremos de las distintas herramientas disponibles y que veremos más adelante en este trabajo.
- **Procesamiento**: En esta fase se procesará la información conseguida para proveerla de un formato de manera que posteriormente pueda ser analizada.
- **Análisis**: Se genera inteligencia a partir de los datos recopilados y procesados. En esta fase se analizará la información obtenida, y que, tras depurarla, tratarla y procesarla, se eliminará aquella que sea inservible debido a que carezca del suficiente valor, sea errónea, o no sea lo suficientemente veraz para incluirla. Para ello, se necesita de un equipo de personas que clasifiquen la información en función de los atributos asociados a la fiabilidad de la fuente, fiabilidad de la información, validez de los datos, pertinencia, relevancia y utilidad.
- **Inteligencia**: Presentar la información conseguida de una manera eficaz, potencialmente útil y comprensible, gracias a informes detallados con diagramas, tablas o figuras para que se puedan sacar las conclusiones pertinentes sobre dicha información.

7.4. Fuentes abiertas

Las fuentes de información que se utilizan en un proceso OSINT pueden ser de diferentes tipos dependiendo de los objetivos que se desean obtener, y a partir de las cuales han sido desarrolladas diversas herramientas para su explotación.

A pesar de que se pueden realizar varios tipos de clasificación, ya que hay herramientas que extraen información de varios tipos de fuentes diferentes, clasificaremos las distintas herramientas en tres fuentes principales. Los distintos tipos de fuentes pueden ser:

- **Motores de búsqueda**: Una de las principales fuentes de información se encuentra en los motores de búsqueda. A parte de las posibles búsquedas que podemos realizar usando algunos de estos motores como Google, Bing, Yahoo o DuckDuckGo, podríamos incluir aquí una gran cantidad de metabuscadores o buscadores personalizados. Una de las principales características de algunos de estos motores es la capacidad que nos da para realizar búsquedas parametrizadas a través de los denominados dorks, lo que se conoce como Google o Bing Hacking.
- **Redes sociales**: Una de las fuentes más importantes para encontrar información sobre un objetivo son las redes sociales. El número de usuarios en las redes sociales crece y crece cada día más y muchos de nosotros solemos publicar muchos datos sobre nosotros mismos o nuestras vidas, los cuales poseen una veracidad extra al ser mostrados por nosotros mismos. Existen varias herramientas que nos permiten detectar la presencia de cualquier persona en estas redes sociales y muchas de estas tienen APIs de uso libre que pueden ser utilizados para la creación de buscadores avanzados.

- **Datos de carácter personal**: Este tipo de datos que podemos encontrar en la red incluyen información muy concreta y precisa sobre nosotros mismos, tales como nombre, edad, lugar de residencia, lugar de trabajo, teléfono, etc. Existen varios buscadores especializados que con solo uno de estos datos pueden obtener toda la información que existe en Internet sobre nuestra persona.
- **Datos corporativos**: También podemos encontrar datos generales sobre las empresas como teléfonos, correos, información sobre empleados u otros datos más avanzados como puede ser la propia infraestructura de red y sistemas a través de aplicaciones como Maltego.

Dependiendo del objetivo que establezcamos en la fase de requerimiento inicial, deberemos elegir entre utilizar una fuente u otra y explotar los datos de unas o de otras.

8. Herramientas

Una vez que conocemos el proceso de generación de inteligencia, nos centraremos en la fase de extracción de datos para analizar las herramientas que tenemos a nuestra disposición para realizar esta función.

La lista de las herramientas que podemos encontrarnos es enorme, ya que existen herramientas para buscar y analizar todo tipo de información. A esto le podemos sumar el hecho de que cualquiera que cuente con los conocimientos necesarios puede desarrollar scripts o aplicaciones para realizar búsquedas avanzadas en cualquier fuente.

Por todo esto, y debido a la existencia de herramientas de pago, en este trabajo nos centraremos en aquellas herramientas que se consideran más útiles

o importantes (y que sean libres de utilizar para cualquier persona) que nos ayuden a extraer información de cada una de las fuentes que hemos estudiado anteriormente y que estén disponibles para todo el mundo.

8.1. Motores de búsqueda

Los motores de búsqueda surgieron a principios de los 90 debido a la necesidad de organizar, clasificar y gestionar la información de Internet ya que cada vez surgían más y más nuevos sitios web llenos de contenido. Estos motores realizan una exploración permanente de Internet, indexando toda la información que encuentran, es decir, crean un índice propio de todo el contenido que son capaces de rastrear. Cada vez que estamos realizando una búsqueda en sitios como Google o Yahoo, estos consultan en su índice con el fin de entregar el resultado que consideran mejor. El motivo de esta indexación tiene que ver con la capacidad de reacción. El hecho de disponer de índice propio permite a los motores de búsqueda dar una respuesta rápida al usuario.

El simple hecho de realizar búsquedas en Google, Yahoo, Bing o DuckDuckGo, supone en sí una de las herramientas más útiles de extraer información en Internet. Sin embargo, muchos de estos navegadores nos ofrecen una serie de operadores avanzados que nos permiten realizar búsquedas parametrizadas para acceder a aquella información que es más difícil de localizar.

Figura 5: Motores de búsqueda

8.1.1. Google/Bing Hacking

Este tipo de hacking es una técnica que utiliza operadores avanzados para filtrar información en buscadores como Google, Bing o en otros motores de búsqueda. Estos operadores son conocidos como dorks y existen una gran cantidad de ellos en función de la información que queremos filtrar.

Estos parámetros especializados pueden dividirse en:

- **Operadores booleanos**: Consiste en el uso de operadores y símbolos para realizar búsquedas combinadas. Algunos de estos son:
  - " ": El introducir una expresión entre comillas, nos permitirá buscar esta expresión literalmente, mostrándote todos los resultados en los que aparece exactamente la expresión introducida.
  - -: El símbolo de "menos" nos permitirá excluir de los resultados obtenidos aquellas páginas que incluyan el término o palabra introducida justo después de este símbolo.
  - +: Al contrario que con el símbolo -, el uso de este operador permitirá realizar búsquedas que incluyan el término que se encuentra después del símbolo.

- or y and: Busca páginas que contengan un término u otro, o un término y el otro, buscando ambos al mismo tiempo.
- *: Actúa como comodín, pudiendo remplazar cualquier palabra.
- . : Al igual que es asterisco, actúa como comodín para una o muchas palabras.

- **Comandos**: Una vez que conocemos estos operadores, podemos combinarlos con los diferentes dorks o comandos que nos ofrecen los motores de búsqueda para encontrar la información que deseemos. Los principales dorks son:
  - *intitle*, *allintitle*: Búsqueda por el título de la página.
  - *inurl*, *allinurl*: Búsqueda por la URL.
  - *filetype*: Este comando sirve para buscar archivos con la extensión que se desee (docs, pdf, etc.)
  - *allintext*: Búsqueda de una cadena dentro del contenido de una página.
  - *site*: Solo busca resultados que provengan del dominio introducido.
  - *link*: Busca páginas que tienen un link a determinada web.
  - *inanchor*: Solo busca páginas que tienen en un texto de enlace el término introducido.
  - *daterange*: Búsqueda por rango de tiempo.
  - *cache*: Muestra el resultado de la cache.
  - *info*: Búsqueda de información de un dominio web.
  - *related*: Devuelve resultados de sitios web relacionados.
  - *insubject*: Búsqueda de páginas por tema.
  - *define*: Búsqueda de definiciones.

- *autor*: Permite realizar búsquedas de obras o artículos por autor.
- *group*: Búsqueda de nombres de Google Groups.

Todos estos comandos pueden ser combinados para obtener resultados mucho más precisos. Con estos dorks podemos obtener todo tipo de información que ha sido indexada por los motores de búsqueda pero que resulta difícil de localizar. Entre los distintos tipos de información que podemos obtener encontramos:

- **Ficheros con usuarios y contraseñas**: Con el uso de estos operadores avanzados podemos encontrar todo tipo de archivos de distintas organizaciones que contienen datos sobre sus usuarios y que en muchos casos incluyen contraseñas. Entre las búsquedas que podemos realizar, por ejemplo, con solo introducir el comando *intitle: "index of" "Index of/" password.txt*, podemos obtener enlaces a directorios dentro de servidores en los que se guardan ficheros de texto con las cuentas de los usuarios.

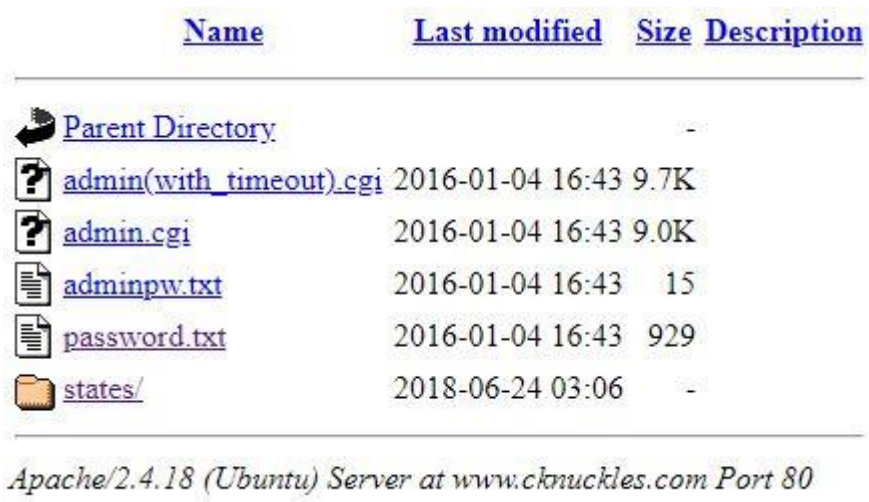

Figura 6: Directorios con ficheros y contraseñas

Una vez que hemos accedido al servidor podemos observar que contiene un fichero de texto denominado password.txt en el que si entramos nos encontraremos con varios usuarios junto a sus respectivas contraseñas.

Figura 7: Ficheros con usuarios y sus contraseñas

También, podemos acceder al código de bases de datos para acceder a los datos de usuarios y contraseñas que hayan sido volcados. Para ello, podemos filtrar la búsqueda para que nos muestre aquellos archivos con extensión sql en los que se haya producido un volcado de memoria y que contengan el campo contraseña o password en alguna de sus tablas. Esto se puede conseguir fácilmente con el siguiente comando: *filetype:sql "dumping data for table" "password varchar".*

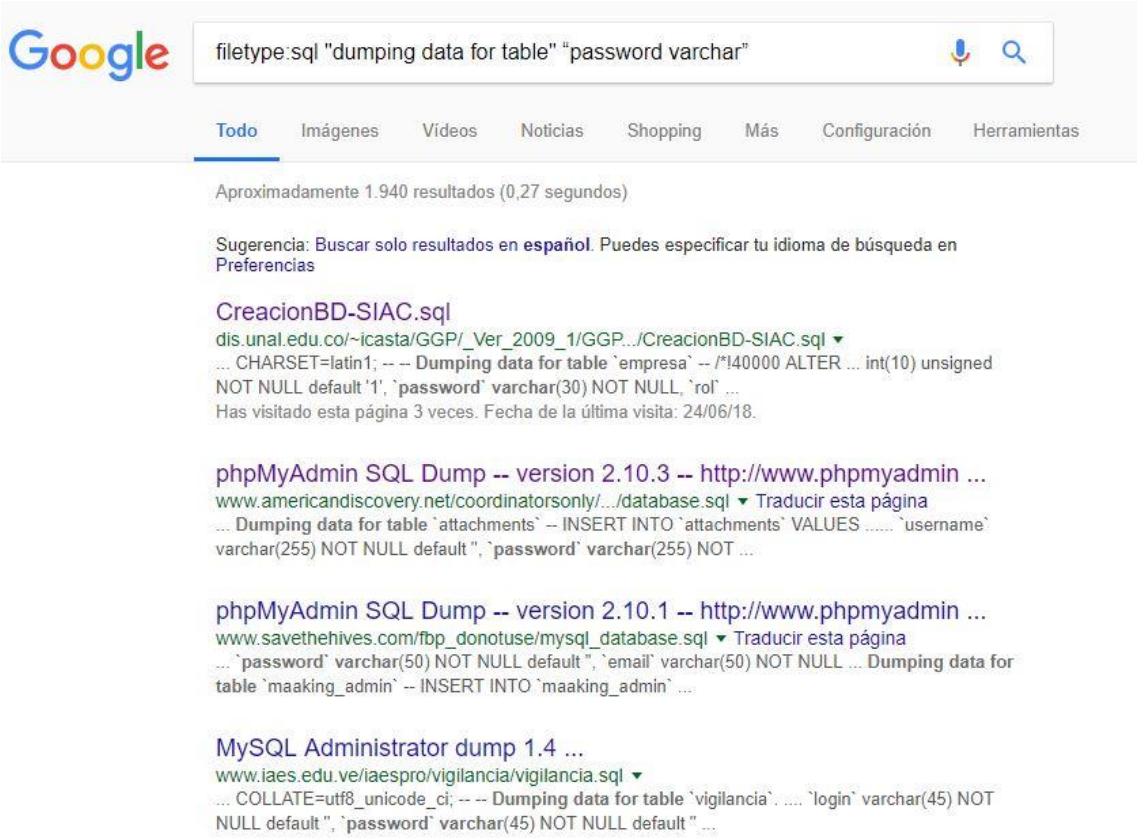

Figura 8: Resultado búsqueda ficheros sql

Como podemos observar, a partir de esta búsqueda podemos encontrar cientos de resultados con ficheros sql en los que aparece información relativa a contraseñas, entre otros tipos de datos.

```
DROP TABLE IF EXISTS `admin`;
CREATE TABLE IF NOT EXISTS `admin` (
  `uid` varchar(20) NOT NULL,
  `password` varchar(32) NOT NULL,
  `old_password` varchar(32) DEFAULT NULL,
  `admin_level` varchar(10) NOT NULL,
  `mobile` varchar(10) NOT NULL,
  `status` int(2) NOT NULL DEFAULT '0',
  `online` binary(1) NOT NULL DEFAULT '0',
  `last_login` datetime DEFAULT NULL,
  `created_by` varchar(20) NOT NULL,
  `created_on` datetime NOT NULL,
  `mod_by` varchar(20) DEFAULT NULL,
  `mod_on` datetime DEFAULT NULL
) ENGINE=InnoDB DEFAULT CHARSET=latin1;

--
-- Dumping data for table `admin`
--

INSERT INTO `admin` (`uid`, `password`, `old_password`, `admin_level`, `mobile`, `status`, `online`
('administrator', '202cb962ac59075b964b07152d234b70', NULL, 'L0002', '0777459969', 0, '0', NULL, '-
```

Figura 9: Código sql con información sobre usuarios

- **Formularios de acceso**: Podemos encontrar páginas que contengan los típicos formularios en los que debemos introducir nuestro usuario y contraseña para el acceso. El hecho de que estas páginas puedan ser encontradas puede permitir realizar ataques de fuerza bruta mediante diccionarios de datos con listas de posibles usuarios y contraseñas, que se van probando en el formulario hasta conseguir entrar. Podemos encontrar estas ventanas de acceso a través de la búsqueda de páginas web que son administradas remotamente con programas como VNC o introducir comandos *inurl:/admin/login.asp,* que nos mostrará miles de páginas diferentes donde podemos hacer login.

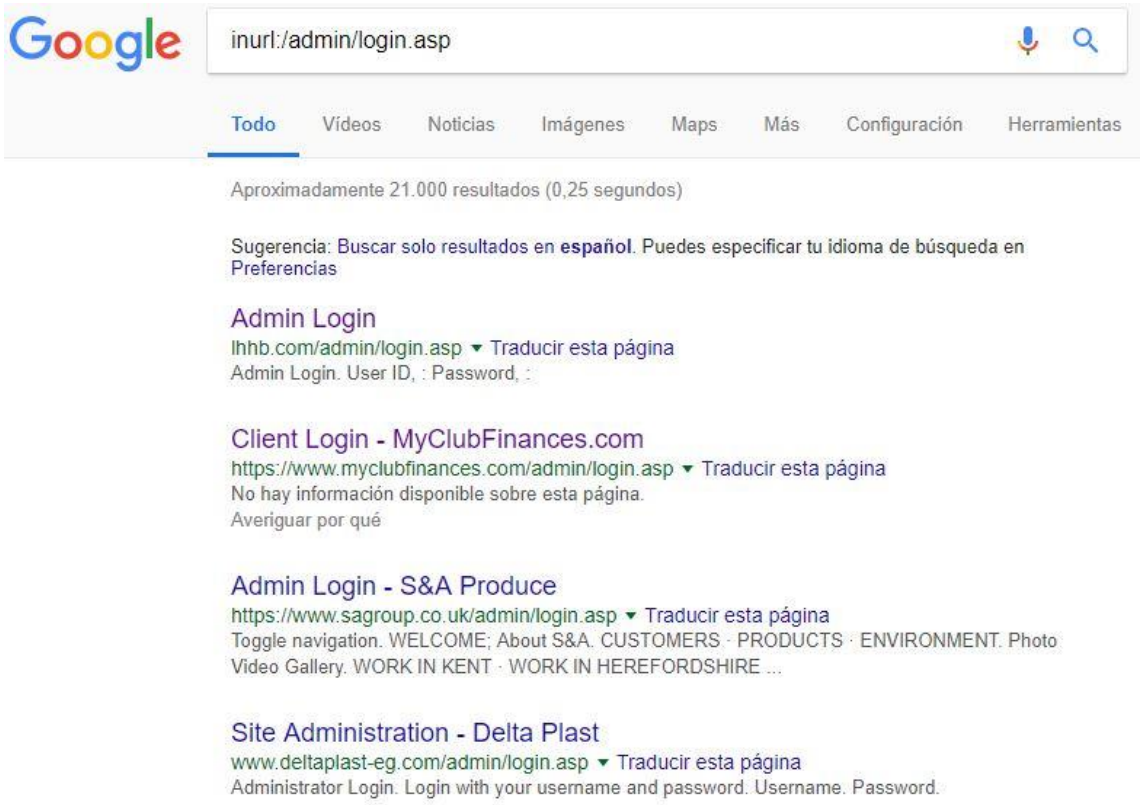

Figura 10: Búsqueda de formularios de acceso

- **Ficheros con nombres de usuario o mensajes de error que revelan información del usuario**: Esto puede facilitar los ataques con diccionarios comentados anteriormente, ya que solo es necesario sacar la contraseña para entrar.

- **Búsqueda de versiones antiguas de servidores web**: Es posible que algunos servidores web o programas instalados en ellos tengan versiones antiguas y desactualizadas. Debido a que las versiones suelen salir habitualmente para actualizar los distintos fallos de seguridad que se van descubriendo en estos entornos, es posible que estas vulnerabilidades existan aún en versiones antiguas. Para ello basta con buscar versiones antiguas de ciertos servidores de los cuales podamos explotar alguna de estas vulnerabilidades.

Una vez detectados, basta con buscar en alguna de las muchas páginas que contienen bases de datos con distintos exploits, y utilizarlos para entrar en el servidor.

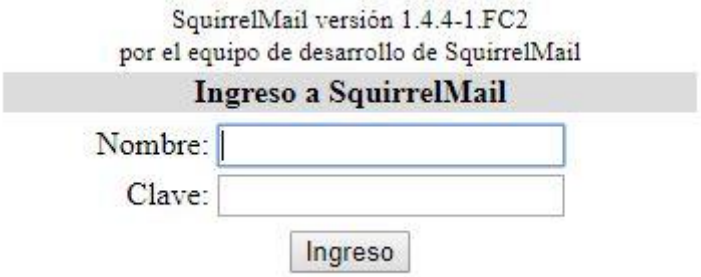

Figura 11: Ventana de acceso a SquirrelMail

En la figura anterior, podemos observar un ejemplo de una ventana de acceso al gestor de correo SquirrelMail, que está instalado en un servidor web. De esta versión 1.4.4 se pueden encontrar varios exploits en varias páginas que pueden permitir a los atacantes a acceder a los correos del servidor.

- **Dispositivos hardware online**: Mediante el uso de dorks podemos encontrar cientos de enlaces a webcams o cámaras de vigilancia que pueden estar desprotegidas permitiendo visualizar todas las imágenes o manejar los distintos controles de direccionamiento, con comandos como *inurl:"ViewerFrame?Mode=".*

Figura 12: Acceso a cámaras de vigilancia mediante Dorks.

8.1.2. Exploits DB

Todos estos dorks que nos permiten encontrar y acceder a páginas web con información crítica debido a la falta de securización de dichas páginas pueden ser encontrados en webs como Exploit-Database.

En este repositorio podemos encontrar todo tipo de exploits para aprovechar las vulnerabilidades de servicios remotos, aplicaciones web o realizar ataques de denegación de servicio. A su vez, incluye una base de datos de Google Hacking, producto del trabajo realizado por el hacker profesional Johnny Long en la pasada década, que, mediante el trabajo conjunto con la comunidad, consiguió catalogar todas estas consultas en la llamada Google Hacking Database. Tras el cierre de esta en 2010, Exploit-Database se encargó de su mantenimiento, añadiéndola como extensión a su web.

En ella podemos encontrar todo tipo de búsquedas para motores como Google, Bing y repositorios como GitHub, en los que podemos encontrar información relativa a archivos, servidores o redes vulnerables, mensajes de error que revelan demasiada información, archivos con contraseñas, portales de acceso y otros tipos de información vistos anteriormente.

Figura 13: Bases de datos de Google Hacking

8.1.3. Robots.txt

El estándar de exclusión de robots utiliza archivos de texto denominados robots.txt, que se crean y suben a los distintos sitios web para impedir que los robots de ciertos buscadores rastreen el contenido que no deseamos que indexen ni muestren en sus resultados.

Es decir, es un archivo público que usamos para indicar a esos rastreadores o arañas que parte o partes no deben entrar a rastrear en indexar en las páginas

web. En él, se suele especificar de manera sencilla, los directorios, subdirectorios, URLs o archivos de nuestra web que no deseamos que se muestren.

Para poder conseguir esto, existen una serie de comandos que se utilizan para aplicar todas las restricciones que deseemos a los sitios web.

Algunos de estos comandos son:

- **User-agent**: Indica que tipo de robot debe cumplir con las directivas que se indiquen a continuación.
- **Disallow**: Deniega el acceso a un directorio o página concreta.
- **Allow**: Funciona al contrario que la directiva Disallow, permitiendo el acceso a determinados directorios o páginas. Se puede utilizar para sobrescribir la directiva Disallow parcial o totalmente.
- **Sitemap**: Indica la ruta donde se encuentra el mapa del sitio en XML.
- **Crawl-delay**: Indica al robot el número de segundos que debe esperar entre cada página. Puede ser útil en casos en los que se necesita reducir la carga del servidor.

Sin embargo, estos archivos están disponibles de forma pública, y su contenido puede ser visto por cualquier persona con un navegador web y conocimientos avanzados. En muchos casos, el incluir un directorio en este archivo, indica la existencia de este, lo que puede ser utilizado por posibles atacantes.

Podemos encontrar archivos robots.txt en muchas páginas web usando dorks como *site:"*" inurl:"robots.txt"* para ello.

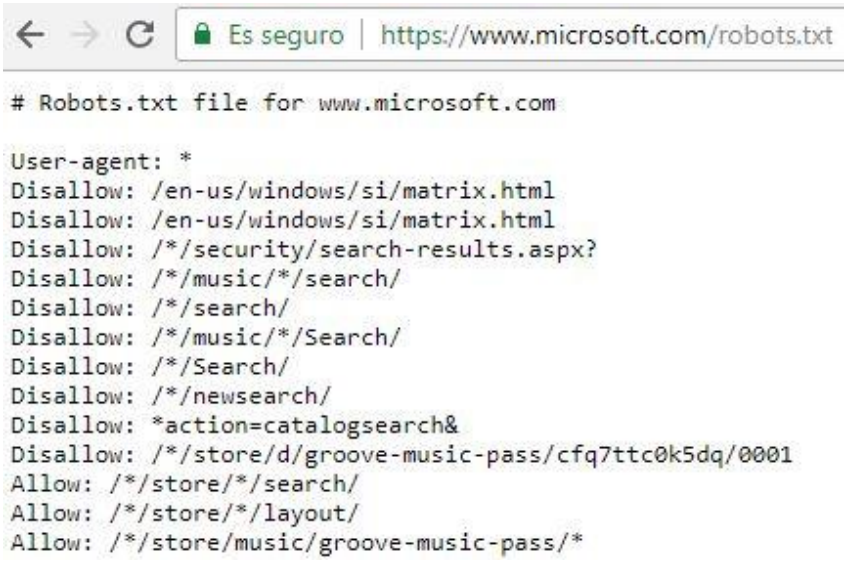

Figura 14: Ejemplo de archivo robot.txt

8.1.4. Crawlers

Un crawler, también conocido como araña o Web Spider, es un programa o webbot que se encarga de recorrer los enlaces de las páginas webs de una forma automática y sistemática.

Para su funcionamiento, parte de un conjunto inicial de URLs, conocidas como semillas, de las cuales va descargando sus páginas web asociadas y buscando enlaces nuevos en dichas páginas. Este proceso se realiza sucesivamente, añadiendo las nuevas URLs encontradas a la lista de direcciones, creando así un índice con todas las páginas que el crawler debe visitar.

Para encontrar estos enlaces, el crawler opta por dos alternativas cuando visita un sitio web:

- Buscar el archivo robot.txt para ver las reglas establecidas.
- Explotar el contenido visible a partir de las etiquetas HTML y los hipervínculos en listados en la página.

Estos crawlers pueden ser utilizados para extraer y recopilar información a partir de la estructura de las páginas web. Una de las herramientas más conocidas es Scrapy, que ofrece un framework para Python para hacer web scraping (recopilar información de forma automática de una web). Mediante Scrapy podemos desarrollar arañas simples para recorrer aquellas webs de las que queramos obtener información.

En la siguiente figura, podemos ver un ejemplo del resultado por un crawler realizado en Scrapy utilizado para buscar todas las ofertas de trabajo de distintas empresas en la página Craig List.

```
link,title
http://sfbay.craigslist.org/nby/npo/3389073434.html,LIFE SKILLS COACH Bilingual Spanish-English
http://sfbay.craigslist.org/sfc/npo/3389057953.html,Take Action! FUNDRAISE! $11-$17/hr
http://sfbay.craigslist.org/eby/npo/3389022509.html,Child and Family Therapist (Bi-lingual -  Spanish)
http://sfbay.craigslist.org/pen/npo/3389014674.html,Counselor
http://sfbay.craigslist.org/sby/npo/3389006779.html,Bell Ringer - Santa Clara
http://sfbay.craigslist.org/eby/npo/3388978152.html,ORGANIZE FOR PEACE & GET PAID!
http://sfbay.craigslist.org/pen/npo/3388974967.html,Grants Administrator (Foundation)
http://sfbay.craigslist.org/sfc/npo/3388967129.html,Development Intern
http://sfbay.craigslist.org/eby/npo/3388947776.html,Associate Director of Admissions
http://sfbay.craigslist.org/sby/npo/3388908896.html,Human Resources Manager
http://sfbay.craigslist.org/sfc/npo/3388840031.html,Chief of Programs
http://sfbay.craigslist.org/eby/npo/3388803144.html,Case Manager - El Cerrito
http://sfbay.craigslist.org/sby/npo/3388777537.html,Case Manager - San Jose
http://sfbay.craigslist.org/eby/npo/3388742615.html,BTSA Coach -- Part-time
http://sfbay.craigslist.org/nby/npo/3388736325.html,Housing Counselor - Rapid Rehousing Specialist
http://sfbay.craigslist.org/pen/npo/3388707460.html,Production Supervisor
http://sfbay.craigslist.org/sfc/npo/3388706641.html,Programs Associate at The Long Now Foundation
http://sfbay.craigslist.org/sfc/npo/3388697772.html,Senior Planner
http://sfbay.craigslist.org/sfc/npo/3388683830.html,Development Coordinator - Street Youth Program
http://sfbay.craigslist.org/eby/npo/3388676307.html,Program Coordinator
http://sfbay.craigslist.org/pen/npo/3388673455.html,Relief Counselor
http://sfbay.craigslist.org/eby/npo/3388664966.html,Detox Assistant Level 3
http://sfbay.craigslist.org/eby/npo/3388661311.html,Detox Assistant Level 2
http://sfbay.craigslist.org/sfc/npo/3388589833.html,Director of Information Technology
http://sfbay.craigslist.org/nby/npo/3388583438.html,    *FT Work for Greenpeace to STOP GLOBAL WARMING - $12-$13/hr*
http://sfbay.craigslist.org/pen/npo/3388575421.html,    *FT Work for Greenpeace to STOP GLOBAL WARMING - $12-$13/hr*
http://sfbay.craigslist.org/eby/npo/3388563471.html,"Therapist -- Outpatient Clinic, Children's Program"
http://sfbay.craigslist.org/sfc/npo/3388464002.html,Make a difference! Become an activist!
http://sfbay.craigslist.org/pen/npo/3388457358.html,Development Manager
http://sfbay.craigslist.org/scz/npo/3388454391.html,7 Positions to defend the enviroment need immediate filling!!!  Start
http://sfbay.craigslist.org/sfc/npo/3388454302.html,7 Positions to protect CA forrests need immediate filling!!!
http://sfbay.craigslist.org/sfc/npo/3388447369.html,7 Positions to defend reproductive rights need immediate filling!
http://sfbay.craigslist.org/sfc/npo/3388441733.html,P/T Bilingual (Cantonese) Residential Counselor
```

Figura 15: Ejemplo de araña con Scrapy

8.1.5. Deep Web

Como ya se ha comentado, la información que podemos encontrar en los buscadores tradicionales solo representa una pequeña parte de todo Internet. La mayor parte del contenido de este se encuentra en la denominada Deep Web,

ya que incluye toda información no indexada por los motores de búsqueda tradicionales. En ella podemos encontrar todo tipo de documentos gubernamentales, informes financieros, científicos o académicos y otro tipo de información que podría considerarse ilegal.

A pesar de que estos datos sean de difícil acceso, podemos tratar de localizarlos navegando por esta DeepWeb. Para ello debemos usar TOR, el cual nos ofrece una red de comunicación distribuida de baja latencia y sobrepuesta sobre Internet, en la que se garantiza nuestro anonimato debido a que el intercambio de mensajes no revela nuestra dirección IP, y además mantiene la integridad y el secreto de la información que viaja por ella.

El uso de TOR permite a los usuarios publicar un sitio web u otros servicios sin necesidad de revelar desde donde se publica. Por ello, la búsqueda en TOR mediante los distintos motores de búsqueda que este ofrece como DuckDuckGo o Not Evil, supone una de las herramientas más importantes para obtener información sobre delincuentes que utilizan la Deep Web para cometer sus acciones.

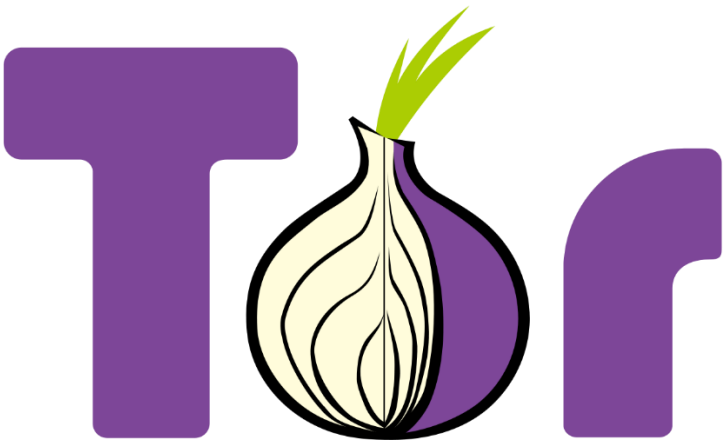

Figura 16: The Onion Router (TOR)

8.2. Redes sociales

En general, cuando hablamos de una red social nos referimos a una estructura social formada por personas o entidades conectadas y unidas entre sí por algún tipo de relación social como amistad o parentesco, intereses comunes, o que comparten conocimientos.

Aunque el concepto de red social pueda parecer algo nuevo, lo cierto es que ya a finales del siglo pasado empezaron a originarse este tipo de sitios web. No fue sin embargo hasta principios de la década de los 2000 cuando empezaron a originarse las primeras redes sociales tal y como las conocemos hoy en día. La gran revolución llegó en el año 2003 con la llegada de MySpace, que a pesar de que con el paso del tiempo ha perdido bastante popularidad, fue la primera red social que nos permitía crear un perfil completo sobre nosotros mismos, donde podíamos publicar nuestros gustos, intereses o preferencias a la hora de conocer gente, además de la posibilidad de compartir fotos o música. Posteriormente, fueron surgiendo las grandes redes sociales con las que estamos más familiarizados, como Facebook o Twitter, y que desde el momento de su aparición han ido incrementando el número de usuarios registrados de manera exponencial.

Figura 17: Redes sociales

Todos estos usuarios generan una gran cantidad de datos. Debido a la capacidad de interacción y difusión que poseen estas redes sociales, toda esta información que publicamos puede estar disponibles para cualquier persona del mundo de manera instantánea.

Estas características permiten que las redes sociales supongan hoy en día uno de los medios más importantes de comunicación y transmisión de datos, lo que las convierte en una de las más importantes y accesibles fuentes de información. Dicha información puede ser utilizada posteriormente como herramienta de marketing o para analizar corrientes de opinión, pero también supone un peligro, ya que se pueden encontrar todo tipo de datos de carácter personal y que quedan de manera pública en Internet.

Por estos motivos, las redes sociales suponen una herramienta de extracción de datos en sí mismas, ya que navegando a través de ellas podemos obtener todo tipo de información que hayan publicado sus usuarios sin tener en cuenta que pueden poner en riesgo su privacidad. En este apartado, analizaremos las herramientas disponibles para el análisis de estas redes sociales, que van desde aquellas que nos permiten detectar la presencia de usuarios en ellas, a las que nos ofrecen la posibilidad de realizar búsquedas avanzadas para obtener resultados más concretos.

8.2.1. Detección de presencia en redes sociales

En la mayoría de estas aplicaciones no utilizamos nuestro nombre real, si no que utilizamos para identificarnos lo que se conoce como nombre de usuario. Este nombre de usuario suele ser utilizado siempre que nos registramos en alguna red social por lo que, por ejemplo, si alguien conoce nuestro usuario en Twitter, pueda utilizar este identificador para ver si formamos parte de otra red social diferente y así encontrar más información sobre nosotros.

Para esto existen herramientas que, a partir de un usuario conocido, buscan en un conjunto de redes sociales para detectar la existencia de usuarios con el mismo nombre. Esta herramienta suele ser utilizada fuera del ámbito de OSINT, para comprobar la disponibilidad de algún nombre de usuario en alguna web y así poder registrarse con él, pero a su vez nos indica todos los sitios web en los que sí existe un nombre usuario, que con una alta probabilidad puede estar asociado a la misma persona.

Una de las herramientas más conocidas y completas que podemos encontrar es KnowEm. Entre las distintas funciones disponibles, KnowEm ofrece un motor de búsqueda de usuarios en redes sociales, un servicio de protección de marcas, una plataforma de marketing, y una red social en sí misma.

Para utilizarla, solo debemos introducir el nombre de usuario que deseemos encontrar, y nos indicará todas las redes sociales en las que detecta un usuario con ese mismo nombre.

Figura 18: Herramienta KnowEm

Una vez introducido el usuario, nos mostrará los resultados en algunas de las redes sociales más populares, pero podremos elegir realizar la búsqueda en en función de la temática de la web, la cuál puede ser:

- **Blogging** (como Blogger, WordPress o Tumblr)
- **Bookmarking** (Pinterest o Instapaper)
- **Negocios** (LinkedIn o eBay)

- **Comunidades** (Facebook o Steam)
- **Diseño** (como por ejemplo Behance)
- **Entretenimiento**
- **Salud** (por ejemplo, MyFitnessPal)
- **Información** (ask.fm o Wikipedia)
- **Microblogging** (Twitter o Foursquare)
- **Música** (SoundCloud, last.fm o Spotify)
- **Noticias** (BuzzFeed o Reddit)
- **Fotos** (Instagram, Imgur, Flickr, Giphy, …)
- **Tecnología**
- **Viajes**
- **Video** (siendo YouTube la más conocida)

Figura 19: Resultado de una búsqueda en KnowEm

Además, la aplicación nos permite realizar búsquedas de dominios en varios países y regiones del mundo, y buscar marcas registradas.

Figura 20: Búsqueda de dominios registrados

A parte de KnowEm, existen otras herramientas alternativas para la detección de presencia en redes sociales como NameChk o CheckUsernames, las cuales funcionan de una manera similar.

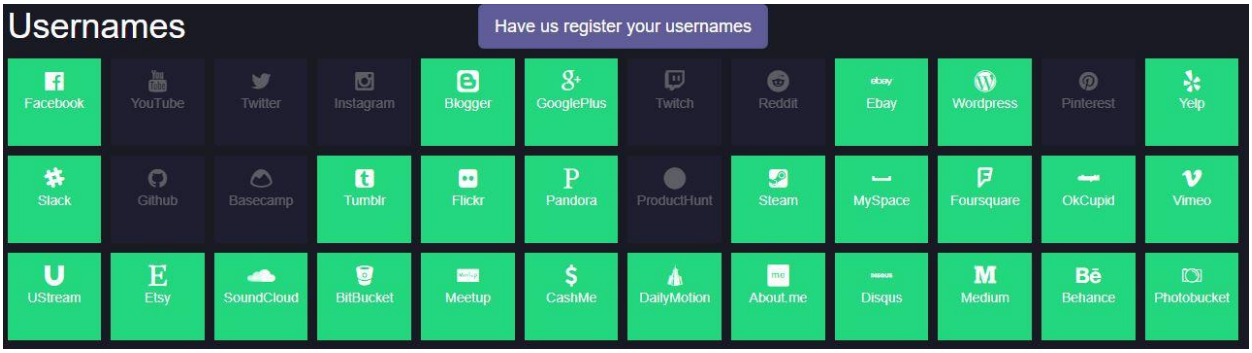

Figura 21: Búsqueda de usuarios con NameChk

Nota: En verde los sitios web donde el nombre no está registrado

8.2.2. Redes sociales de uso masivo

A pesar de la existencia de una gran cantidad de diferentes redes sociales, muchas de estas están especializadas en un ámbito específico, por lo que su uso es reducido.

Por ello, nos centraremos en las redes sociales con las que la mayor parte de las personas convive día a día, pasando buena parte de su tiempo en compartir información con el resto de los usuarios.

Estas redes sociales suponen una importantísima fuente de información, ya que están dedicadas en gran parte a la publicación de datos, casi siempre de carácter personal y realizándose dicha publicación muchas veces de forma incontrolada o inconscientemente de que se está poniendo en riesgo nuestra privacidad.

8.2.2.1. Twitter

La plataforma Twitter, consiste en un servicio de microblogging, desarrollado en el año 2006 en la ciudad de San Francisco (Estados Unidos) por algunos de los empleados de Odeo, entre los que se encontraban extrabajadores de Google como Evan Williams y Biz Stone, con la colaboración de Jack Dorsey, Evan Henshaw-Plath y Noah Glass. Ese mismo año fue lanzada la versión definitiva, la cual tuvo unos comienzos difíciles hasta que fue adquirida por la nueva compañía, Twitter, Inc que se independizó de su gestora en 2008. Desde entonces Twitter ha ido ganando ha ido ganando adeptos rápidamente, sobre todo tras la aparición de la plataforma en otros lenguajes como el español en 2009. Actualmente cuenta con alrededor de 500 millones de usuarios que generan 65 millones de tuits al día y realizan más de 800000 peticiones de búsqueda diarias.

El microblogging es una variante de los blogs (diarios personales online), que se caracterizan por la brevedad de sus mensajes y la facilidad de publicación, pudiéndose enviar desde diversas plataformas como móviles, aplicaciones de escritorio o desde la página web.

La red de Twitter permite enviar mensajes de texto plano de corta longitud, con un máximo de 280 caracteres (140 en sus orígenes), denominados tuits o tweets, que por defecto se muestran a los demás usuarios de manera pública. Además de la publicación de estos tweets, los usuarios pueden seguir a otros usuarios para poder ver los tweets que estos publiquen en su página principal, retweetear para compartir mensajes publicados en otra cuenta, u observar los trending topics o temas más mencionados en ese momento.

El problema viene cuando utilizamos la aplicación para publicar información sobre nosotros. Muchos son los usuarios que tienen la opción de geolocalización activada, que nos indica la ubicación desde la que están escribiendo, utilizan su nombre real, añaden sus gustos a sus biografías o publican fotos y tweets que pueden darnos datos personales.

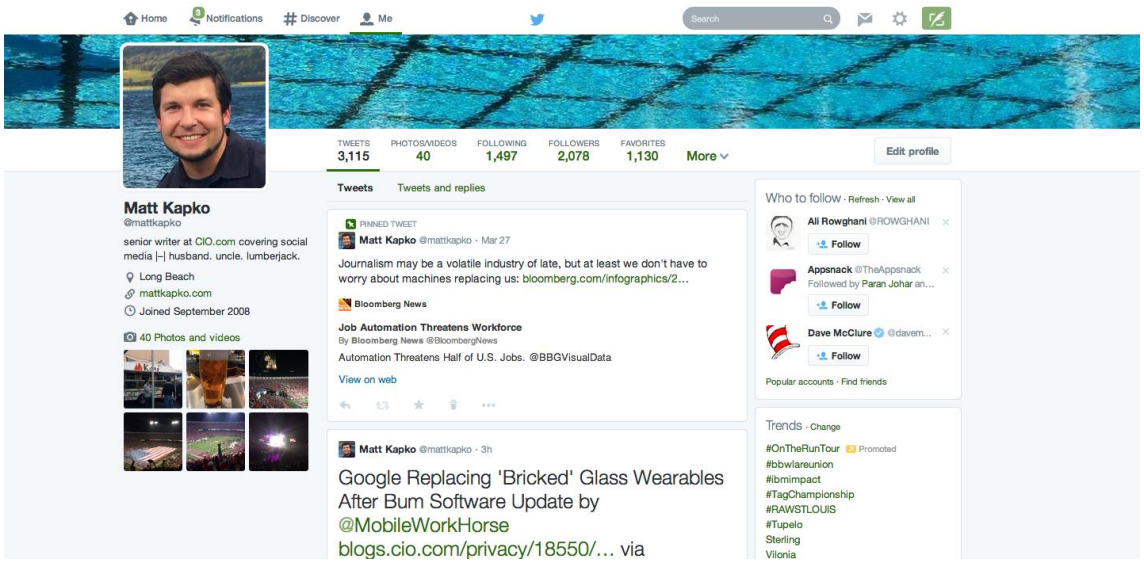

Figura 22: Ejemplo de cuenta de Twitter

Como podemos observar en la Figura 22, solo con visitar la página principal de un usuario podemos encontrar todo tipo de información. Entre los datos que podemos extraer, encontramos su foto, su nombre real (Matt Kapko), su lugar de residencia (Long Beach), su profesión (escritor senior en la página CIO.com) o que está casado (Aparece la palabra “husband” en su biografía), todos datos que podrían ser usados para realizar un perfil sobre su persona. Por otro lado, el usuario cuenta con 3115 tweets y 40 fotos, además de sus listas de seguidores, tweets favoritos y lista de cuentas a las que sigue, que podrían ser analizados para obtener más información como opiniones, aficiones, o personas relacionadas con el objetivo.

Búsquedas avanzadas en Twitter

Para poder acceder a toda esta información de manera más fácil, existen varios buscadores avanzados que nos permiten filtrar el contenido que queremos encontrar a través de una serie de parámetros. La propia aplicación de Twitter nos ofrece un buscador avanzado en el que podemos buscar palabras o frases exactas en cualquier idioma, lo que se puede utilizar para identificar delitos de odio o para saber que se está comentando sobre un tema en concreto, además de buscar tweets publicados en alguna cuenta en particular, escritos desde algún lugar o en una fecha específica.

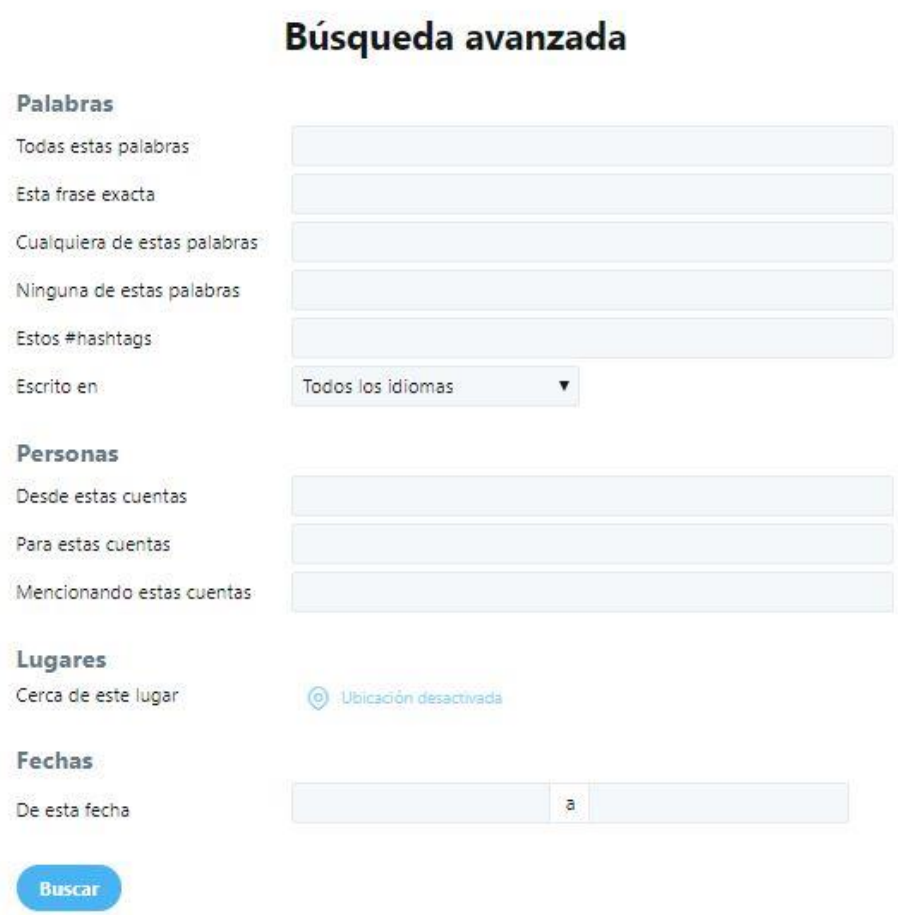

Figura 23: Búsquedas avanzadas en Twitter

Twitter ofrece, al igual que el resto de grandes redes sociales ofrece un sistema REST (Transferencia de Estado Representacional). Cuando hablamos de un sistema REST, nos referimos a cualquier interfaz entre sistemas que use HTTP para obtener datos o generar operaciones sobre estos datos en todos los formatos posibles, como pueden ser XML o JSON.

Esta API REST es abierta para desarrolladores, y permite acceder a leer o escribir datos de Twitter. Es capaz de obtener información sobre los distintos campos que conforman un tweet, como el lugar y la fecha de publicación, el usuario que lo escribió, el número de retweets, menciones a otras cuentas, hashtags utilizados, etc. Además, cuenta con una API de streaming que proporciona acceso a un alto volumen de tweets a cambio de una baja latencia.

Tinfoleak

Una de las herramientas más conocidas actualmente en el análisis de esta red social es Tinfoleak. Desarrollada por Vicente Aguilera, consiste en una herramienta de código abierto, que automatiza la extracción de datos de Twitter y facilita su posterior análisis. Podemos encontrarla para su descarga en la página personal del desarrollador, además de estar disponible en versión web.

Para su utilización, introducimos una cuenta de usuario, y recibiremos a la dirección email que seleccionemos el análisis completo de toda la información extraída sobre dicho usuario. Entre estos datos podemos encontrar:

- Información básica sobre el usuario como imagen de perfil, estado de verificación, nombre de usuario, descripción de la cuenta, ID, seguidores, usuarios a los que sigue, ubicación (si está activada), zona horaria, idioma, etc.

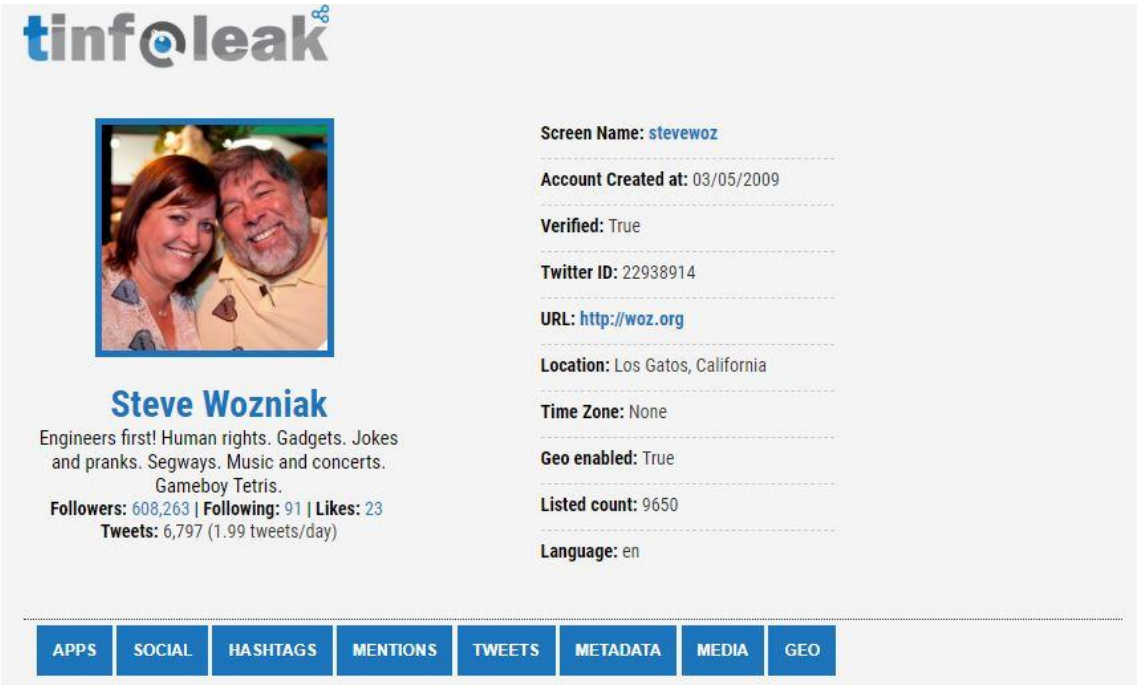

Figura 24: Ejemplo Tinfoleak

- Aplicaciones utilizadas por el usuario para publicar los tweets, organizándolas por porcentaje de uso, número de usos en cada una, fecha de uso, primer y último tweet publicado, etc.

CLIENT APPLICATIONS

Source	Uses	Percentage	First Use	First Tweet	Last Use	Last Tweet
Foursquare	181	90.5 %	02/19/2018	view	07/05/2018	view
Twitter for iPhone	6	3.0 %	06/05/2018	view	06/27/2018	view
OS X	8	4.0 %	02/28/2018	view	06/22/2018	view
Twitter Web Client	5	2.5 %	05/06/2018	view	06/15/2018	view

Figura 25: Análisis de aplicaciones

- Análisis de geolocalización: Fecha y hora de publicación de los tweets, coordenadas y localización desde la que se publica cada tweet, aplicación utilizada para su publicación, etc.

GEOLOCATION INFORMATION

TWEETS WITH GEOLOCATION ENABLED

Date	Time	Coordinates	Media	App	Tweet	Location
07/05/2018	00:59:18	39.53024439, -119.81573582		Foursquare	view	Reno
07/04/2018	02:32:59	38.12049, -120.46972764		Foursquare	view	California
07/03/2018	23:18:36	38.77254221, -121.26808165		Foursquare	view	Roseville
07/03/2018	23:17:56	38.77125955, -121.26619146		Foursquare	view	Roseville
07/03/2018	03:03:45	37.42630395, -122.07944747		Foursquare	view	Mountain View
07/02/2018	05:18:04	37.68702449, -122.1310782		Foursquare	view	San Leandro
07/02/2018	00:17:09	38.0884893, -122.5534492		Foursquare	view	Novato
07/01/2018	18:49:37	37.21581003, -121.96437453		Foursquare	view	Los Gatos
07/01/2018	01:04:05	37.766756, -122.43056333		Foursquare	view	San Francisco
07/01/2018	00:25:43	37.7905481, -122.4226117		Foursquare	view	San Francisco
06/23/2018	02:13:26	33.5342441, -111.96466371		Foursquare	view	Paradise Valley
06/23/2018	02:13:01	33.53365025, -111.96487986		Foursquare	view	Paradise Valley
06/22/2018	04:35:31	33.45179743, -112.07665354		Foursquare	view	Phoenix

Figura 26: Análisis de geolocalización

- Análisis del contenido de los tweets, como palabras utilizadas, menciones a otras cuentas, hashtags, contenido multimedia, metadatos, etc.

USER MENTION DETAIL

Date (since)	Date (until)	RT's	Likes	Count	Name	Mention
03/29/2018	07/03/2018	6	64	2	Cheesecake Factory	@Cheesecake
07/03/2018	07/03/2018	7	110	1	Shoreline Amp	@ShorelineAmp
07/02/2018	07/02/2018	3	57	1	In-N-Out Burger	@innoutburger
04/21/2018	07/01/2018	7	91	2	RuthsChrisSF	@RuthsChrisSF
06/27/2018	06/27/2018	0	2	1	Nelson Mandela Vz	@NelsonMandelaVz
06/27/2018	06/27/2018	0	2	1	jose rivas	@joserivas788
06/27/2018	06/27/2018	0	2	1	camil@	@corin80301423
06/27/2018	06/27/2018	0	2	1	Franyer Bolivar	@franyerbolivar2
06/27/2018	06/27/2018	0	2	1	Gregory David Myers	@Gregory69021620

Figura 27: Análisis de menciones en Tweets

8.2.2.2. Facebook

Facebook es una compañía estadounidense que ofrece servicios de redes sociales y medios sociales en línea, cuya sede se encuentra en Menlo Park (California). La red social fue lanzada en el año 2004 por Marck Zuckerberg junto con otros estudiantes de la Universidad de Harvard. Aunque en sus inicios fue creada para el conjunto de los estudiantes de Harvard, posteriormente se fue ampliando a otras universidades e instituciones de los Estados Unidos, hasta el año 2005, donde con 6 millones de usuarios era ya mundialmente conocida. La aparición de la web en varios idiomas, la accesibilidad desde cualquier tipo de dispositivo (ordenadores personales, portátiles, tabletas o móviles), y la posibilidad para registrarse de toda persona mayor de 13 años, así como las distintas funcionalidades añadidas a la red en los últimos años, han convertido a Facebook en la red social más utilizada con más de 2 billones de usuarios y que le han otorgado en el mercado un valor de casi 500 mil millones de dólares.

Entre los principales servicios que Facebook ofrece podemos encontrar:

- Lista de amigos: Podemos agregar a cualquier persona que conozcamos y este registrada, siempre que acepten nuestra invitación.
- Chat: Servicio de mensajería instantánea.
- Grupos y páginas: Permite reunir personas con intereses comunes. En ellos se pueden añadir fotos, videos, mensajes, etc.
- Fotos: Según datos de la compañía, Facebook almacena unos 5 mil millones de fotos de usuarios.
- Botón “Me gusta”: Permite a los usuarios valorar si el contenido de un usuario es de su agrado.

- Aplicaciones y juegos: A través del App Center, Facebook ofrece una herramienta para desarrolladores con la que pueden añadir sus juegos y aplicaciones a la página web.

Todos estos servicios derivan en una gran cantidad de información proveniente de los usuarios, lo que ha supuesto para la compañía una constante cobertura mediática sobre su privacidad, así como una intensa presión ante la cantidad de noticias falsas, páginas de incitación al odio y representaciones de violencia que prevalecen en estos servicios.

A pesar de que Facebook ha aumentado sus opciones de privacidad ante estas situaciones y los recientes escándalos como el de la minería de datos de Cambridge Analytica, donde se manipularon los datos de 50 millones de usuarios para influir en las elecciones estadounidense, la información publicada por los usuarios sigue siendo accesible si no se toman las medidas de privacidad necesarias.

Figura 28: Ejemplo de perfil de Facebook

Facebook nos ofrece un apartado dentro de la opción de Configuración General de la Cuenta, donde podremos descargar la información relativa a nuestro perfil. Con esto podemos hacernos una idea de la cantidad de información que

los usuarios pueden publicar, que es almacenada por la aplicación, y entre la que podemos encontrar:

- **Publicaciones**: Fotos, videos, texto o actualizaciones de estado que comparten los usuarios, así como aquellos en los que un usuario ha sido etiquetado, encuestas creadas, y publicaciones de otras personas en la biografía del usuario.
- **Comentarios realizados** en publicaciones de otras personas o grupos.
- **Me gusta y reacciones**: Publicaciones, comentarios, páginas o grupos donde un usuario indica que le gustan.
- **Amigos**: Personas con las que cada usuario está conectado en Facebook.
- **Personas o páginas de organizaciones o negocios** a los que sigue un usuario.
- **Mensajes intercambiados** a través del chat.
- **Eventos** creados.
- **Grupos y páginas**: Donde se agrupan personas que comparten intereses comunes.
- **Historial de pagos** realizados a través de Facebook.
- Una **lista de las localizaciones** creadas.
- **Información del perfil**, como información de contacto (direcciones postales, números de teléfono, direcciones de correo electrónico), información personal (nombre, fecha de nacimiento, empleo, formación académica, lugares visitados, etc.), libretas de direcciones de contacto agregadas a los amigos, videos visualizados o la opción de reconocimiento facial.
- **Aplicaciones y sitios web** en las que se inició sesión en Facebook

- **Anuncios**: Intereses, interacciones y relación con los anunciantes que influyen en los anuncios mostrados en la página.
- **Historial de búsqueda**: Historial de palabras, frases o nombres buscados.

Todos estos datos que podemos obtener de los usuarios son almacenados por Facebook, y pueden ser fácilmente obtenidos realizando un análisis de la página de perfil de un usuario si este no ha tomado las medidas de privacidad necesarias.

Además, podemos encontrar en Internet distintos buscadores avanzados que nos permitirán extraer la información que deseemos filtrando la búsqueda mediante distintos campos como nombres, genero, idioma, religión, ideología política, etc. Un ejemplo de este tipo de páginas puede ser Netbootcamp.

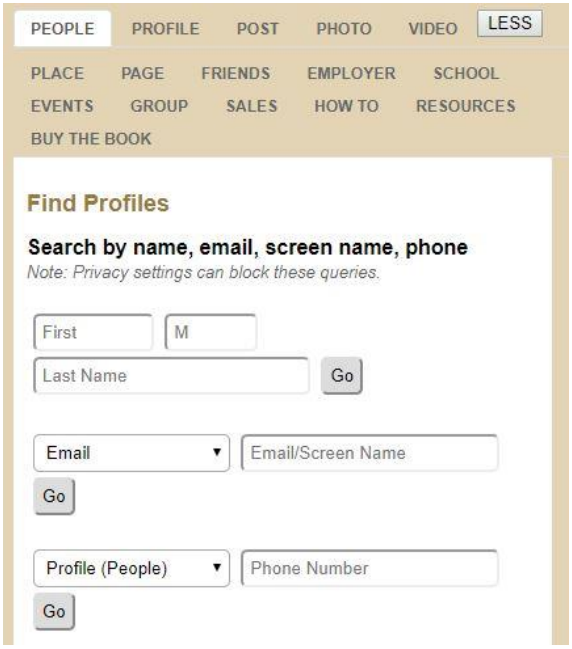

Figura 29: Netbootcamp

8.2.2.3. Instagram

En el año 2010, Kevin Systromy y Mike Krieger, dos estudiantes de la Universidad de Stanford, crearon la red social Instagram. A diferencia del resto de las redes sociales más conocidas que surgieron como sitios web, fue diseñada originalmente para smartphones, apareciendo primero en iPhone y posteriormente para dispositivos Android en 2012, cuando fue adquirida por Facebook.

Instagram fue una de las redes sociales que mayor éxito tuvo desde sus inicios, al dedicarse exclusivamente a la publicación de fotos y vídeos, una de las aficiones más extendidas entre la población. También permitía acercar la fotografía a toda persona que estuviese interesada en ella, pero no tuviera suficientes conocimientos, mediante la posibilidad de aplicar efectos fotográficos a las imágenes, además de ofrecer una forma de difundir este contenido a través de compartirlo en otras redes sociales como Facebook, Twitter o Flickr.

Actualmente, con sus más de 800 millones de usuarios activos, Instagram ha dejado de ser una red social dirigida únicamente a los aficionados a la fotografía, para convertirse en un sitio utilizado por los usuarios para publicar fotos y vídeos de su vida cotidiana.

Estos hechos la han convertido en una gran fuente de datos de carácter personal, en la que, a partir de las fotos, videos, historias publicadas o las recién incorporadas encuestas, podemos obtener todo tipo de información, principalmente asociada a lugares visitados por el usuario, gustos, aficiones y personas relacionadas con dicho usuario que pueden aparecer etiquetadas en sus fotos.

Al igual que en las redes sociales vistas anteriormente, existen varios buscadores avanzados que permiten realizar búsquedas filtrando por hashtags, que son las etiquetas que ponen los usuarios a sus publicaciones. Una de los

más conocidas es WEBSTAGRAM donde, por ejemplo, podemos buscar por el hashtag #trip con el que podemos visualizar todas las publicaciones que han realizado los usuarios sobre los lugares en los que pasan las vacaciones.

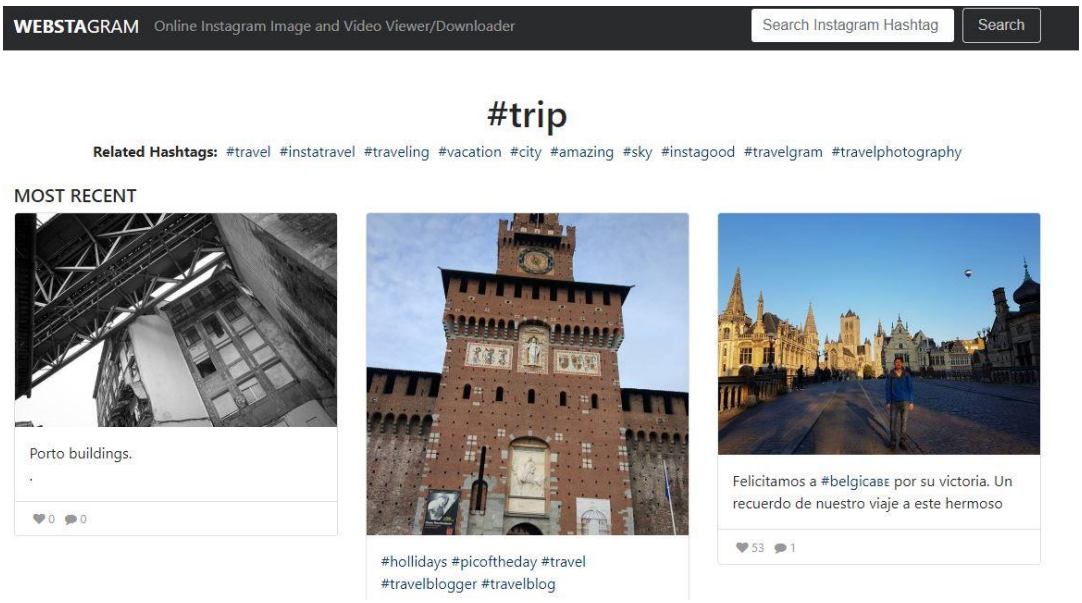

Figura 30: Búsqueda avanzada en Instagram

8.2.2.4. LinkedIn

Esta red social fue creada en 2002 por Reid Hoffman y otros estudiantes, y lanzada en mayo del siguiente año. Consiste en una comunidad social orientada a las empresas, a los negocios y al empleo. En ella, los usuarios revelan libremente su experiencia laboral y destrezas, poniendo en contacto así a millones de empresas y empleados.

Entre la información que podemos encontrar en esta red social tenemos:

- Una red de contactos para los usuarios construida mediante las conexiones de estos con otras personas.
- Los usuarios pueden subir su CV, mostrando así sus experiencias laborales y habilidades profesionales, donde se puede indicar también el lugar de trabajo actual.

- Información acerca de los puestos de trabajo que busca alguna empresa.
- Fotos de perfil con las que poder identificar a los usuarios.
- Preferencias de los usuarios a la hora de buscar trabajo.

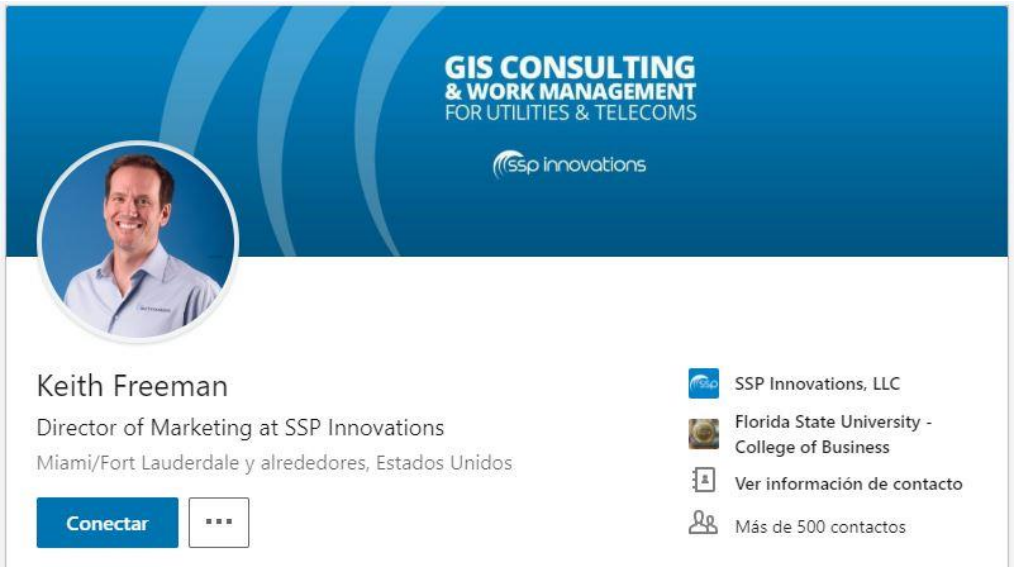

Figura 31: Perfil de LinkedIn

Para acceder a esta información solo hay que estar registrado en la red social, ya que algunos usuarios pueden elegir la opción de visualizar quien está visitando sus perfiles.

8.2.3. Redes sociales de uso reducido.

Como hemos visto con LinkedIn, tras el éxito que tuvieron las redes sociales en sus primeros años, fueron surgiendo nuevas páginas que se segmentaron en función de su temática. Podemos encontrar webs dedicadas a deportes, música, fotografía, viajes, etc. Aunque la cantidad de información que podemos encontrar en estos sitios pueda ser menor, nos pueden ser de gran utilidad para identificar aficiones o intereses de los usuarios, lugares visitados u opiniones del objetivo. Podemos llegar a saber cuál es el tipo de música que

alguien escucha a través de sus cuentas de Spotify o YouTube, si les gusta viajar, hacer deporte, si son aficionados de algún equipo o descubrir otros hobbies como la fotografía o la tecnología a través de cuentas en Pinterest o GitHub.

8.2.3.1. Páginas de citas o preguntas

Algunas de las aplicaciones donde podemos encontrar más información acerca de una persona son las páginas de preguntas como ask.fm o CuriosCat donde los usuarios responden a las preguntas hechas por otros usuarios, o las páginas de citas como Badoo. En estas redes de contactos los usuarios suelen introducir datos personales que reflejan sus aficiones, preferencias a la hora de buscar pareja, sexualidad, tipo de relación que buscan, vivienda, edad y por supuesto apariencia a través de la publicación de fotos.

Figura 32: Badoo, una red social de citas

8.2.3.2. Redes sociales de turismo.

También suponen una fuente de información importante las redes sociales dedicadas a viajes, donde se permite a los usuarios realizar reseñas de aquellos lugares, hoteles o restaurantes donde han estado. Un ejemplo de este tipo de red social puede ser TripAdvisor. En su página web podremos acceder a los perfiles de sus usuarios a través de la URL:

https://tripadvisor.es/members-reviews/{{username}}

Una vez dentro del perfil, podremos acceder a las opiniones realizadas de cada lugar visitado, las puntuaciones realizadas, fotos publicadas que pueden contener su imagen, y un mapa en el que podemos observar todos estos lugares visitados.

Figura 33: Lugares visitados en TripAdvisor

De una temática similar encontramos redes como Foursquare, la cual consiste en un servicio basado en localización aplicado a redes sociales. De esta forma, la web irá guardando todos los lugares específicos donde uno se encuentra,

recompensando al usuario cuando se descubren nuevos lugares. Se crean así listas con los lugares favoritos o más visitados por el usuario, pudiendo visualizarlos en un mapa.

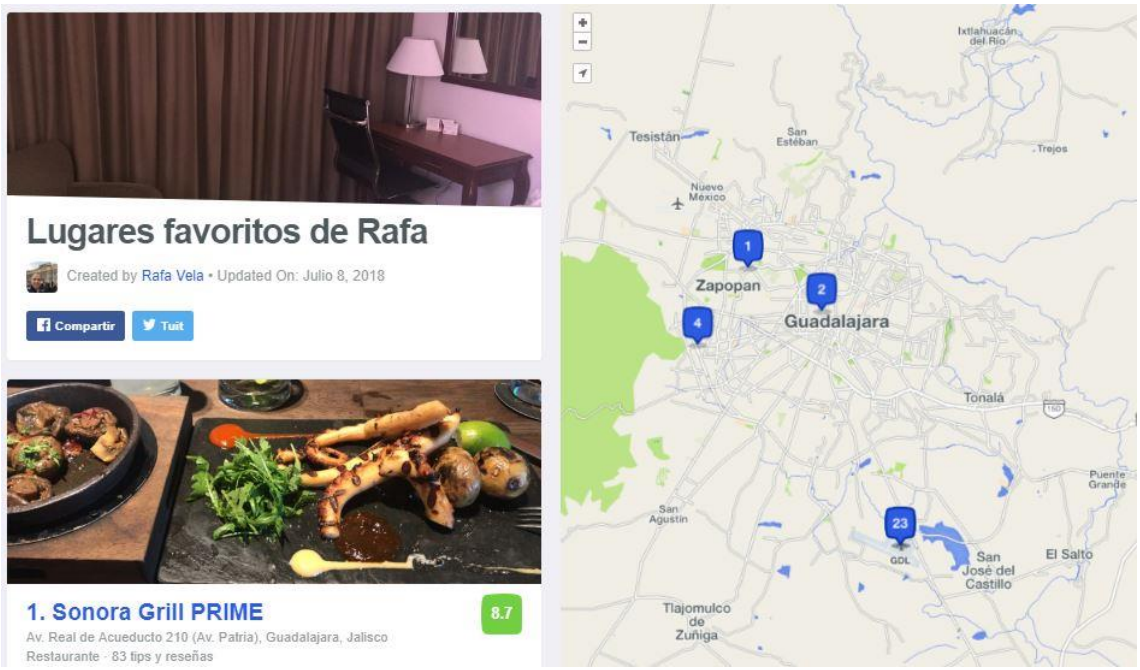

Figura 34: Lugares favoritos en Foursquare

8.2.3.3. Aplicaciones de deportes.

Otras redes sociales que pueden darnos información sobre localizaciones son las aplicaciones deportivas, como Strava o Endomondo. En ellas podemos ver el recorrido que realizan los usuarios cuando salen a hacer ejercicio, lo que nos puede mostrar entre otros datos el lugar de residencia del usuario.

Figura 35: Rutas en Strava

La aplicación Strava lanza cada año un mapa de calor en el que se muestran todas las salidas registradas por los miles de usuarios de su comunidad, y que podemos encontrar en el siguiente enlace:

https://labs.strava.com/heatmap

Estos mapas han generado mucha polémica al quedar expuestas bases militares, con las rutinas de los propios militares, y bases secretas en zonas de conflicto que no aparecen en otros mapas.

Figura 36: Bases militares en Strava

8.3. Herramientas de análisis de datos personales.

Todos estos datos que podemos extraer mediante las redes sociales pueden servir de fuente para otras aplicaciones, que a partir de ellos permiten completar o analizar en mayor profundidad la información obtenida.

8.3.1. Metabuscadores

Una de las herramientas más utilizadas para encontrar la huella digital de una persona son los denominados metabuscadores. Un metabuscador es un sistema que localiza información de diferentes motores de búsqueda, utilizando las bases de datos de estos, y mostrando una combinación de las mejores páginas que le devuelve cada uno, es decir, un metabuscador es un buscador de buscadores.

Podemos encontrar varios de ellos en distintas páginas web, siendo Pipl uno de los más utilizados. En esta página podremos buscar datos sobre cualquier persona a partir de su nombre, usuario, email, localización o teléfono. Tras buscar toda la información sobre la persona introducida en distintos buscadores, creará un perfil completo con todos los datos recopilados, siendo algunos como teléfonos o emails solo accesibles desde la versión Pro para la que habrá que registrarse.

Entre los posibles datos que se nos mostrarán aparecerá la foto de perfil, nombre completo, edad, género, carrera profesional, educación, nombre de usuario, teléfonos, emails, localizaciones y personas relacionadas por parentesco, así como enlaces a todas las redes sociales en las que el usuario este registrado.

En la siguiente figura podemos observar un ejemplo del resultado obtenido en Pipl tras pedirle toda la información que pueda encontrar sobre un objetivo concreto.

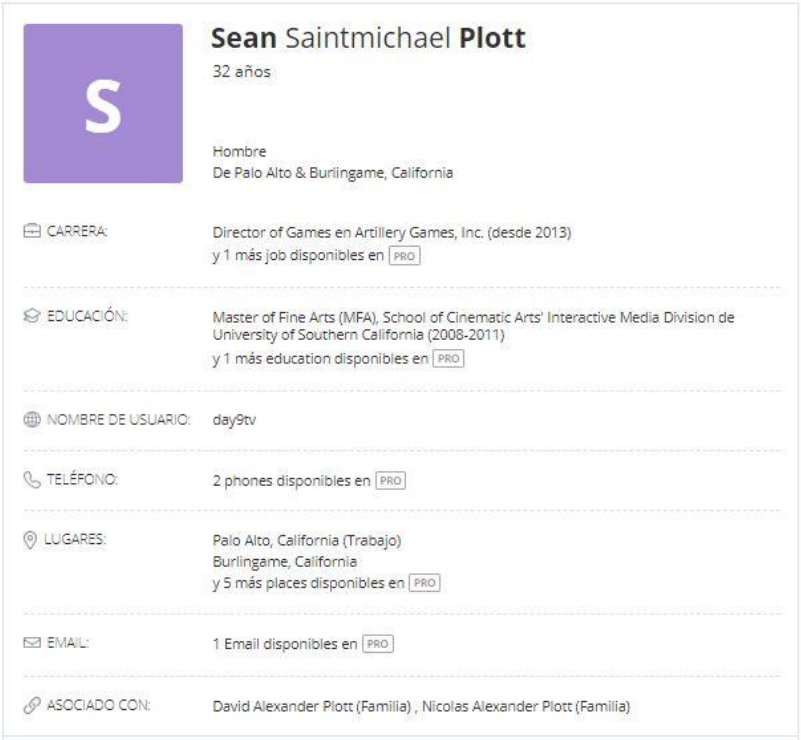

Figura 37: Búsqueda en Pipl

Otra forma de extraer información personal es la plataforma Yasni. Esta página busca a partir de un nombre todas las páginas web donde este aparece, pudiendo así encontrar enlaces a redes sociales, noticias o artículos en los que se menciona dicha persona, documentos, y todo tipo de enlaces relacionados. Sin embargo, toda esta información necesitará de un posterior análisis para determinar cuáles de estos enlaces se corresponden realmente con la persona a la que estamos buscando y no otra distinta que puede llamarse igual.

También puede resultar interesante la aplicación ThatsThem. Su funcionamiento es similar al de los otros sitios web comentados anteriormente, pero añade información adicional interesante como direcciones IP o un test de personalidad en la que puntúa el nivel de riqueza de la persona o si es propenso a realizar donaciones, compras online, viajar, adquirir nuevas tecnologías, o cuidar el medio ambiente.

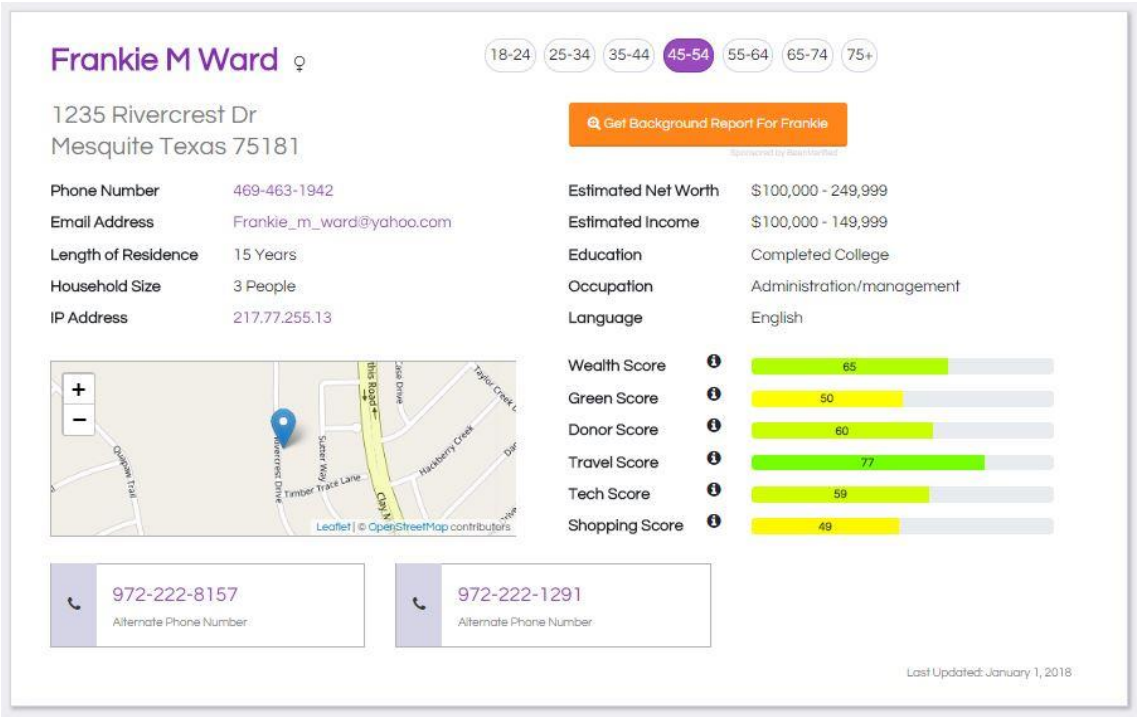

Figura 38: Ejemplo ThatsThem

8.3.2. Teléfonos y Emails

Además de estos metabuscadores, también existen páginas web dedicadas a la búsqueda de datos personales como emails o teléfonos. Para los números de teléfono, existen varios directorios en Internet como las Páginas Blancas o España-Directorio, donde encontramos miles de personas con su correspondiente número de teléfono y lugar de residencia.

Nombre	Número de teléfono	Ciudad
[redacted]	921-443289 / 921443289	(40006) Segovia
	922-181025 / 922181025	(38711) La Polvacera, Santa Cruz De Tenerife
	918-151877 / 918151877	(28692) Villafranca Del Castillo-La Mocha Chica, Madrid
	917-084565 / 917084565	(28230) Molino Hoz, Madrid
	952-751433 / 952751433	(29313) Villanueva Trabuco, Málaga

Figura 39: Directorios de teléfono

En el caso de los emails, existen varias herramientas web que nos permiten conocer si alguna dirección de correo electrónico se ha visto comprometida en algún leak conocido. Estos leaks se producen cuando un atacante obtiene datos de un sistema el cual ha sido comprometido previamente, explotando alguna vulnerabilidad de este. De estos ataques se suelen extraer datos, normalmente de empleados o usuarios, que son vendidos o difundidos públicamente, encontrándose muchos de ellos en la Deep Web. Uno de los casos más conocidos ocurrió en 2012 con LinkedIn, cuando se vendieron más de 167 millones de cuentas de sus usuarios con sus respectivas contraseñas.

Figura 40: Bases de datos con cuentas robadas en la darknet (LinkedIn)

La herramienta de este tipo más conocida es ;-- have i been pwned?. En su página web no solo podemos averiguar si un email se encuentra comprometido, si no que nos ofrece información sobre los leaks más recientes o importantes que se han producido.

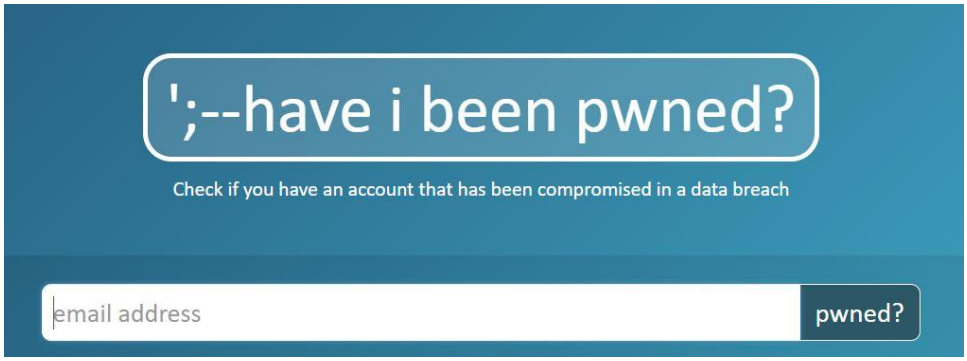

Figura 41: He sido hackeado?

Figura 42: Email comprometido

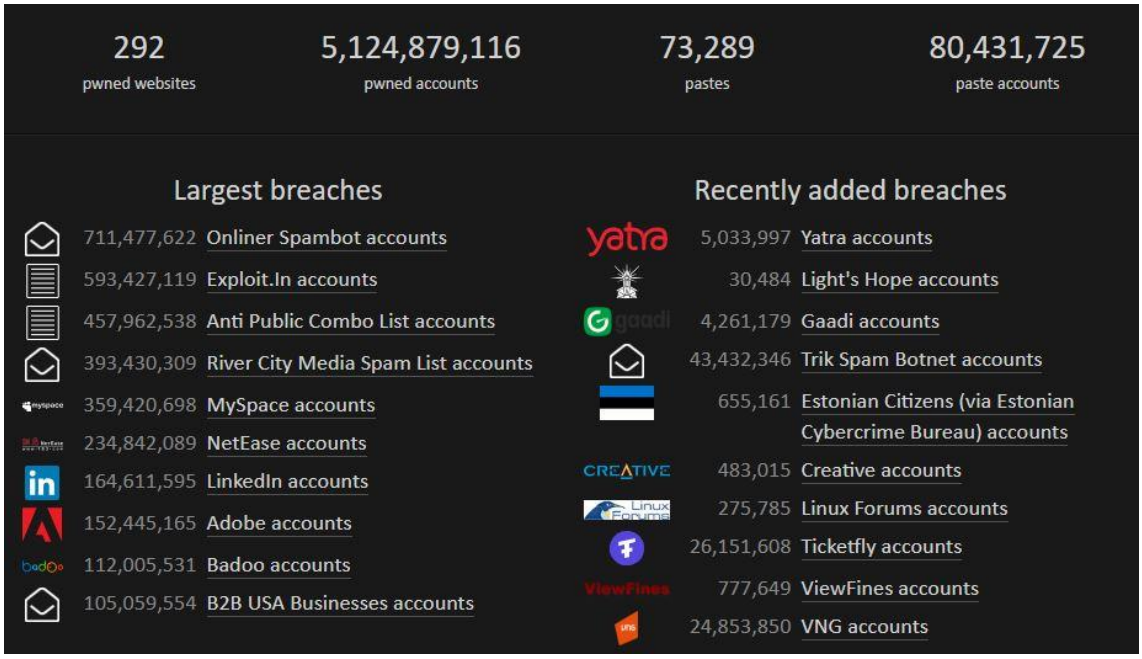

Figura 43: Información sobre leaks.

8.3.3. Geolocalización

La localización es uno de los principales datos que obtenemos al realizar búsquedas en redes sociales u otras aplicaciones, ya que muchas de estas solicitan al usuario compartirla, lo que muchos aceptan sin darse cuenta del riesgo que supone.

Existen varias herramientas que utilizan esta información sobre localizaciones que es difundida por Internet para analizarla y presentarla de una forma más accesible y visual. Además de la herramienta TinfoLeak vista anteriormente, herramientas como GeoSocial Footprint nos permiten obtener un mapa de calor de todas las ubicaciones compartidas por el usuario.

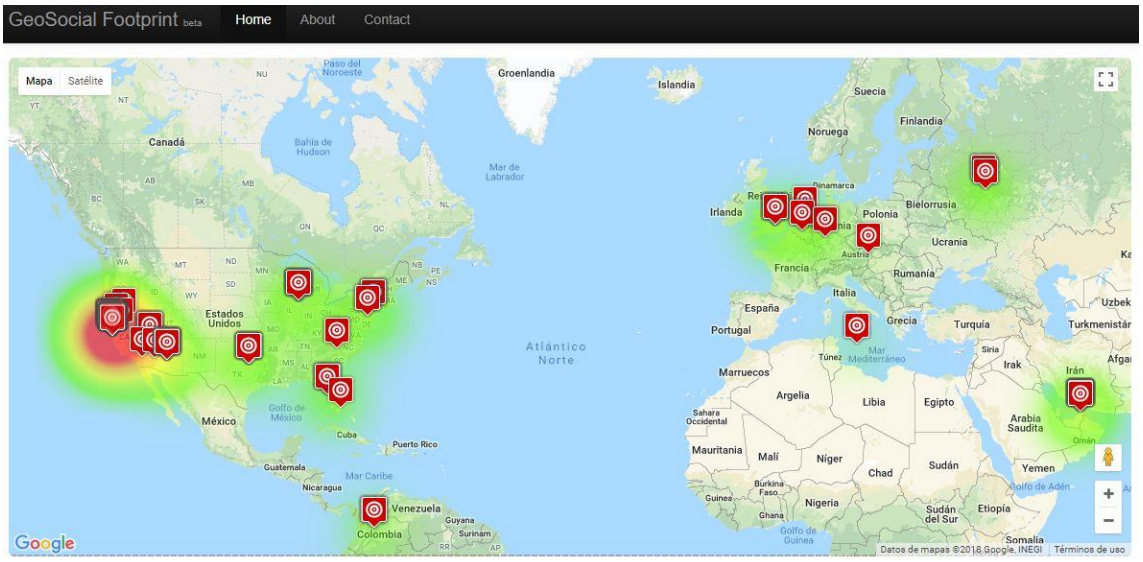

Figura 44: GeoSocial FootPrint

Es muy utilizada también la herramienta de ingeniería social Cree.py, desarrollada en Python en 2013, que permite geolocalizar a los usuarios de servicios web como Twitter, Flickr, Instagram y Google+, a partir de la información GPS de teléfonos móviles, tweets con ubicación activada y triangulación basada en la IP desde la que se realiza una publicación. De esta forma consigue extraer toda la información de localizaciones y de fechas de estas cuentas, para posteriormente generar bases de datos en las que visualizar las coordenadas obtenidas.

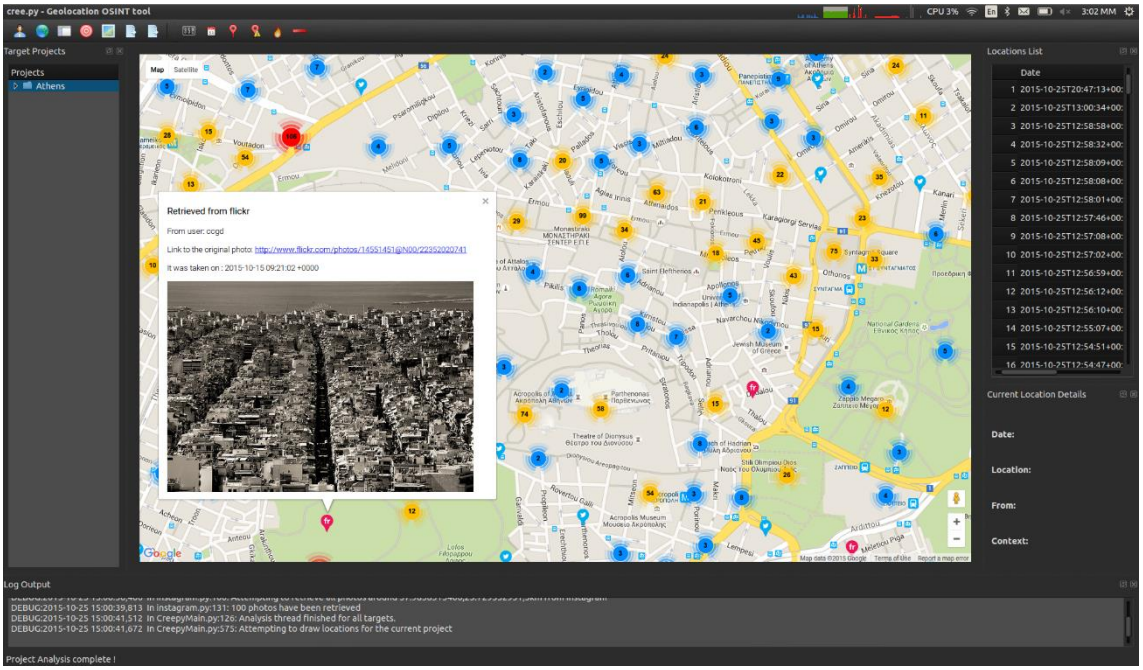

Figura 45: Aplicación Cree.py

8.3.4. Imágenes

Normalmente nuestro usuario en Internet suele estar acompañado de una foto de perfil que ayude a identificarnos. Tenemos a nuestra disposición varias herramientas que nos permite utilizar estas fotos para buscar todos aquellos sitios web en la que esta pueda aparecer. Páginas como TinyEye nos permiten subir una imagen o insertar una URL a la misma, mostrando como resultado los enlaces de todos los sitios web donde ha sido encontrada.

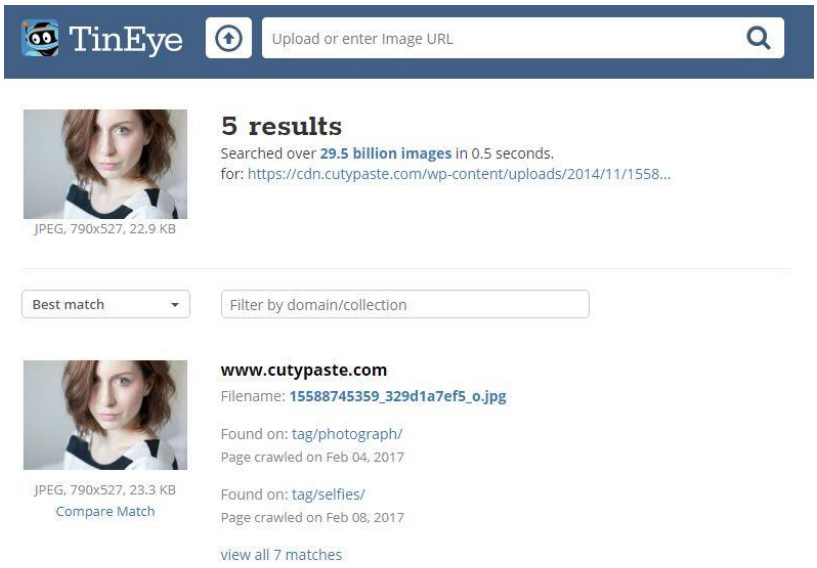

Figura 46: Ejemplo TinEye

Por su parte Google incluye un modo de búsqueda por imágenes, cuyo funcionamiento es similar al de buscadores como TinEye. Con estos buscadores de imágenes podemos identificar las distintas páginas donde se puede encontrar la persona, además de poder identificar lugares en los que las fotos han sido tomadas para establecer ubicaciones.

A continuación, se muestra un ejemplo de cómo podemos identificar la localización desde la que se ha realizado la siguiente foto, mediante una búsqueda por imágenes en Google.

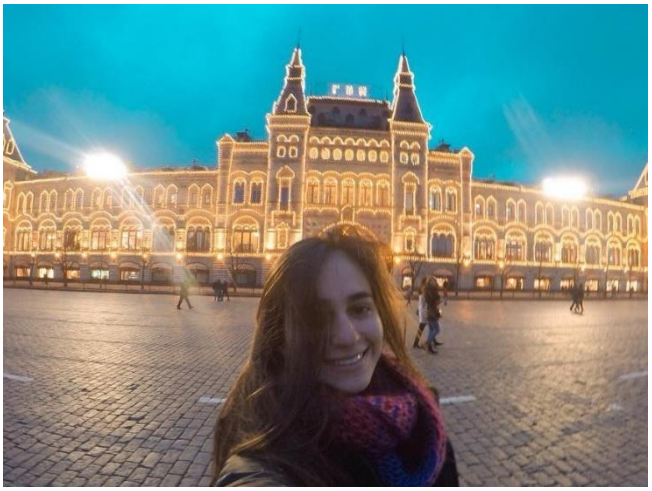

Figura 47: Ejemplo de imagen

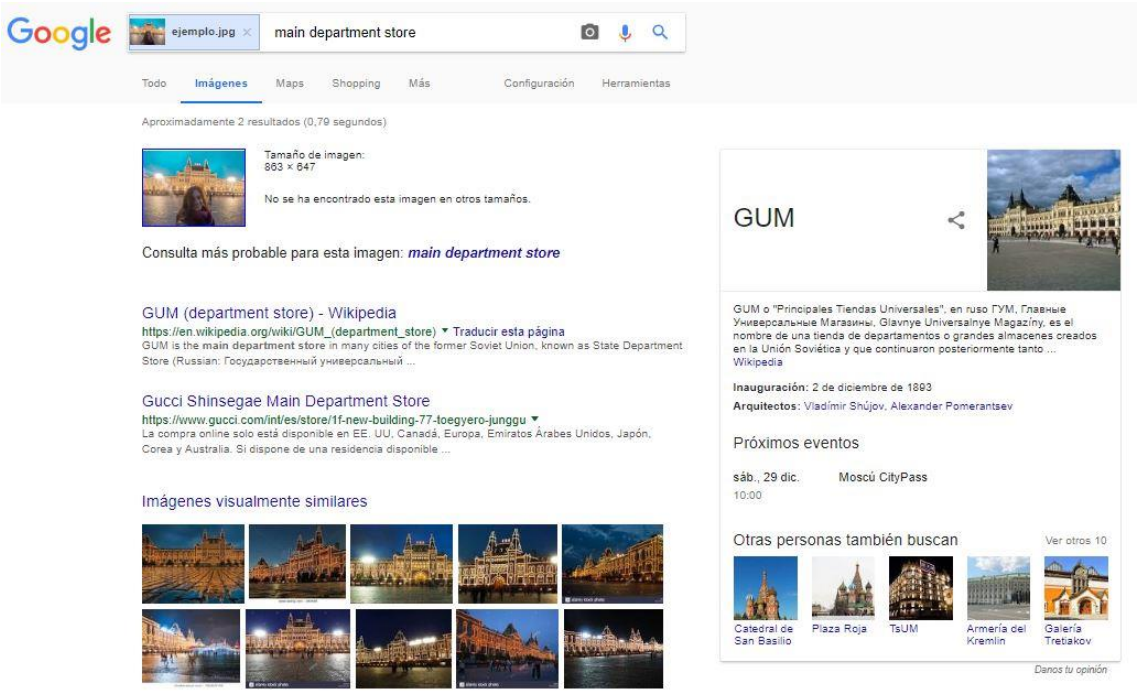

Figura 48: Búsqueda por imagen en Google

8.4. Herramientas de análisis de datos corporativos.

En Internet no solo podemos encontrar datos sobre personas en fuentes abiertas. También han sido desarrolladas aplicaciones que nos ayudan a extraer información sobre instituciones, que pueden ser utilizadas por las mismas empresas para realizar test de intrusión, o por atacantes para conocer algunas de las vulnerabilidades de estas.

8.4.1. Maltego

Maltego es un software utilizado para la inteligencia en fuentes abiertas y análisis forense, desarrollado por Paterva. Se centra en proporcionar al usuario una biblioteca de transformaciones para el descubrimiento de datos en fuentes

abiertas, y visualizar esa información en formato de gráfico, adecuado para su análisis y minería de datos.

La herramienta nos permite crear entidades personalizadas, para representar cualquier tipo de información, además de los tipos de entidades básicas que integran la aplicación. Las principales entidades con las que podemos trabajar son:

- **Personas** (nombres o correos electrónicos)
- **Redes sociales**
- **Compañías**
- **Organizaciones**
- **Sitios Web**
- **Infraestructura de Internet**
- **Documentos**

El principal objetivo de la aplicación es analizar las relaciones existentes entre la información relativa a estas entidades que se encuentra pública en Internet.

Entre sus fuentes de datos se encuentran registros de DNS, registros de Whois, motores de búsqueda, redes sociales, varias APIs online y diversos metadatos.

Podemos encontrar la aplicación de forma gratuita instalada en Kali Linux, aunque puede ser descargada en otras plataformas a través del siguiente enlace:

https://www.paterva.com/web7/downloads.php

En esta ocasión realizaremos un ejemplo en el que crearemos todas las transformaciones posibles a partir del dominio de la Universidad de Alcalá de Henares: *uah.es*.

Para ello añadiremos un nuevo proyecto, y crearemos dentro de este una nueva entidad, que será el dominio que queremos analizar.

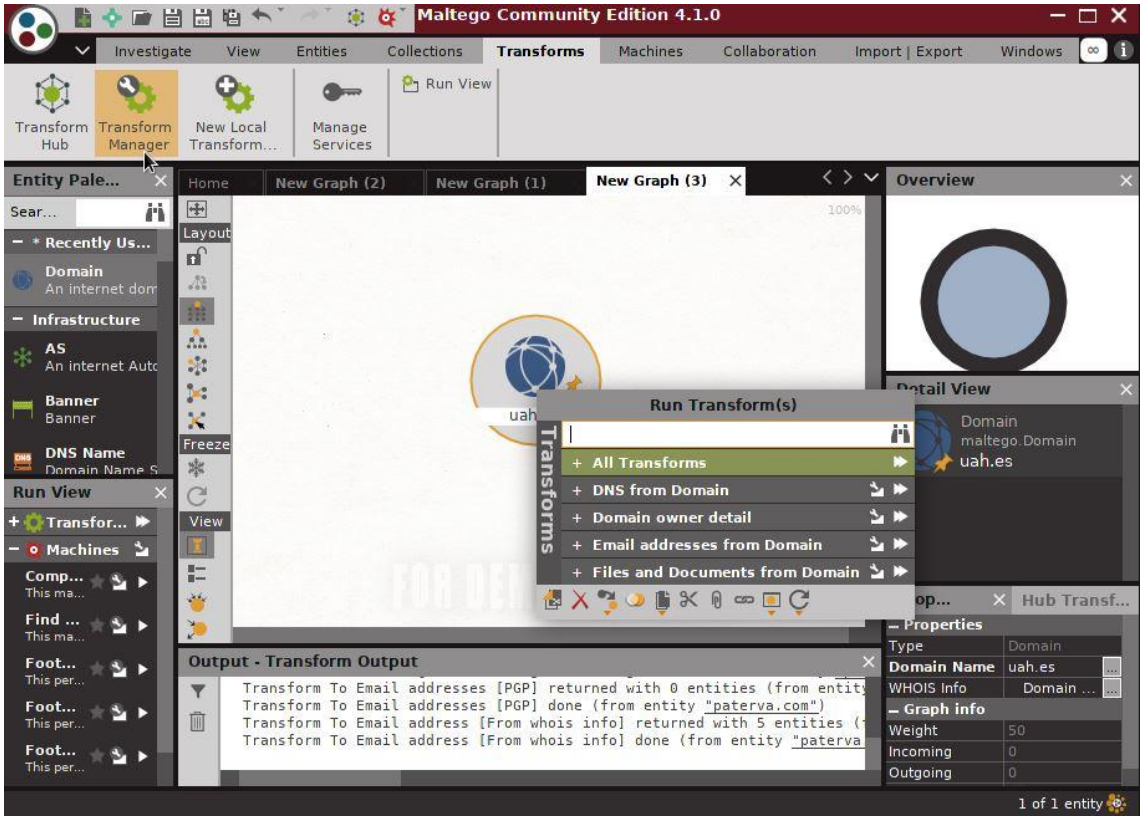

Figura 49: Crear gráfico en Maltego

Al ejecutar todas las transformaciones para ese dominio, Maltego generará un grafo con toda la información encontrada sobre subdominios, direcciones de correo, servidores DNS o páginas web asociadas al dominio uah.es.

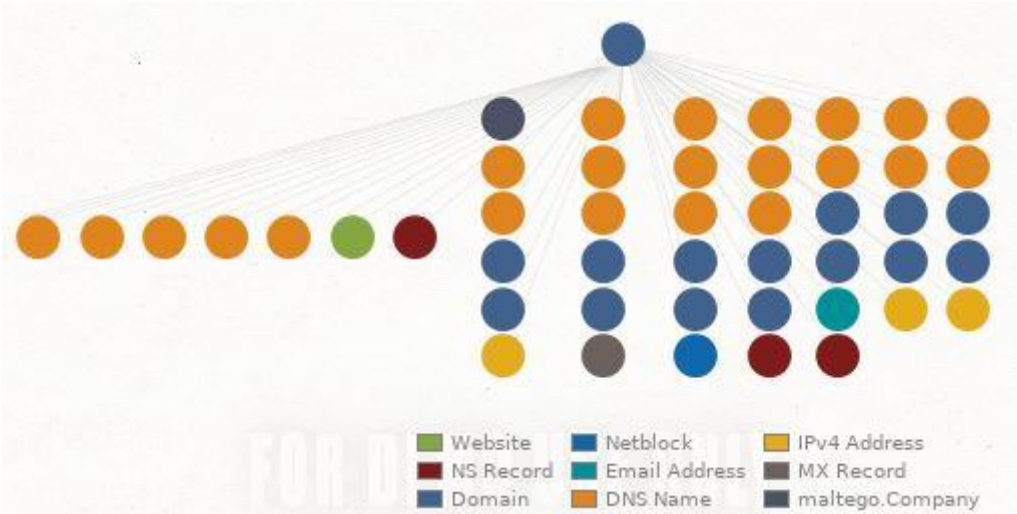

Figura 50: Grafo generado a partir del dominio uah.es

De la misma forma, podríamos seleccionar un solo tipo de transformación, para poder visualizar con mayor claridad, por ejemplo, el nombre de los servidores DNS utilizados.

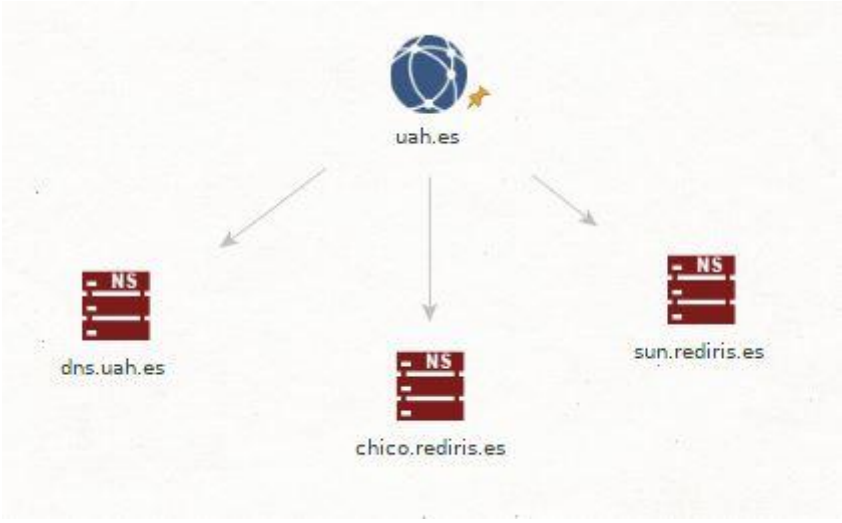

Figura 51: Servidores DNS en uah.es

8.4.2. TheHarvester

TheHarvester es una herramienta que recopila información pública de diferentes medios como buscadores o redes sociales:

- Direcciones de correo electrónico.
- Subdominios.
- Hosts virtuales.
- Puertos.
- Información de empleados.

Se trata de una aplicación desarrollada en Python la cual se puede encontrar instalada por defecto en máquinas con Kali Linux o descargar desde GitHub:

https://github.com/laramies/theHarvester

Cuando iniciemos la herramienta introduciendo el comando “theharvester” en la consola, se nos desplegará un menú de ayuda con todas las opciones disponibles.

Figura 52: TheHarvester

Para probar su funcionamiento, realizaremos una búsqueda de toda la información que pueda encontrar en todas las fuentes de datos disponibles (Google, Bing, LinkedIn, Twitter, etc.) del dominio uah.es

Para ello deberemos introducir el siguiente comando:

theharvester -d uah.es -b all

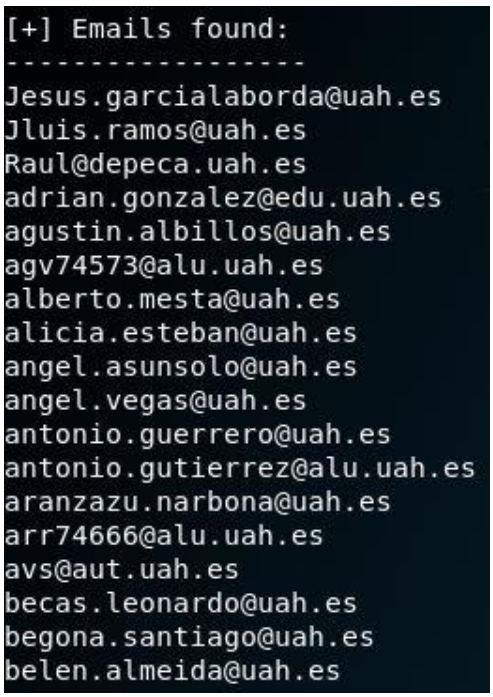

Figura 53: Direcciones de correo en uah.es

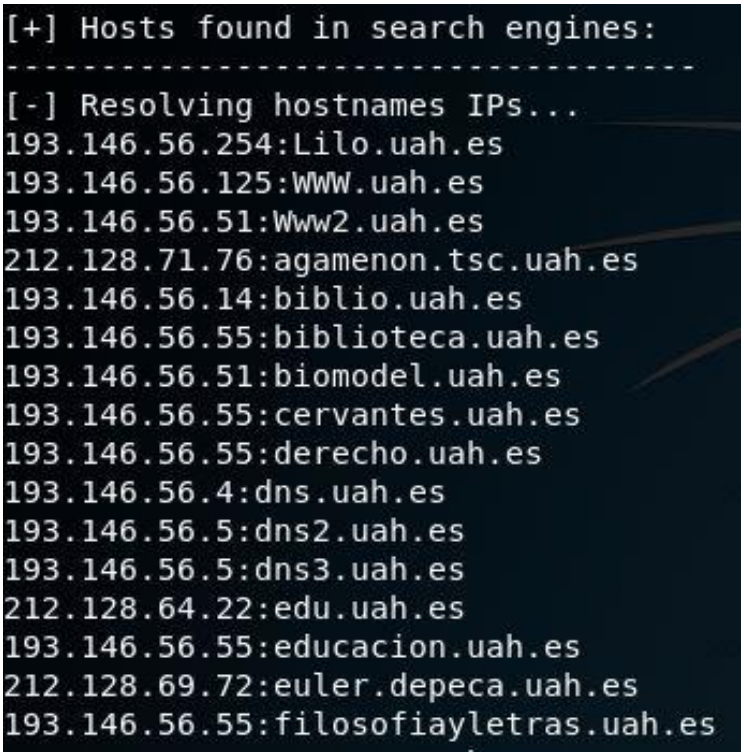

Figura 54: Subdominios de uah.es

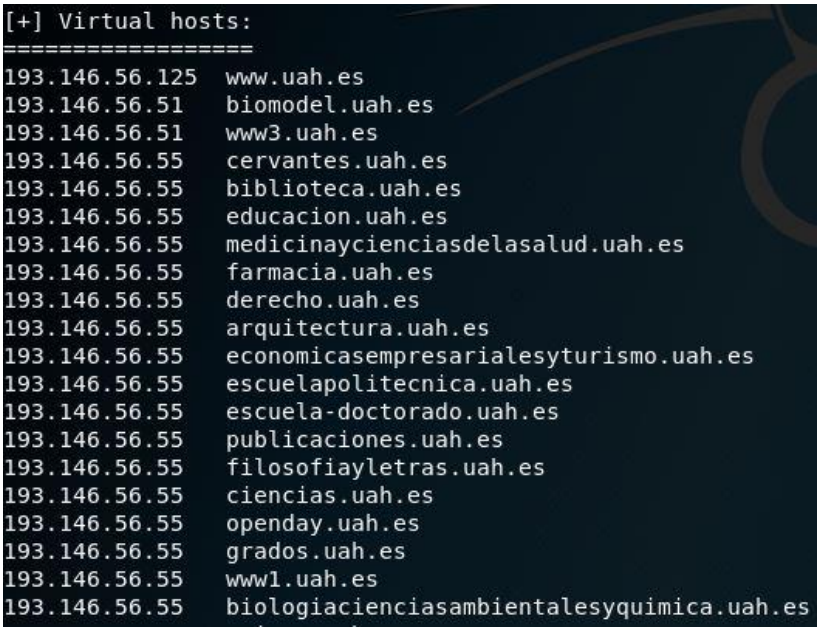

Figura 55: Hosts virtuales en uah.es

De esta forma, con TheHarvester podemos conocer la estructura de diferentes dominios, así como las direcciones de correo electrónico de estos dominios que podrían ser utilizados para enviar ataques personalizados a todos los empleados de una organización mediante ingeniería social, aumentando la superficie de ataque y las probabilidades de éxito.

8.4.3. Hunter

Como podemos observar en la Figura 53, las direcciones de correo electrónico suelen seguir un formato, en este caso formado por nombre.apellido1@uah.es.

Estos formatos en los emails suelen ser utilizados por muchas organizaciones para proporcionar a sus empleados un correo electrónico de empresa.

Con herramientas como TheHarvester podemos analizar los emails obtenidos para realizar tablas con las distintas cuentas que poseen los empleados en la empresa, y posteriormente intentar acceder a información confidencial de la organización a través de alguna de ellas.

La aplicación web Hunter, disponible en https://hunter.io/, permite encontrar y verificar estos formatos en las direcciones de correo electrónico de un dominio, a la vez que nos provee de los enlaces en los que se ha encontrado cada uno de esos correos.

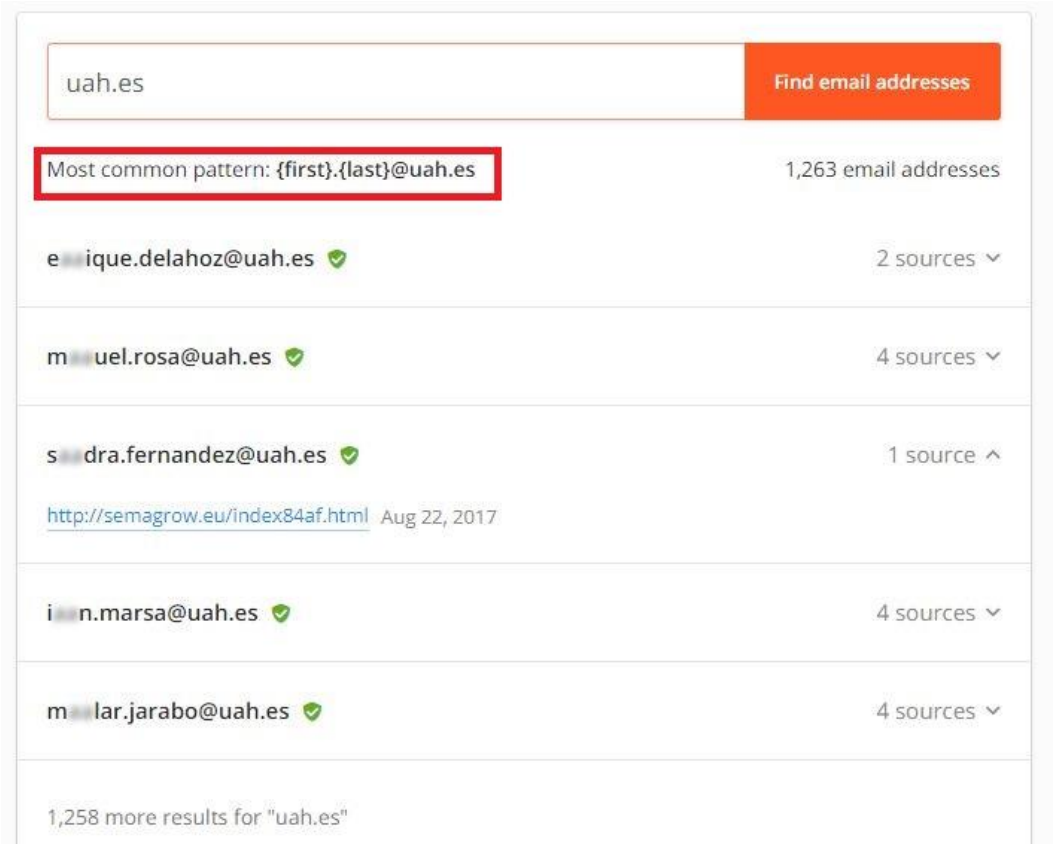

Figura 56: Formatos de correo en uah.es

8.4.4. FOCA

FOCA (Fingerprinting Organizations with Collected Archives) es una herramienta utilizada principalmente para extraer y analizar los metadatos e información oculta en los documentos que examina, los cuales suelen estar situados en páginas web.

Estos documentos pueden ser de varios tipos, siendo los más comunes los archivos de Microsoft Office, Open Office o ficheros PDF.

La búsqueda se realiza utilizando los motores de búsqueda de Google, Bing y DuckDuckGo, cuya unión permite que se consigan un mayor número de documentos.

La herramienta extrae los metadatos de estos ficheros y los analiza para saber quién los creó, modificó, el tipo de software que lo genera, que documentos han sido creados desde cada equipo, que servidores y los clientes que se pueden inferir de ellos.

Una vez instalada, podemos crear un nuevo proyecto a través del icono de la esquina superior izquierda. Tras situarnos en la ventana del proyecto, debemos introducir el nombre de dicho proyecto, el dominio que queremos analizar, y la carpeta donde lo queremos guardar. En este ejemplo, utilizaremos el dominio de la Escuela Politécnica de la Universidad de Alcalá.

Figura 57: Creación de proyecto con FOCA

Una vez creado el proyecto, elegimos la extensión del fichero que queramos buscar (en este caso PDF) y comenzaremos la búsqueda.

A medida que va encontrando los archivos, estos se nos irán mostrando en la pantalla, pudiendo parar la búsqueda en cualquier momento.

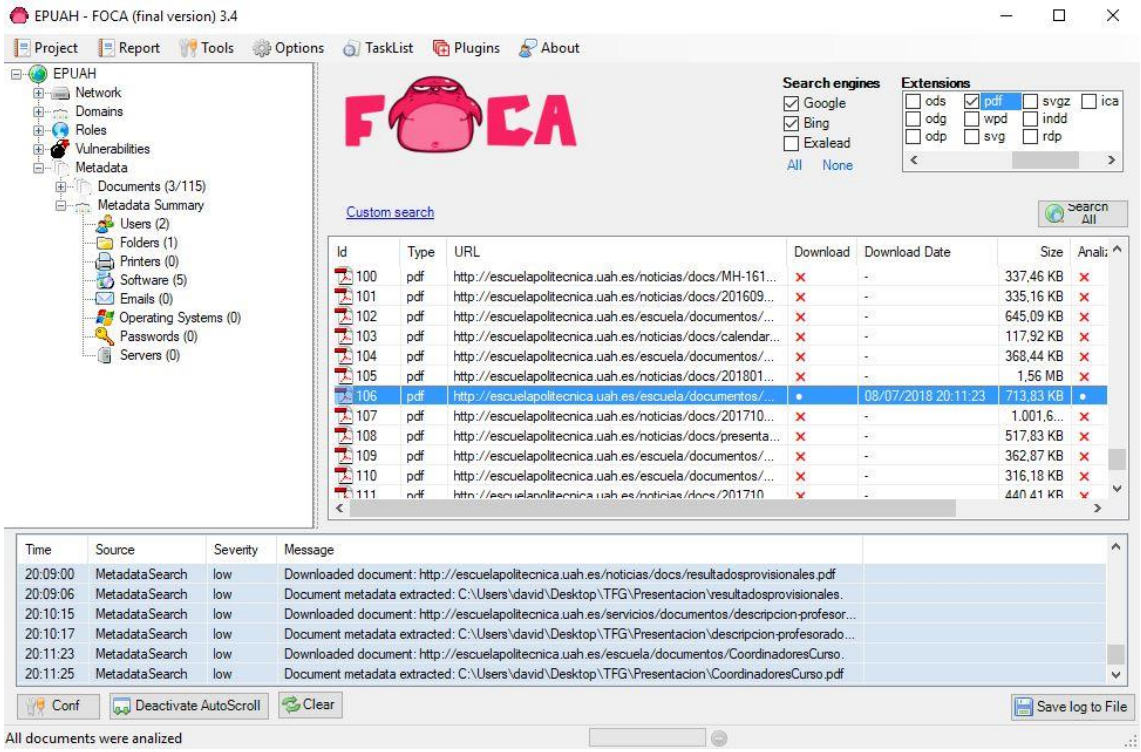

Figura 58: Búsqueda de archivos PDF

Cuando la búsqueda haya concluido, podremos seleccionar un archivo para su descarga en la carpeta del proyecto. Una vez descargado, extraemos los metadatos del archivo.

Attribute	Value
File Information	
URL	http://escuelapolitecnica.uah.es/escuela/documentos/CoordinadoresCurso.pdf
Local path	C:\Users\david\Desktop\CoordinadoresCurso.pdf
Download	Yes
Analized	Yes
Download date	08/07/2018 20:23:09
Size	713,83 KB
Users	
Username	SONIA
Dates	
Creation date	08/06/2018 11:34:59
Modified date	08/06/2018 11:34:59
Other Metadata	
Application	Microsoft: Print To PDF
Application	Microsoft Office XP
Title	Microsoft Word - 20171027Composicion CCurso
Software	
Microsoft: Print To PDF	
Microsoft Office XP	

Figura 59: Metadatos del archivo

De esta forma podemos observar el propietario del fichero, las fechas de creación y modificación del mismo o las aplicaciones utilizadas.

Además, descargando estos archivos podemos acceder a ciertos datos como nombres de profesores con sus respectivas direcciones de correo u otros tipos de datos que pueden resultar útiles.

8.4.5. Shodan

Shodan es un motor de búsqueda de servicios que permite al usuario buscar equipos conectados a Internet a través de una gran variedad de filtros. A diferencia de otros buscadores convencionales, Shodan se caracteriza por buscar más allá de servicios con interfaz web.

El buscador lee las cabeceras de los servicios para obtener información acerca de ISP, hostnames, países, puertos, servicios, protocolos, etc.

Puede utilizarse de manera gratuita, pero la adquisición de una licencia permite el acceso a la API, aumenta los resultados de búsqueda, elimina limitaciones de consultas diarias y proporciona acceso a todos los filtros de búsqueda.

Hoy en día es utilizado por hackers para localizar y acceder a infraestructuras críticas a través de los sistemas SCADA (control de supervisión y adquisición de datos) utilizados para gestionar dichas infraestructuras en tiempo real.

Muchos de estos sistemas están desprotegidos al utilizar credenciales de autenticación por defecto, servicios web sin securizar, falta de mantenimiento o exceso de privilegios a determinados usuarios.

De esta manera, Shodan nos ofrece una herramienta para la recopilación de información de forma activa de un objetivo.

Entre los distintos filtros que podemos utilizar están:

- *City*: Dispositivos ubicados en una determinada ciudad.
- *Country*: Terminales cuya localización coincida con el país buscado.
- *Geo*: Búsqueda de direcciones IP por coordenadas.
- *Hostname*: Búsqueda por nombre de dispositivo.
- *OS*: Buscar por sistema operativo.
- *Port*: Búsqueda por el número de puerto indicado.
- *Net*: Busca información vinculada a una IP dada. Puede realizarse de dos formas:
  - IP directa: por ejemplo, net:111.11.11.11
  - Rango de subred: por ejemplo, net:111.11.11.0/24
- *Before/After*: Resultados en un rango de tiempo determinado
- *Org*: Búsqueda por nombre de organización
- *Product*: Búsqueda por producto, por ejemplo, MySQL.

El uso de Shodan por tanto, es muy parecido a lo que hemos visto con los dorks en Google Hacking, y a partir de estos comando podemos acceder por ejemplo, a servicios con contraseñas y usuarios por defecto a través de country:ES “default password”

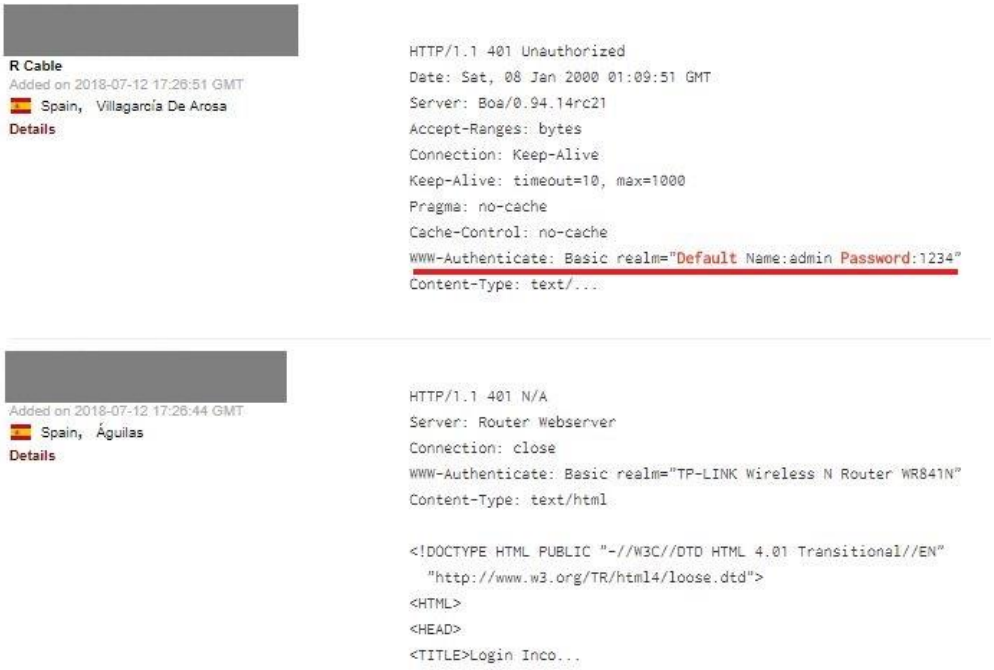

Figura 60: Ejemplo de búsqueda en Shodan

A pesar de que buscar en Shodan no es delito, debemos tener cuidado con el acceso no permitido a algunos de los resultados que este ofrece, sobre todo en el acceso e interacción con los sistemas SCADA

9. Frameworks

Los distintos frameworks dedicados a la inteligencia en fuentes abiertas que podemos encontrar suponen una herramienta muy útil para la extracción de datos en fuentes abiertas ya que agrupan las funcionalidades de varias aplicaciones para facilitarnos el trabajo de investigación. En este apartado analizaremos algunos de los frameworks más utilizados.

9.1. OSINT framework

Esta herramienta se centra en ayudar a las personas a encontrar recursos y herramientas OSINT gratuitos. Para ello, clasifica los distintas herramientas y recursos disponibles según su temática y la muestra en forma de árbol.

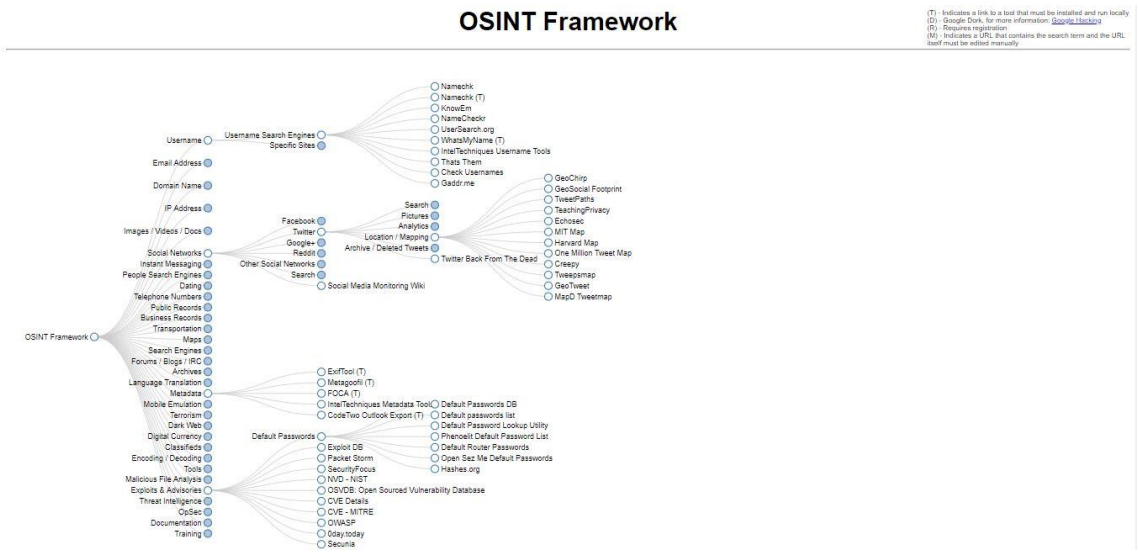

Figura 61: OSINT Framework

En la parte superior del árbol podemos encontrar herramientas que nos pueden ayudar a la extracción de datos personales como nombres de usuarios, direcciones de correo, análisis de recursos, direcciones IP, redes sociales, motores de búsqueda de personas, números de teléfono, etc.

La parte inferior, sin embargo, clasifica fuentes de información orientadas a contenido no personales. Aquí podemos encontrar fuentes y herramientas dedicadas a la búsqueda de vulnerabilidades, análisis de código, análisis de malware, metadatos, etc.

Sin embargo, lo que convierte a este framework en una herramienta tan útil, es su facilidad de uso, y el hecho de que nos permite hacernos una idea de la cantidad de información e investigaciones diferentes que podemos realizar dentro del campo OSINT.

9.2. Intel Techniques

Cuando hablamos de frameworks de OSINT, el más extendido actualmente es Intel Techniques. Creado Michael Bazzell, no solo nos permite conocer las distintas herramientas disponibles para las distintas fuentes de información

como hace OSINT Framework, si no que implementa una interfaz que recoge todas estas herramientas para su uso directo a través del framework.

Como punto a tener en cuenta, todas las búsquedas realizadas con Intel Techniques quedarán almacenadas en sus bases de datos, por lo que deberemos tener cuidado a la hora de extraer información personal.

Todas estas herramientas pueden ser encontradas en la ventana de Tools en la página web de la aplicación.

Figura 62: Intel Techniques

Una vez dentro podemos observar los distintos tipos de recursos y fuentes de información a partir de los cuales podemos extraer información.

Figura 63: Fuentes de información en Intel Techniques

Como podemos observar, entre las opciones disponibles, tenemos la posibilidad de seleccionar OSINT Links, donde tendremos enlaces directos a

una gran cantidad de herramientas para cada una de las fuentes disponibles, así como enlaces a los buscadores personalizados que ofrece este framework.

Estos buscadores personalizados, permiten buscar toda la información disponible en fuentes públicas, a partir de los datos que aparecen en la Figura 64.

Por ejemplo, podemos elegir la opción de búsqueda por email, donde introduciendo una dirección de correo electrónico, esta será buscada en todas las herramientas de las que dispone la aplicación, y que podemos ver en la siguiente figura.

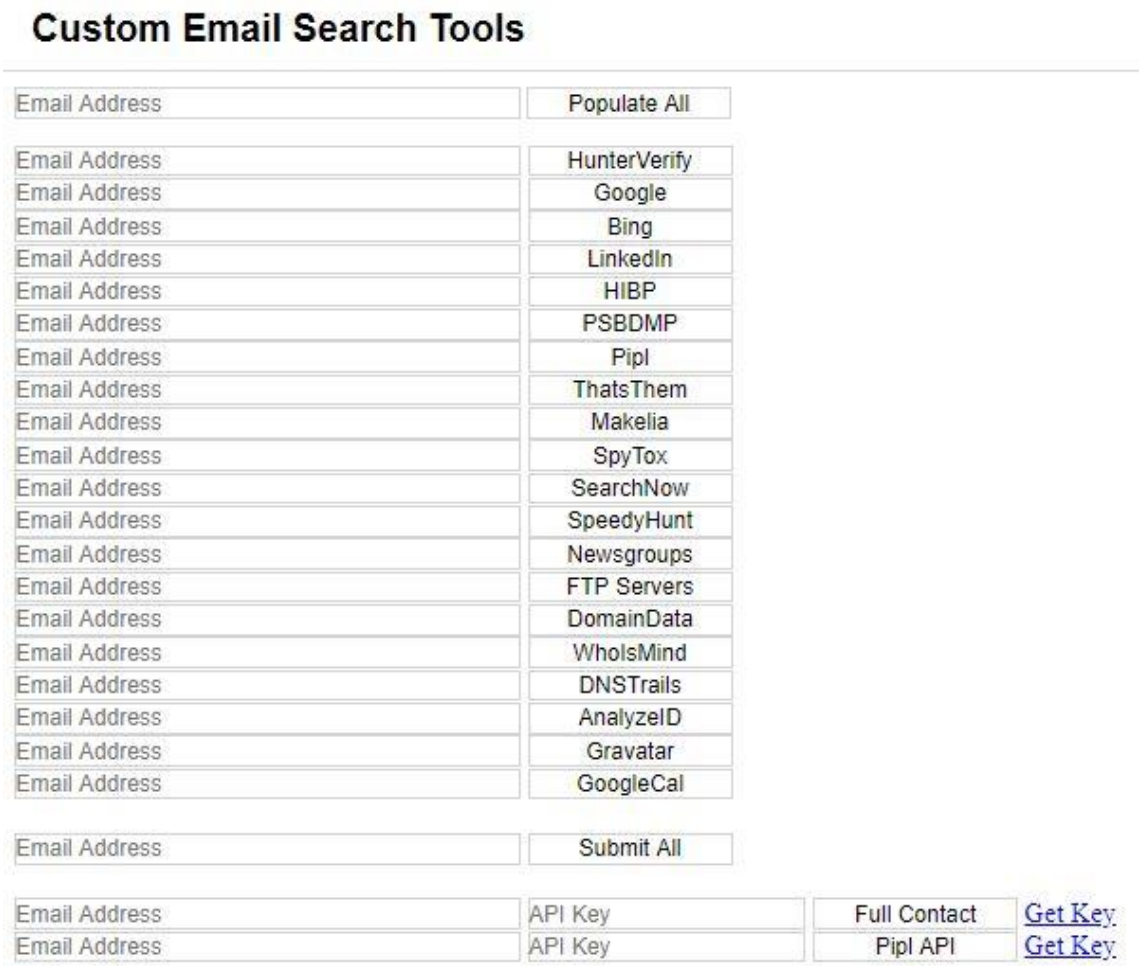

Figura 64: Búsqueda por Email

En resumen, podemos decir que Intel Techniques es una de las herramientas más completas que podemos encontrar para la extracción de datos en fuentes abiertas, ya nos ofrece una gran cantidad de posibilidades para la búsqueda de información a partir de diversas fuentes, reuniendo en una sola web las aplicaciones OSINT más importantes, facilitando la investigación y el uso de estas al ser directamente implementadas en sus buscadores personalizados.

10. Conclusión

Tras la realización del trabajo, y habiendo estudiado las distintas formas de extraer datos de fuentes abiertas, podemos llegar a las siguientes conclusiones:

- No es de extrañar el crecimiento que está teniendo el campo de la ciberinteligencia, especialmente las técnicas de OSINT, dentro del ámbito de la ciberseguridad. La gran cantidad de información útil que se puede obtener en fuentes públicas puede ser diferencial en todo tipo de actividades de prevención de riesgos e investigaciones de objetivos.
- Una vez que hemos estudiado algunas de las herramientas más utilizadas, podemos concluir en que ninguna de ellas es perfecta. Si bien cada una de ellas es muy útil para cada tipo de fuente, no todas acaban de ofrecer resultados 100% fiables.
- En relación con el punto anterior, las distintas herramientas disponibles pueden ofrecernos una gran cantidad de datos, pero lo que es realmente importante es como procesamos toda esta información y la inteligencia que le aportemos para saber que datos debemos recoger y cuales excluir.
- La existencia de tantos tipos de fuentes abiertas de información y herramientas disponibles para su explotación nos lleva a pensar la falta de privacidad del mundo en el que vivimos, en muchos casos debido a la poca concienciación o interés de la población, y que podría suponer un riesgo importante si toda esta información que publicamos cae en malas manos.

11. Trabajo a futuro

Con la creciente importancia de la inteligencia en fuentes abiertas, cada vez son más los interesados en sacar provecho de la información que puede conseguirse con OSINT, lo que impulsará la creación de nuevas fuentes de información y herramientas dedicadas a la extracción de datos.

Sin embargo, la tarea pendiente que tendrán estas nuevas aplicaciones será la de aumentar la fiabilidad de la información obtenida. A día de hoy existen herramientas muy útiles para la búsqueda de información, pero muchas de ellas son herramientas en desarrollo que son elaboradas en la mayoría de los casos por una sola persona o por pequeños grupos de personas, motivos por los que su funcionamiento no siempre es óptimo.

Con el apoyo de grandes organizaciones, se podrán crear herramientas cada vez más precisas y que a su vez puedan abarcar mayores cantidades de datos. Además, los avances en el campo de la inteligencia artificial podrían combinarse con estas herramientas para poder automatizar el proceso de extracción de datos, sobre todo a la hora la generación de inteligencia, para que estas puedan encargarse de la correcta selección de la información obtenida.

12. Bibliografía

[1] Asier Martínez (28 de mayo de 2014), *"OSINT – La información es poder"*, Blog Instituto Nacional de Ciberseguridad de España. Online: https://www.certsi.es/blog/osint-la-informacion-es-poder

[2] Selva Orejón (8 de febrero de 2018), *"OSINT o como pescar en la red"*, Blog Inesdi, Digital Business School. Online:

https://www.inesdi.com/blog/osint-pescar-red/

[3] Travis Lishok (3 de abril de 2018), *"Part I: An Introduction To OSINT Research For Protective Intelligence Professionals"*, Blog Instituto Nacional de Ciberseguridad de España. Online:

https://www.protectiveintelligence.com/blog/

[4] Germán Realpe (11 de agosto de 2016), *"Fuentes abiertas: Herramientas para hacer inteligencia en la red"*, Blog Enter.co. Online:

http://www.enter.co/chips-bits/seguridad/herramienta-inteligencia-internet/

[5] Álvaro Vállega y Jorge Alcaín (2018), *"OSINT: Atacando con la información pública"*, Taller Prosegur Ciberseguridad, Jornadas de Seguridad y Ciberdefensa Ciberseg, Universidad de Alcalá.

[6] Marco Varone, Daniel Mayer, Andrea Melegari (15 de marzo de 2016), *"OSINT: definition of Open Source INTelligence"*, Blog Expert System. Online:

https://www.expertsystem.com/what-is-osint/

[7] Paula Rochina (18 de octubre de 2016), *"Nuestra huella digital en Internet: ¿Hasta dónde saben de mí?"*, Blog Revista Digital INESEM. Online:

https://revistadigital.inesem.es/informatica-y-tics/huella-digital-internet/

[8] Marcos Polanco (21 de abril de 2016), *"La ciberinteligencia como habilitador de la ciberseguridad"*, Blog Magazcitum. Online:

http://www.magazcitum.com.mx/?p=3205#.W0ZmQtIzbIV

[9] Curso de inteligencia (16 de marzo de 2016), *"¿Qué es la Ciberinteligencia? La inteligencia en materia de Ciberseguridad"*, Blog ASINT 360º. Online:

http://www.asint360.com/que-es-la-ciberinteligencia-la-inteligencia-en-materia-de-ciberseguridad/

[10] Gabriel Bergel (28 de marzo de 2016), *"Las Fases de la Ciberinteligencia"*, Blog Eleven Path. Online:

http://blog.elevenpaths.com/2016/03/las-fases-de-la-ciberinteligencia.html

[11] Equipo InboundCycle (23 de marzo de 2014), *"Indexación: primer paso para aparecer en los buscadores", Blog de Inbound Marketing*. Online:

https://www.inboundcycle.com/blog-de-inbound-marketing/bid/194390/indexaci-n-primer-paso-para-aparecer-en-los-buscadores

[12] Juan Antonio Calles (2014), *"Open Source Intelligence y la unión de los mundos virtual y físico"*, Zink Security S.L. Online:

http://www.isaca.org/chapters7/Madrid/Events/Documents/Forms/AllItems.aspx?utm_referrer=direct%2Fnot%20provided

[13] Antonio González (2 de marzo 2012), *"Google Hacking (46 ejemplos): cómo consigue un hacker contraseñas usando Google. Google puede ser tu peor enemigo."*, Blog personal de Antonio González. Online:

https://antoniogonzalezm.es/google-hacking-46-ejemplos-hacker-contrasenas-usando-google-enemigo-peor/

[14] Daniel González, Jesús Alcalde (2018), *"OSINT, la verdad está ahí fuera"*, Blog ZeroLynx. Online:

https://www.zerolynx.com/downloads/OSINT_RootedCON_2018.pdf

[15] Marie Perod (2018), *"Pasado y presente de las redes sociales"*, Artículo en MuyInteresante Online:

https://www.muyinteresante.es/tecnologia/articulo/pasado-y-presente-de-las-redes-sociales-711496244493

[16] Yolanda Corral (5 de septiembre 2016), *"Qué es OSINT: fases, fuentes y herramientas"*, Blog de Yolanda Corral. Online:

https://www.yolandacorral.com/que-es-osint-fases-fuentes-herramientas/

[17] Enlaces a Herramientas OSINT, CiberPatrulla, Online:

https://ciberpatrulla.com/links/

[18] Páginas oficiales de las distintas herramientas analizadas.

Universidad de Alcalá
Escuela Politécnica Superior

ESCUELA POLITECNICA
SUPERIOR