ANÁLISIS DE TÉCNICAS DE RASPADO DE DATOS EN LA WEB – APLICADO AL PORTAL DEL ESTADO NACIONAL ARGENTINO

Roxana Martínez[1], Rocío Rodríguez[1], Pablo Vera[1], Christian Parkinson[1]

[1] Centro de Altos Estudios en Tecnología Informática (CAETI).
Universidad Abierta Interamericana (UAI), Ciudad Autónoma de Buenos Aires, Argentina
{Roxana.Martinez; Rocioandrea.Rodriguez; Pablomartin.Vera;
Christian.Parkinson}@uai.edu.ar

Resumen. Entender la importancia de los datos como herramienta es fundamental para el avance en la tecnología. Garantizar la calidad de los datos no es sencillo, por lo que es vital estar atentos a los procesos de recolección de los mismos. Este trabajo permitirá obtener datos precisos, actualizados y completos. Para ello, se analizan distintas herramientas de raspado de datos (web scraping) que existen en el mercado y se estudian las modalidades de extracción de datos de cada una de ellas en base al caso de estudio del sitio web del Estado Nacional Argentino (Ministerio de Modernización de Argentina), con el fin de extraer datos de los trámites y servicios disponibles para los ciudadanos.

Palabras Claves: Extracción de Datos, Scraping, Big Data, Datos Abiertos.

1. Introducción

La innovación en la tecnología informática se encuentra en pleno auge y cada vez es más estrecha su alcance y relación en los distintos ambientes que utiliza la población para los entornos tecnológicos.

Uno de los puntos a tener en cuenta es que los datos son el centro básico de transformación digital y que sólo pueden brindar su máximo potencial, si se exploran correctamente las innovaciones tecnológicas que los utilizan. "*Los datos ahora se han convertido en el activo más valioso del mundo, más que el petróleo*" [1].

Una de las características más sobresalientes en la actualidad es el concepto de "*algoritmos y la acumulación de grandes volúmenes de información, los así llamados Big Data. Se trata de un proceso de automatización que, lejos de remitir a la vieja idea de lo automático como repetición, genera incesantemente diferencia, establece rangos de acción, permite niveles cada vez mayores de interacción y, por ello mismo, suscita nuevas formas, sutiles y sofisticadas, de control social*" [2].

Con respecto a la innovación de los datos, existen varias observaciones a tener en cuenta, pero las más destacadas son "*las cuatro V de la innovación de los datos: volumen, la cantidad de datos; velocidad, la rapidez con que se crean; variedad, los tipos de datos involucrados; y veracidad, su precisión*" [3]. La comprensión de cada

uno de estos ítems es fundamental para un correcto aprovechamiento de la información.

Los datos abiertos posibilitan el conocimiento abierto, al que las personas pueden acceder sin restricciones, utilizándolo en forma libre y gratuita. Este paradigma busca generar soluciones que sean beneficiosas para los ciudadanos y generar un bien público que se lleve a cabo de manera colaborativa junto con datos abiertos. Los datos abiertos "*son una infraestructura básica para la creación de negocios y de productos y servicios. Para hacer un análisis de su utilidad hay que tener en cuenta que no es igual el acceso a los mismos y su difusión que su reutilización*" [4].

1.1 Web Scraping

El método de web scraping es una técnica utilizada mediante programas de software para extraer información de sitios web. "*Usualmente, estos programas simulan la navegación de un humano en la World Wide Web ya sea utilizando el protocolo HTTP manualmente, o incrustando un navegador en una aplicación*" [5].

Básicamente si se realiza la copia de datos de una página web y se almacenan en una base de datos, se considera que es un proceso de extracción de datos. Si, en lugar de hacerlo de forma manual, se utilizan robots o bots que automatizan todo el procedimiento anteriormente comentado, se hablará de "web scraping".

El web scraping está muy "*relacionado con la indexación de la web, la cual indexa la información de la web utilizando un robot y es una técnica universal adoptada por la mayoría de los motores de búsqueda. Sin embargo, el web scraping se enfoca más en la transformación de datos sin estructura en la web (como el formato HTML) en datos estructurados que pueden ser almacenados y analizados en una base de datos central*" [6].

1.2 Web Scraping versus Web Crawling

Existen dos técnicas que por lo general se suelen confundir. Si bien, poseen una relación entre ambas y parte de la técnica es similar, hay una diferencia bien marcada en cuanto a una metodología comparada con otra. En el caso de Web Scraping, se conoce como el "*raspado web*" o bien "raspado de datos" y en el caso de Web Crawling se conoce como el "*rastreo web*".

Particularmente el "*rastreo web*" tiene que ver con el proceso de lograr ubicar diversos datos en Internet con el fin de almacenar todas las palabras relevantes o bien palabras claves de búsqueda, y que las mismas se puedan alojar en una base de datos. Mediante esta técnica se tiene como resultado un pequeño repositorio de palabras en un almacenamiento de acceso, para luego lograr manipularlas. Por otra parte, cada una de las palabras extraídas, posee una identificación, que permite saber de qué link (hipervínculo) fue obtenido, por lo que esta técnica, además, permite que se guarden dichos links como parte del proceso de rastreo. Los mismos pueden complementarse con una exploración como parte de la indexación de la base de datos, enlazados a un tema particular. El proceso de rastreo web se lleva a cabo con un software que permite acceder a diversos sitios web, y luego, realizar una lectura de la página en

forma completa para crear un índice de motores de búsqueda. Los principales motores de búsqueda del mercado que utilizan esta técnica son: Google o Bing, los cuales poseen un programa de este estilo, lo que también se conoce más comúnmente como: “araña web” o "bot". Estos programas permiten generar un índice para luego poder realizar consultas contra el índice generado, para así localizar páginas web que coincidan con la consulta de filtros realizadas. Otro de los fines que se le puede dar a este tipo de método es un uso de minería de datos para el análisis de diversas propiedades estadísticas sobre los datos extraídos. Finalmente, se puede lograr un servicio más sofisticado de monitoreo de datos para generar un aviso o alerta a los usuarios que requieran determinada información del rastreo.

Como desafío de esta técnica se puede optar por un modelo de extracción de datos donde se busca información nueva o bien actualizada de manera que surja un comportamiento proactivo de esta actividad.

En resumen, el rastreo web es el método que realizan los motores de búsqueda, es decir, se busca cualquier tipo de información, en cambio, el raspado web está apuntado a determinados sitios web específicos para localizar datos determinados.

A continuación, se muestra la Figura 1 en la que se puede ver una comparativa de pasos básicos entre ambos métodos. En la parte izquierda, se muestra Web Crawler, se observa que se rastrea y se visitan los distintos sitios web para luego armar la lista y así lograr indexarla. Como paso final, se almacenan los datos en una base de datos, la cual es utilizada posteriormente. En la parte derecha, se muestra el Web Scraping, el cual permite analizar un sitio web y así recolectar los datos específicos que son interesantes a identificar del sitio web, como paso final, éstos se almacenan en diversos formatos como ser: XML, SQL o bien formato de archivo Microsoft Excel.

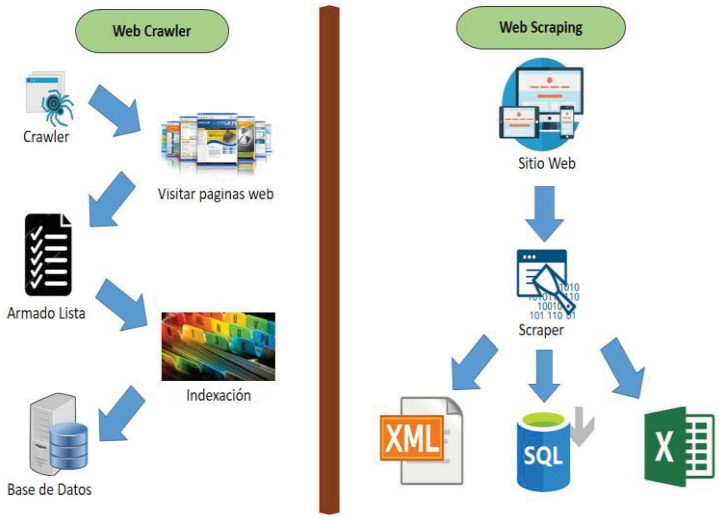

Fig. 1. Comparativa de los pasos, entre el método de Web Scraping y Web Crawling.

2 Herramientas de Raspado de Datos

Las herramientas de raspado web cumplen la función de extracción de información de sitios web, también se los conoce como herramientas de recolección/extracción de datos web. Esta técnica no requiere el método de copiar y pegar, sino que apunta a una forma de extraer los datos en un formato determinado para que luego sea accesible.

La idea central es localizar datos puntuales de los sitios web y almacenarlos para su posterior utilización. Es decir, de un sitio web, puede ser interesante obtener sólo algunos datos, y el resto de lo que se analiza puede ser descartado. Existen algunos servicios de monitoreo frente a los cambios de datos en sitios web, lo que permite mantener actualizado lo previamente recuperado con la técnica de web scraping.

Existen en el mercado diversas herramientas disponibles de raspado de datos. Mediante Google Trends se consideraron las más populares para generar una lista acotada que es la que se presenta en la Tabla 1. Cabe destacar que totas las herramientas listadas tienen licenciamiento gratuito o bien que permitían tener una versión trial para testearlas y compararlas entre sí.

Tabla 1. Herramientas de Raspado de Datos para extraer datos de sitios web.

Herramienta	Descripción
Import.io [8]	Elabora sus propios datasets mediante la importación de los datos de una página web específica. Raspado de miles de páginas web. No requiere de programación, ya que es una plataforma automatizada de extracción. Además, en su versión paga, posee varias funcionalidades más, como ser: Pericias, escalabilidad Integración e Informes.
Web Scraper [9]	Es un Plugin, extensión de Google Chrome llamado Web Scraper. Diseñado para quienes no tienen conocimiento de programación, posee funciones básicas de extracción de dato y no es tan amigable.
Dexi.io [10]	Esta herramienta permite la recopilación de datos de cualquier sitio web. Posee un editor basado en navegador para configurar rastreadores y extraer datos en tiempo real. Es bastante complejo y poco amigable su utilización, es utilizada por usuarios avanzados.
ParseHub [11]	Gestiona tareas de extracción de datos y administra páginas web que usan JavaScript y AJAX. Para el reconocimiento de documentos complejos utiliza aprendizaje automático. Posee una versión web y de escritorio.
Outwit Hub [12]	Se destaca por su interfaz, ya que es muy amigable, lo que la hace fácil de usar, y a su vez, posee grandes características de reconocimiento de datos. Es una herramienta genérica, con un amplio espectro de uso, que va desde la extracción de datos ad hoc sobre temas de investigación específicos hasta la toma extensiva diaria de datos en línea para poblar sitios web.
Scrapestorm [13]	Modo inteligente: basado en algoritmos de inteligencia artificial, esta herramienta identifica de manera inteligente los datos de lista, datos tabulares y botones de paginación sin tener que establecer reglas manualmente.

Herramienta	Descripción
Octoparse [14]	Es utilizada por programadores y analistas de datos. Tecnología de aprendizaje automático. Posee una extensión Cloud Extraction que permite raspado programado en tiempo real. Rotación de IP automática: cuando se configura una tarea de extracción para ejecutarse en la nube, las solicitudes se realizan en el sitio web de destino a través de varias IP, lo que minimiza las posibilidades de ser rastreado y bloqueado.

En la Tabla 2, se muestran los criterios de comparación entre las herramientas de extracción de datos, para ello se tiene en cuenta: Si son herramientas instalables o bien se accede sólo por web; Si la interfaz es amigable y posee varias funcionalidades desde la aplicación; Se identifican los formatos en los que se pueden exportar los datos; y si brinda documentación para manipular la herramienta.

Tabla 2. Comparativa de Herramientas de Raspado de Datos.

Herramienta	Instalable/ Web/Plugin	Interfaz amigable	Funciones	Monitoreo de Cambios	Formatos	Documen-tación	Gratis/ Pago
Import.io [8]	Sitio Web	SI	Muy completas	SI	XLSX; CSV y JSON	SI	Ambos
Web Scraper [9]	Plugin Chrome	SI	Básicas	NO	CSV; JSON	Poca	Gratis
Dexi.io [10]	Sitio Web	NO	Completas	SI	CSV; JSON	SI	Pago (tiene Trial)
ParseHub [11]	Instalable (MAC, Windows, Linux)	SI	Muy Completas	NO	JSON; EXCEL	SI	Ambos
Outwit Hub [12]	Instalable (MAC, Windows, Linux)	SI	Básicas	SI	EXCEL; JSON; XML; CSV; SQLLite; TXT; HTML; SQL	SI	Gratis
Scrapestorm [13]	Instalable (MAC, Windows, Linux)	MEDIA	Muy Completas	SI	EXCEL; CSV; TXT; HTML; MySQL; MongoDB, SQL Server, PostgreSQL y WordPress	SI	Ambos
Octoparse [14]	Instalable Windows Se requiere Framework (.NET3.5 SP1)	SI	Muy Completas	SI	EXCEL; JSON; CSV; HTML; DB	SI	Ambos

2.1 Consideraciones ante la elección de una técnica de Web Scraping

Ante la elección de una mejor técnica de Web Scraping, es necesario tener presente los siguientes puntos:

a) ¿Sólo es importante extraer información de una página web?
b) Los datos extraídos: ¿serán utilizados directamente en otra página web, o bien serán analizados en una PC?
c) ¿Es importante hacer una sincronización de actualización con el sitio web que se está realizando la técnica de extracción de datos?

Si se responde a cada una de estas preguntas se logrará entender la opción que más conviene para llevar a cabo para una minería de datos. Cada una de estas preguntas, permite identificar que se necesita para realizar un raspado de datos.

Si se necesita recuperar datos una única vez y luego procesar dicha información, entonces, es posible utilizar las herramientas de los navegadores, ya que permiten obtener el resultado en un formato CSV y XML, lo cual es totalmente manejable.

Si lo que se pretende es utilizar la extracción de datos como herramienta para tener actualizada la información y como siguiente paso, utilizar ésta en una página web, entonces se puede pensar en dos opciones: a) Utilizar un servicio web como ser: import.io [8], Dexi.io [10], ParseHub [11] o bien otra opción que permita el raspado constante de la información en un sitio web. Cada uno de ellos requiere conocimiento de lenguajes de programación, ya que, en este caso, no se utilizan las herramientas de los navegadores (plugin). b) Otra opción es el scraping web a medida, por medio de programación. Es decir, se realiza el programa de software de raspado de página web a medida, con el fin de extraer sólo los datos que sean necesarios, luego es fundamental realizar la integración y esta información pueda ser actualizada en un determinado período. Tanto para este caso, como para el anterior, también es necesario tener conocimientos de lenguaje de programación y de integración de API. Cabe destacar que en este trabajo se realizó un análisis de las herramientas completas, no así de las opciones de extracción de datos que brinden servicio de API únicamente.

3 Caso de Estudio

A continuación, se propone un caso de estudio en el que se desea extraer datos del sitio web del Estado Nacional Argentino (Ministerio de Modernización de Argentina) [15]. Los datos a extraer son puntualmente los 14 trámites y servicios que pueden realizar los ciudadanos argentinos desde el sitio web, catalogados por categorías junto con su descripción y link. Se extraerán los datos mediante la técnica de raspado de datos. Para esto se aplicarán las distintas herramientas presentadas en la sección anterior. En la Figura 2 se muestran los 14 trámites y servicios a extraer con las herramientas de raspado de datos.

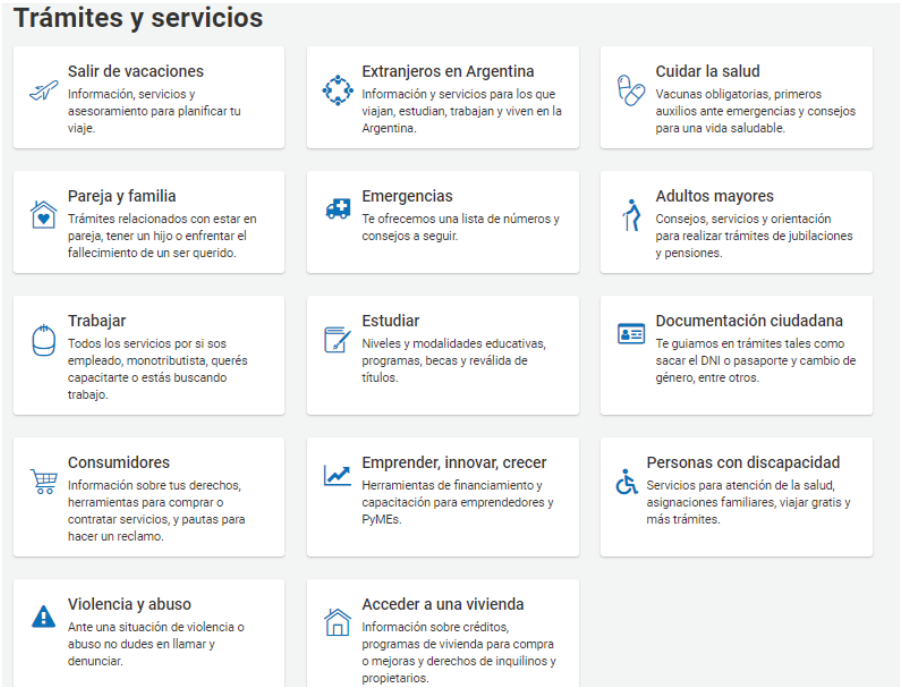

Fig. 2. Trámites y servicios del sitio web del Estado Nacional Argentino (Ministerio de Modernización).

4 Resultados de la comparativa

En base a los resultados arrojados, se muestra a modo de resumen, el análisis de las herramientas de raspado de datos teniendo en cuenta los siguientes criterios en la Tabla 3: a) ¿Logra extraer los 14 trámites y servicios disponibles?; b) Tiempo de extracción tomando como “Rápido” a un rango de 1 segundo a 5 segundos; “Medio” en un rango de 6 segundo a 10 segundos; y “Alto” en un rango de más de 10 segundos; c) Nivel de trabajo de depuración para obtener los datos limpios extraídos de la herramienta, tomando como “Alto”; “Medio” y “Bajo”; d) Dificultad a nivel de extracción.

Para comparar los resultados de la extracción de datos se ha exportado todo a CVS que es el formato común de todas las herramientas comparadas. Algunas herramientas permitían exportar a JSON (tal como se indicó en la tabla 2) pero para comparar lo obtenido fue necesario exportar a CVS que todas lo tienen habilitado.

Tabla 3. Comparativa de Resultados de las Herramientas de Raspado de Datos.

Herramienta	Efectividad de Extracción	Tiempo [Rápido/ Medio/ Lento]	Separación de los datos	Depurar resultado [Alto/Medio/ Bajo]	Dificultad de Extracción [Alta/Media/ Baja]
Import.io [8]	14 registros	MEDIO	","	MEDIO: Se muestra el texto plano exportado correctamente. Se exporta una columna adicional en la que repite el URL. Se exportan 2 columnas extras.	BAJA
Web Scraper [9]	43 registros	RÁPIDO	","	ALTO: Se extrae más datos de los solicitados, como ser números de IDs e información redundante.	MEDIA
Dexi.io [10]	14 registros	MEDIO	","	MEDIO: Se muestran caracteres especiales en lugar de acentos. Se exporta una columna adicional de control de errores. No se extrajeron links.	MEDIA
ParseHub [11]	14 registros	RÁPIDO	","	MEDIO: Se muestran caracteres especiales en lugar de acentos. Se exporta una columna adicional que repite el URL. Tiene problemas si los datos originales tienen coma ya que lo toma como separador de campo.	BAJA
Outwit Hub [12]	16 registros	RÁPIDO	"," y ";"	MEDIO: Tiene dos separadores "," (coma) y ";" (punto y coma); pero se extrae información de más, provenientes de la sección anterior.	BAJA
Scrapestorm [13]	17 registros	MEDIO	","	ALTO: Se muestran caracteres especiales en lugar de acentos. Se exportan registros de más que surgen del raspado.	BAJA
Octoparse [14]	1 registro (con 15 tramites, uno duplicado)	MEDIO	","	ALTO: Todos los datos son exportados juntos en un mismo registro. No se comprende bien la división si los datos extraídos tienen "," (coma). No se extrajeron links.	MEDIA

Para seleccionar una de estas herramientas, se descartan aquellas que no logran recuperar los 14 registros. Luego se analiza también los inconvenientes que se presentan al depurar el resultado obtenido, por ejemplo, aquellas herramientas que tienen problemas con los acentos ó las comas del texto original. El caso de estudio

considerado, tiene el nombre de un servicio con dos comas, y así también tiene acentos en nombres o descripciones; lo cual genera problemas en la exportación con algunas herramientas. Algunas herramientas traen registros vacíos, duplicados, etc. Lo que no permite una extracción directa de lo obtenido. Como puede observarse en la tabla 3, sólo 3 herramientas permitieron traer correctamente los 14 registros esperados. Por los resultados obtenidos se selecciona como herramienta para la extracción de datos a Import.io [8], ya que, si bien es una herramienta que requiere de un trabajo de limpieza de datos de un nivel medio, posee una buena identificación de los registros extraídos para su reutilización e identificación de datos abiertos. A comparación de las otras herramientas, no presenta caracteres extraños en los acentos, además, permite una extracción de datos en forma amigable y con múltiples funciones, a través de una licencia gratuita. Por otra parte, se destaca el tiempo de rapidez en la extracción. En la Figura 3 se muestran los datos raspados de la herramienta Import.io [8], identificando los 14 trámites y servicios del sitio web.

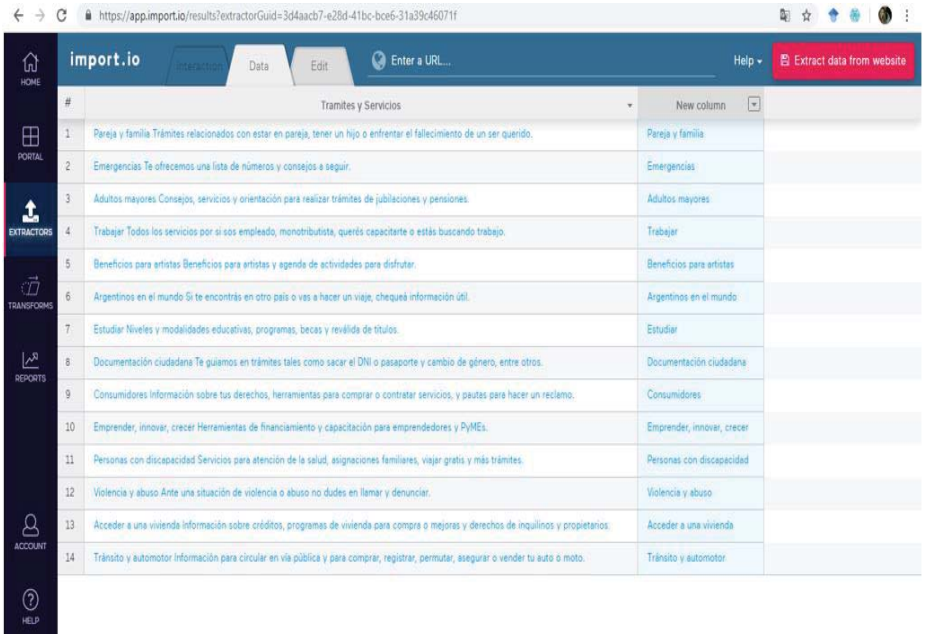

Fig. 3. Se muestran los trámites y servicios raspados del sitio web.

5 Conclusiones y Trabajos Futuros

Si bien existen diversas herramientas en el mercado no todas han dado por resultado una extracción correcta de los registros a recuperar en el caso de estudio seleccionado. Por lo que es importante analizar las herramientas existentes en cuanto al resultado generado, lo cual se realizó analizando los CSV exportados. Luego se contemplan otros parámetros como la dificultad y el tiempo de extracción, siendo el parámetro más decisivo la efectividad en el resultado obtenido. Como se trató en este

trabajo, el scraping es una técnica que se utiliza para extraer datos de cualquier sitio web, siempre y cuando se arme el mapa de los datos a extraer. Mediante la extracción de datos de los sitios web gubernamentales, se puede exportar a un formato más sencillo y así lograr analizarlos y cruzarlos con mayor facilidad, disponibilizándolos como datos abiertos. Para recopilar automáticamente y mostrar esta información, los scrapers, permiten la manipulación de los datos con el fin de obtener un almacenamiento de datos abiertos y que puedan ser compartidos para su posterior análisis estadístico. Estos datos compartidos podrían ofrecerse en un entorno donde a través de crowdsourcing los usuarios puedan escribir comentarios sobre estos trámites o servicios, realizar aportes en caso de detectar errores en la información ofrecida. La manipulación de estos datos extraídos es una línea de trabajo futuro.

Referencias

[1] Sinead Garvan, BBC, "Cambridge Analytica: cómo Netflix retrata el mayor escándalo de privacidad en las redes sociales en *Nada es privado*", Disponible en: https://www.bbc.com/mundo/noticias-49122905

[2] Costa, F., & Rodríguez, P. (2018). ALGORITMOS, BIG DATA Y AUTOMATIZACIÓN SOCIAL. AVATARES de la Comunicación y la Cultura, (15).

[3] Estrada, J. C. H., Silva, I. A. M., & Páez, J. O. B. (2018). Big Data: Ventajas y desventajas-aplicaciones y tecnologías para implementar el servicio. COMITÉ CIENTÍFICO CICOM 2018, 44.

[4] Abella, A., Ortiz-de-Urbina-Criado, M., & De-Pablos-Heredero, C. (2018). Indicadores de calidad de datos abiertos: el caso del portal de datos abiertos de Barcelona. El profesional de la información (EPI), 27(2), 375-382.

[5] Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. PloS one, 9(5), e93949.

[6] Andrés, O. R., Pulido, J. R. G., Guillermo, A., & Morales, J. R. H (2019). Recuperación de metadatos e indicadores de impacto para publicaciones científicas mediante servicios de Google académico.

[7] Google Trends, "Descubre qué está buscando el mundo", Disponible en: https://trends.google.com/trends/?geo=US

[8] Import.io, "Proporcionar los datos web que informan las decisiones comerciales cotidianas", Disponible en: https://www.import.io

[9] Web Scraper, "More than 250,000 users proud of using our solutions", Disponible en: https://webscraper.io/

[10] Dexi.io, "Web Automation Software [Scraping ETL API AI]", Disponible en: https://dexi.io/

[11] ParseHub, "A web scraping tool that is easy to use", Disponible en: https://www.parsehub.com/

[12] Outwit Hub, "OutWit Hub explores the depths of the Web for you, automatically collecting and organizing data and media from online sources", Disponible en: https://www.outwit.com/products/hub/

[13] Scrapestorm, "AI-Powered Visual Web Scraping Tool", Disponible en: www.scrapestorm.com

[14] Octoparse, "Extraiga fácilmente cualquier dato web", Disponible en: https://www.octoparse.com/product

[15] Estado Nacional Argentino (Ministerio de Modernización de Argentina), "Argentina.gob.ar", Disponible en: https://www.argentina.gob.ar/