Web application in Shiny for the extraction of data from profiles in Google Scholar

Danny Murillo, Mastría[1], Dalys Saavedra, Maestría[1] , Robinson Zapata, Maestría[1]
[1]Universidad Tecnológica de Panamá, Panamá, danny.murillo@utp.ac.pa, dalys.saavedra@utp.ac.pa
[2]Secretaría Nacional de Ciencia y Tecnología e Innovación, Panamá, rzapata@senacyt.gob.pa

***Abstract* – The need to measure the contribution of researchers through academic profiles is of great importance, which is why in 2018 we created an algorithm in R language to dynamically extract data from individual and institutional public profiles in Google Scholar Citations. Although the algorithm has been of great use in the automatic extraction of data, allowing statistical reports and analyzes to be carried out with this data, it is only possible to use it if the user knows the R language, due to the multiple functions that the R language has integrated. algorithm.**

In this work we show the creation of a web application integrating the algorithm to extract data from Google Scholar Citations but improving the ease of use of these scripts using the R Shiny package, which integrates web components from Rstudio but maintaining the programming characteristics of the language. .

Shiny converts scripts into interactive web applications, without any knowledge of HTML, CSS or Javascript, making it easy for users to use, manipulate, view, and allow for future updates to improve functionality.

The results of the tests and tasks carried out in this work show that the use of the web application in Shiny, the extraction algorithm could be integrated without difficulty, improving the extraction time in seconds and minutes, because the user does not interact with it. R code but with the Web interface allowing users new to R who are dedicated to the analysis of Google Scholar data to use it.

Keywords— Google Scholar Citations, shiny, web application, web scraping, bibliometric indicators.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2022.1.1.235
ISBN: 978-628-95207-0-5 **ISSN:** 2414-6390

Aplicación web en Shiny para la extracción de datos de perfiles en Google Scholar

Danny Murillo, Mastría[1], Dalys Saavedra, Maestría[1] , Robinson Zapata, Maestría[1]
[1]Universidad Tecnológica de Panamá, Panamá, danny.murillo@utp.ac.pa, dalys.saavedra@utp.ac.pa
[2]Secretaría Nacional de Ciencia y Tecnología e Innovación, Panamá, rzapata@senacyt.gob.pa

Resumen – La necesidad de medir el aporte de investigadores a través de los perfiles académicos es de mucha importancia, es por ello que en el año 2018 creamos un algoritmo en lenguaje R para extraer dinámicamente datos de perfiles públicos individuales e institucionales en Google Scholar Citations. Aunque el algoritmo ha sido de gran utilizad en la extracción automática de datos, permitiendo realizar con estos datos, reportes y análisis estadísticos, solo es posible hacer uso de él si el usuario conoce el lenguaje R, debido a las múltiples funciones que tiene integrado el algoritmo.

En este trabajo mostramos la creación de una aplicación web integrando el algoritmo para extraer datos de Google Scholar Citations pero mejorando la facilidad de uso de estos scripts utilizando el paquete R Shiny, el cual integra componentes web desde Rstudio pero manteniendo las características de programación del lenguaje.

Shiny convierte scripts en aplicaciones web interactivas, sin ningún conocimiento de HTML CSS o Javascript, lo que facilita su uso, manipulación, visualización por parte de los usuarios y permite futuras actualizaciones para mejorar su funcionalidad.

Los resultados de las pruebas y tareas realizadas en este trabajo muestran que el uso de la aplicación web en Shiny, el algoritmo de extracción se pudo integrar sin dificultad mejorando el tiempo de extracción en segundos y minutos, debido a que el usuario no interactúa con el código R sino con la interfaz Web permitiendo que usuarios noveles en R que se dedican al análisis de datos de Google Scholar puedan utilizarlo.

Palabras claves— Google Scholar Citations, shiny, aplicación web, web scraping, indicadores bibliométricos.

I. Introducción

Desde el siglo pasado a través de diversas investigaciones, Internet quedó firmemente establecido como un recurso accesible para las personas que ofrece una amplia variedad de beneficios potenciales, entre ellos, el acceso a todo tipo de información, donde, tanto los gobiernos, negocios y entidades educativas disponen de una plataforma para obtener información o brindar algún servicio [1]. Aunque el acceso y la sobreabundancia de información en la red es uno de los principales componentes de su éxito, el tratamiento de estos datos para generar conocimiento exige una enorme cantidad de tiempo y energía a fin de cribar la calidad de los datos sumergidos en tan enorme repositorio de documentos [2], [3]. Los datos organizados, estructurados y visibles en páginas web, no siempre permiten su extracción debido al formato inadecuado en que se presentan o porque existe limitaciones de descarga, esto evita su uso y análisis posterior de esta información.

En el ámbito académico y de investigación, una de las plataformas más utilizadas para búsqueda de información es Google Académico o Google Scholar (GS). Según datos del portal Social Media en Investigación, en el 2015, el 75% de los investigadores iniciaban sus primeras búsquedas de información en esta plataforma [4], [5] ya que integra documentos de diversas fuentes en la web. Para enero del 2020 según el portal de estadísticas web Alexa.com, el portal scholar.google.com (GSc), era uno de los sitios más visitados por día. Este sitio era consultado casi por segundo y tenían 12 minutos más por día en consulta, que el total de minutos acumulados de otras plataformas de contenidos académicos como Microsoft Academic, Researchgate y Wikipedia.

GS fue lanzado en noviembre de 2004 como buscador especializado de contenidos académicos y científicos, desde el 2012 ofrece a los académicos la posibilidad de crear su perfil de investigadores utilizando Google Scholar Citations (GSC) [4]. El GSC permite listar sus publicaciones e integrarlas de forma manual o automática de trabajos que puede provenir de diversas fuentes en la web. Esta plataforma tienen el objetivo de visibilizar la producción científica con acceso a indicadores de forma gratuita, esto lo convierte en un competidor de otras plataformas de pago, como Web of Science (WoS) y Scopus [6]. El perfil en GSC genera indicadores Bibliométricos tales como, el hindex [7] citas, número de publicaciones, siendo actualmente estos indicadores objeto de estudio por investigadores, buscando esclarecer su potencialidad, confiabilidad e importancia, como instrumento de recuperación de información científicas y medición [8].

Los datos generados en GSC, son de mucha importancia para la comunidad científica, sin embargo, extraer esta información sobre todo los perfiles de una institución para su análisis, no resulta fácil. Una de las formas que utilizan los usuarios para acceder a los datos de las plataformas web es utilizando API´s. Las API o Interfaz de Programación de Aplicaciones, en español, es un conjunto de subrutinas, funciones y procedimientos generadas por el proveedor, que ofrecen la posibilidad de extraer los datos de forma directa y gratuita. Actualmente GS no ha desarrollado ninguna interfaz de acceso a los datos, sin embargo si existe una API desarrollada por la empresa SerApi llamada Google Search API que permite extraer los datos de búsqueda de GS según palabras, fecha o fuente de datos[9] , pero no extrae datos del GSC, añadiendo que el API fue desarrollada por una empresa

externa a GS, su uso tiene un costo mensual según el tipo de usuario y número de consultas.

Otras de la opción encontrada para extraer datos de GSC, fue utilizando el software gratuito, Publish or Perish que permite extraer indicadores bibliométricos por perfil y exportarlos [10]. Aunque esta herramienta es muy robusta, no es posible extraer el listado de perfiles en GSC de una institución, por lo que la tarea de extraer varios perfiles a la vez se debe realizar de manera semiautomática, ya que se deben conocer los perfiles de cada institución y extraer sus indicadores de manera individual y después unirlos en una sola tabla.

Para ayudar con este problema, en el año 2018, un grupo de investigadores de la Universidad Tecnológica de Panamá, creó un algoritmos utilizando el lenguaje R para extraer datos de GS utilizando la técnica de Web Scraping [11], [12]. El algoritmo que integra varias funciones, utiliza está técnica no estructurada de minería de datos para escanear el código HTML de una página web [13] y de manera dinámica extrae los datos que se muestran estructurados o no y los transforma en formato de tabla para su depuración, visualización, acceso y posterior exportación.

Utilizando este algoritmo se han realizado 8 trabajos académicos y publicaciones, donde se han extraído más de 40000 perfiles de diferentes universidades probando su eficiencia en la extracción, entre ellos: Estudio de indicadores científicos de perfiles en Google Académico de universidades en Centroamérica y el Caribe, Perfiles de Investigadores de Panamá con perfil público en Google Scholar (Julio 2019), Analysis of Bibliometric Indicators of Panama Magazines to Know Their Scope and Projection of Citations and Impact in the National Context, existe un inconveniente para quienes utilicen el algoritmo, es necesario tener conocimiento del lenguaje R y Rstudio, por lo que limita su uso y aprovechamiento de esta herramienta para algunos usuarios noveles de R.

En este artículo mostramos los resultados de crear una aplicación web integrando el algoritmo para extraer los datos de Google Scholar Citatiosn pero mejorando la facilidad de uso de estos scripts a través de la creación de una interface web que no requiere que el usuario tenga conocimiento del lenguaje R. Para diseñar esta aplicación nos enfocamos en el uso del paquete Shiny de R [14], el cual integra varios componentes basado en un marco de ejecución vía web. Shiny integra las características de programación del lenguaje R y el IDE Rstudio [15], lo que hace que sea más fácil de convertir scripts creados en R en una aplicación web interactiva, fácil de usar, manipular, visualizar y actualizar de forma automática para mejorar funcionalidades posteriores.

El objetivo del trabajo fue desarrollar una aplicación web para extraer los perfiles individuales e institucionales de GSC que facilitara al usuario su uso a través de una interface sencilla, pero, manteniendo la rapidez de extracción del algoritmo utilizando Rstudio. La importancia de su creación es que pueda ser utilizada por los investigadores, instituciones académicas, como de investigación para extraer datos de indicadores bibliometricos proveniente de esta plataforma.

A. Lenguaje R

R es un entorno y lenguaje de programación diseñado por Ross Ihaka y Robert Gentleman, especialmente desarrollado para el análisis de datos, cálculos estadísticos y representaciones gráficas. La sintaxis de este lenguaje de programación es parecida a otros lenguajes como C, C++ y Phyton por lo que facilita su aprendizaje a quienes conozcan estos lenguajes. R es un lenguaje de script, por lo que no es necesario compilar su código para ejecutar y mostrar los resultados de las líneas programadas. Es un software libre y constantemente se crean paquetes por diferentes autores que se pueden integrar a R. Está habilitado para múltiples plataformas como Windows, Mac y Linux[16].

B. Rstudio

Este software es una IDE de programación que contiene la consola del RCommander, componente donde se programa las líneas de código del lenguaje. Rstudio permiten contar con una interacción más fluida ya que su apariencia es un entorno amigable de ventanas y paneles que tiene como principales ventajas, el orden de sus funciones y la visualización de los procesos que son llevados a cabo con R de manera simultánea. Su interface facilita la visualización del código, listado de variables y funciones utilizadas, ejecución del código script, archivos utilizados por el usuario y la ventana de gráficos desde una sola interface [17].

C. Web Scraping

Es una técnica de la minería de datos que consiste en la extracción del código HTML de una o varias páginas web de forma automatizada. El objetivo de extraer el código es leer su estructura a través de nodos o selectores HTML para separar el código fuente de los datos y así guardar los contenidos de forma estructurada con datos que resulten con información relevante para su análisis posterior. Este proceso también llamado Web crawler es el mismo proceso que hace Google para integrar las páginas web a su buscador [13].

D. Algoritmo en R para scraping en GS

El algoritmo en R para extraer datos de Google Scholar fue creado en el año 2018 por investigadores de la Universidad Tecnológica de Panamá, con el fin de facilitar la extracción de datos de perfiles en Google Scholar y exportar datos en formato CSV, fue actualizado en el año 2019 añadiendo funciones para extraer el gráfico de citas del perfil y las palabras claves de los perfiles. El algoritmo utiliza la técnica de Web Scraping y contiene 6 funciones relacionadas con carga de librerías a utilizar, extraer perfil individual, extraer perfiles institucionales, extraer publicaciones de perfil institucional, grafico de citas por año y ordenar perfiles por citas y documentos, estas deben ejecutarse en el entorno de Rstudio, por lo que se requiere de conocimiento básico del lenguaje R.

El proceso de extracción manual de datos de 55 perfiles y 1400 publicaciones podría durar entre 2 a 4 horas, añadiendo el tiempo de error humano al copiar y pegar, al utilizar el

algoritmo el tiempo promedio de extracción es de 3 a 5 minutos y generando los datos en formato de tabla [18].

E. Shiny

Es un paquete de R que proporciona un entorno web fácil y potente para implementar aplicaciones web usando R. Shiny convierte la programación y análisis de datos, en aplicaciones webs interactivas, sin ningún tipo de conocimiento HTML CSS o Javascript [19]. Su estructura se crea utilizando dos archivos en R que se comunican entre sí: un archivo de interfaz de usuario (UI), que controla diseño y apariencia, está es la interface que se mostrará en la web; y un script de servidor (Server) que incorpora instrucciones para entrada de usuario, funciones que permiten el procesamiento de datos y salidas de tabla de datos en diferentes formatos o exportar imágenes de gráficos generados. Estos scripts pueden estar en un solo archivo llamado app.R, utilizado para lanzar la aplicación a un entorno web [20].

D. Launcher de Web Shiny App

El Launcher o lanzamiento, es el proceso de habilitar una aplicación desarrollada en Shiny en formato de página web. Esta interface contiene elementos visuales amigables que ocultan el código R al usuario. Para lanzar una aplicación Shiny en R, es necesario utilizar el comando runApp(), en el archivo app.R. Con este único código la aplicación se muestra en una nueva ventana en Rstudio con formato web o en una ventana del navegador web predeterminado, eliminando la complejidad de ejecutar los scripts en modo de comandos. Esta interface permite interactuar con otras librerías o funciones creadas en el script del Server, el cual contiene el código que interactúa con las decisiones del usuario en la interface web [21]. La aplicación web puede ser integrada al servidor de Shiny para no tener dependencia de R o Rstudio y que pueda ser usada por varios usuarios de forma simultánea, pero se debe hacer pruebas de compatibilidad.

E. Indicadores Bibliométricos en GSC

Los indicadores bibliométricos permiten cuantificar el comportamiento de la producción bibliográfica y la comunicación científica, esto es posible gracias a la Bibliometría. Existen diversos criterios de clasificación que convergen en dos agrupaciones esenciales: una división en indicadores de productividad, visibilidad o impacto y colaboración, y otra en unidimensionales [22]. En el caso de los indicadores de productividad se menciona el número de documentos generado por investigador, en el indicador de visibilidad e impacto se contempla el número de citaciones como el Hindex[23].

GSC genera 3 indicadores a través de los perfiles de investigadores para medir la producción de los autores e instituciones, ya que esta plataforma integra documentos de portales de revistas de acceso abierto y repositorios que no son tomados en cuenta por Web of Science y Scopus. Para medir el impacto de las revistas se utiliza el Google Scholar métrics y el indicador hindex[24].

II. METODOLOGÍA

A. Recursos

Se utilizó como herramientas principales el lenguaje R y diferentes paquetes en Rstudio:

- R version 3.6.2.
- Rstudio para Windows versión 1.2.5.
- Paquete Shiny 1.4.0
- Paquetes en R: Shinyjs, Shinythemes, Stringr, Rvest, xml2, ggplot2, DT, Tidyverse.
- Computador con Windows 7 de 64 Bits, Dual Core de 2.2 GHz, y Memoria RAM de 3 GB.
- La velocidad de Internet en periodo de pruebas fue de 26.96 Mb de descarga y 14.56 de carga.

B. Pruebas realizadas para comparación de métodos

Para las comparaciones del uso y funcionamiento de la extracción de datos utilizando los métodos, algoritmo en R.

1. Prueba de extracción de publicaciones de perfil individual: se utilizó la URL del perfil de un investigador en Google Scholar, extraer los datos de las publicaciones con los dos métodos, guardar los datos en formato CSV, generar y guardar gráfico de citas por año. Se contabilizó el tiempo en segundos con la función proc.time() de R.

2. Prueba de extracción de perfil / publicaciones institucionales: se utilizó la URL del perfil institucional en GS ya sea a través de la búsqueda por el nombre o el dominio URL de su sitio web institucional. Se extrajeron los datos de los perfiles y todas las publicaciones utilizando las dos herramientas, se guardaron los datos en formato CSV. Se contabilizó el tiempo en minutos con la función proc.time(). Esta prueba como la anterior. fue realizadas por personas con conocimiento tanto del lenguaje R como de Rstudio.

3. Prueba a usuarios de extracción de perfil individual / institucional: se seleccionaron 10 usuarios para utilizar ambos métodos para extraer los datos de los perfiles y publicaciones, tanto del perfil individual como institucional realizando 5 pruebas, extraer publicaciones de perfil individual (Extraer Pub. Ind), Extraer perfiles de perfil institucional (Extraer Perfil Inst.), Extraer publicaciones de perfil individual (Extraer Pub. Inst.), guardar datos, generar y guardar gráficos de citaciones por año.

C. Población seleccionada en las pruebas

1. Prueba de extracción de publicaciones de perfil individual, se seleccionaron 10 perfiles públicos en GS de investigadores con diferentes número de citaciones de 10 universidades diferentes (Martinha Pereira del Instituto Politécnico do Cavado e do Ave, Paul Anderson del George Fox University, Victoria Castro Rojas de la Universidad Alberto Hurtado, Raul Ramos-Pollan de la Universidad de Antioquia, Suzana Maria Ratusznei del Instituto Mauá de Tecnología, José Otero Parra de la Univ. Europea Miguel de Cervantes, Raul Becchio de la Universidad Nacional de Salta, Consuelo Lobato-Calleros de la Universidad Autónoma Chapingo, Daniel Alejandro Barrio de la Universidad

Nacional de Río Negro y Reinhardt Pinzón Adames de la Universidad Tecnológica de Panamá), donde se extrajeron todas las publicaciones por perfil, guardando el ID en GS. Las universidades se seleccionaron de forma aleatoria del Ranking Webometrics (RW) de enero 2020 [25].

2. Prueba de extracción de publicaciones de perfil institucional, se seleccionaron 5 perfiles públicos en GS de universidades de 5 países con diferentes números de citaciones globales según el RW, World Maritime University, Suecia, Universidad del Atlántico, Colombia, Universidad Tecnológica de Panamá, Panamá, Instituto Mauá de Tecnología, Brazil y George Fox University, Estados Unidos.

3. Prueba a usuarios de extracción de perfil individual / institucional, se seleccionaron 10 usuarios sin ningún conocimiento del lenguaje R, pero sí de uso de herramientas informáticas para ejecutar 6 tareas relacionadas con la extracción, filtrado, visualización, y guardado de los datos y gráfico. Se les explicó el funcionamiento básico tanto del algoritmo en Rstudio como el funcionamiento de la aplicación web en Shiny. No se contabilizó el tiempo sino las fallas de los usuarios por cada tarea.

D. Desarrollo de la aplicación en Shiny e integración de algoritmos

Para el desarrollo de la aplicación se utilizó la estructura de Shiny que utiliza dos archivos de scripts, uno para mostrar la interface web y el otro para ejecutar las funciones del lado del servidor. En la figura 1, se muestra el esquema de la estructura de la aplicación Shiny desde una interfaz web.

El script UI en el archivo *APP*, contiene el código en Shiny para mostrar los componentes de una interface web que solicita un solo dato al usuario, la URL de GSC del perfil individual o institucional. Una vez se introduce la URL se selecciona el tipo de información que desea extraer y se selecciona el botón SCRAPER para iniciar se inicia la extracción de los datos. El formato de URL individual es la que genera Google Scholar (https://scholar.google.com/citations?hl=es&user=QObKt9IAAAAJ) contiene un código de 12 dígitos (QObKt9IAAAAJ) como identificador único, esta forma de codificación no se muestra en las URL del perfil institucional, lo que permite validar que tipo de URL introduce el usuario antes de hacer el proceso de Extracción. El algoritmo transforma los datos extraídos en formato de tabla y según la opción de extracción seleccionada, se visualizará, el grafico de citaciones por año del perfil individual, la tabla de resultados y botones de descargar datos e imagen del gráfico. Todas las opciones permiten descargar los resultados en formato CSV.

El algoritmo desarrollado en R para extracción de datos fue estructurado en diversas funciones para que pudiera ser ejecutado por cada uno de los diferentes eventos al presionar los botones de las interfaces. Estas funciones fueron integradas en el archivo *APP* en la sección, script *SERVER*.

Fig. 1. Estructura del funcionamiento de aplicación en Shiny para extraer datos de GS

En la figura 2, se muestra un esquema de los componentes de Shiny integrados en el script UI, los componentes del script SERVER que corresponde a todas las funciones generadas del algoritmo en R y los datos de entrada. Solo hay dos elementos

externos, la URL que introduce el usuario, la cual es validada para la aplicación, siendo este dato utilizado en el proceso de web Scraping, donde, dependiendo de la opción de extracción se ejecuta una de las 3 funciones. Los datos extraídos se almacenan en varios DataFrames que se muestran en formato de tabla al usuario, si el usuario lo decide descarga las tablas y los gráficos. Las 5 funciones son almacenadas en un archivo diferente a *APP* para mantener el código de scraping separado, siendo las funciones creadas para extraer los datos: PubGs_library(), carga las funciones que utiliza el algoritmo como Shiny, PubGS_perfil(), extrae publicaciones por perfil, PubGS_research(), extrae listado de Investigadores, PubGS_publications(), extrae publicaciones del listado de perfiles, graficoIndividual(), muestra el gráfico de citas por año del perfil y genera una imagen de descarga, la función PubGs_afiliacion(), extrae datos afiliación del perfil individual (nombre, universidad). Una vez se muestren los resultados en la interfaz, el usuario puede generar otra búsqueda, sin embargo sino se descargaron los datos, hay que volver a hacer la extracción , ya que los datos no se guardan en ninguna base de datos, solo en memoria.

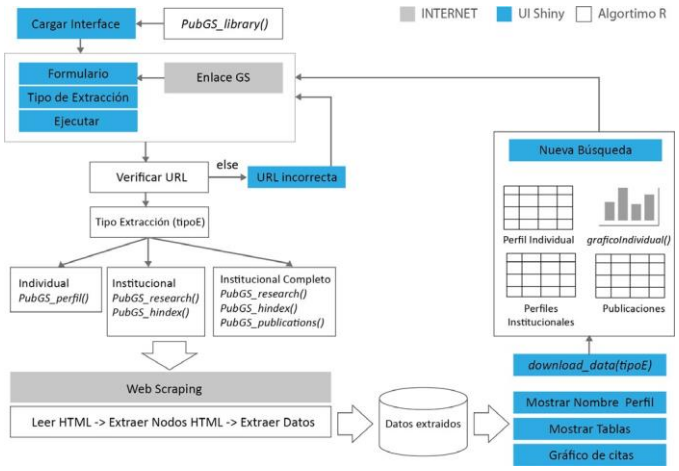


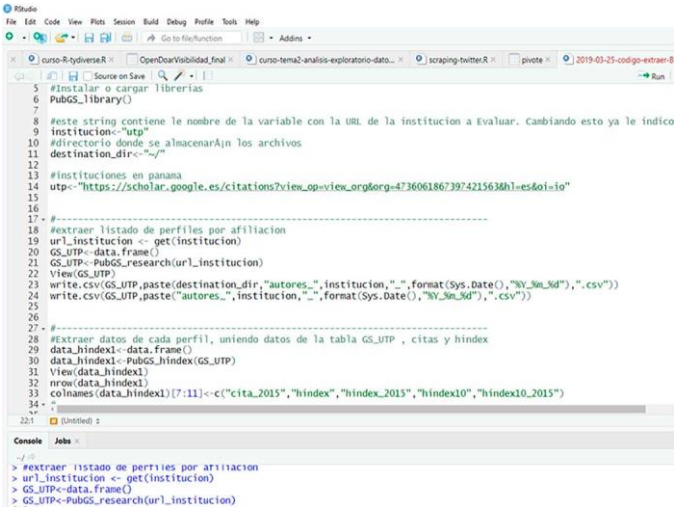
Fig. 2. Estructura de los componentes de la aplicación en Shiny y las funciones que ejecuta el algoritmo

III. Resultados

Los resultados presentados en esta sección muestran la comparación al extraer datos de GSC utilizando dos métodos, el algoritmo en Rstudio a través de código en consola y la interface Web en Shiny para extraer datos de GSC. En esta comparación se muestran indicadores del tiempo de extracción que le tomó a cada usuario en utilizar ambas opciones, imágenes de la interface y el uso por parte de los usuarios favorece a la interface Web en Shiny.

1. Métodos de extracción utilizando la consola en R

El principal inconveniente al extraer los datos utilizando el algoritmo en Rstudio es que el proceso de ejecución resulta complejo por su interfaz poco amigable. En la figura 3, se muestra los códigos del algoritmo en R y la consola del proceso y resultados que realiza el script, este código contiene más de 1000 líneas de código. Al terminar de utilizar la función para la extracción del listado de perfiles en GSC se debe utilizar los datos extraídos y ejecutar otras funciones para extraer las publicaciones del perfil, luego visibilizar y guardar los datos utilizando comando en R, lo que puede complicar el uso de este script para un usuario que no conoce el lenguaje.



```
#Instalar o cargar librerias
PubGS_library()

#este string contiene le nombre de la variable con la URL de la institucion a Evaluar. Cambiando esto ya le indico
institucion<-"utp"
#directorio donde se almacenaran los archivos
destination_dir<-"~/"

#instituciones en panama
utp<-"https://scholar.google.es/citations?view_op=view_org&org=4736061867397421563&hl=es&oi=io"

#-------------------------------------------------------------------------
#extraer listado de perfiles por afiliacion
url_institucion <- get(institucion)
GS_UTP<-data.frame()
GS_UTP<-PubGS_research(url_institucion)
View(GS_UTP)
write.csv(GS_UTP,paste(destination_dir,"autores_",institucion,"_",format(Sys.Date(),"%Y_%m_%d"),".csv"))
write.csv(GS_UTP,paste("autores_",institucion,"_",format(Sys.Date(),"%Y_%m_%d"),".csv"))


#-------------------------------------------------------------------------
#Extraer datos de cada perfil, uniendo datos de la tabla GS_UTP , citas y hindex
data_hindex1<-data.frame()
data_hindex1<-PubGS_hindex(GS_UTP)
View(data_hindex1)
nrow(data_hindex1)
colnames(data_hindex1)[7:11]<-c("cita_2015","hindex","hindex_2015","hindex10","hindex10_2015")
```

```
> #extraer listado de perfiles por afiliacion
> url_institucion <- get(institucion)
> GS_UTP<-data.frame()
> GS_UTP<-PubGS_research(url_institucion)
```

Fig. 3. Interfaces en Rstudio con el código del algoritmo de extracción

2. Métodos de extracción utilizando web Shiny

Para desarrollar la aplicación en Web Shiny, fue necesario programar las interfaces utilizando el paquete R Shiny y cargar las funciones del algoritmo creado en el lenguaje R. En esta aplicación, no es necesario que el usuario conozca cómo funciona el código para utilizar la aplicación web, solo cual es el archivo app.R que debe ejecutar. Lo que verá el usuario al lanzar la aplicación es la interfaz Web Shiny como aparece en la figura 4. En la interfaz, solo se muestra un único campo que es la URL del perfil individual en GSC o institucional en GS, el cual debe introducir el usuario. En función de la URL que se introduzca el programa selecciona el tipo de extracción.

.

Fig. 4. Interfaces en Web Shiny para extraer datos en GS

Al seleccionar el botón SCRAPER de la interfaz web, la aplicación ejecuta la función para leer el código HTML de la página web en GS mostrada en la figura 5a, el algoritmo ejecuta el proceso de scraping, extrae los datos de cada perfil e indicadores y los estructura en formato de tabla como se muestra en la figura 5b. La imagen representa el listado de perfiles en Google Scholar y los datos de los perfiles de una institución.

	Nombre	citas	word_key	cita_2016	hindex
3	Reinhardt Pinzon Adames	1606	Climate Change, Downscaling climate projections, C...	640	11
4	Elida de Obaldia	1213	Material Science, Renewable Energy, Diamond thin f...	267	12
5	Graciela Cecilia Sánchez Hidalgo	1101	Renewable energy, Climate Change, Sustainability	253	12
6	Gilberto Axel Chang, PhD	753	Structural Engineering, Seismic Engineering	356	6
7	Lilia Muñoz	722	Sistemas de Información, Aplicaciones de la Informá...	381	13
8	Oscar M. Ramirez	721	Structural and Earthquake Engineering	280	6
9	Lisseth Sandoval	689	Bioingeniería y Biotecnología	231	9
10	Gabriel Ortega	677	Tecnología Mecanica	577	17
11	Daniel Alejandro Saavedra Gallardo	676	Ingeniería, Industria, Productividad, Control de Calid...	179	12
12	Héctor Montes Franceschi	658	Robotics, Control, Automation, ICT	312	14
13	Rodney Delgado-Serrano	487	Astrophysics, Astronomy, Physics, Mathematics, Inf...	130	5
14	Vladimir Villarreal	486	Ambient Intelligence, Ubiquitous Computing, Ambie...	200	12
15	Danilo Cáceres Hernández	478	Autonomous Navigation, Real Time Application	312	13

(b)

Fig. 5. (a) Listado de perfiles en GS de una universidad, (b) Tabla de datos en después de la extracción utilizando SHINY en R.

Al realizar el proceso de scraping con la aplicación en esta pantalla se habilitan dos botones, uno de color naranja para guardar las publicaciones en CSV y el otro de color azul para guardar el grafico de citas por año, los cuales se muestran en la figura 6. El gráfico mostrado también puede ser descargado con el botón derecho del mouse. Adicional, se muestran los resultados a través de una tabla dinámica que permite filtrar las filas según datos, como también buscar dentro de la tabla, esto hace más fácil la ejecución de las tareas para el usuario, filtro que no se puede hace en el algoritmo, ni tampoco en el perfil de GS

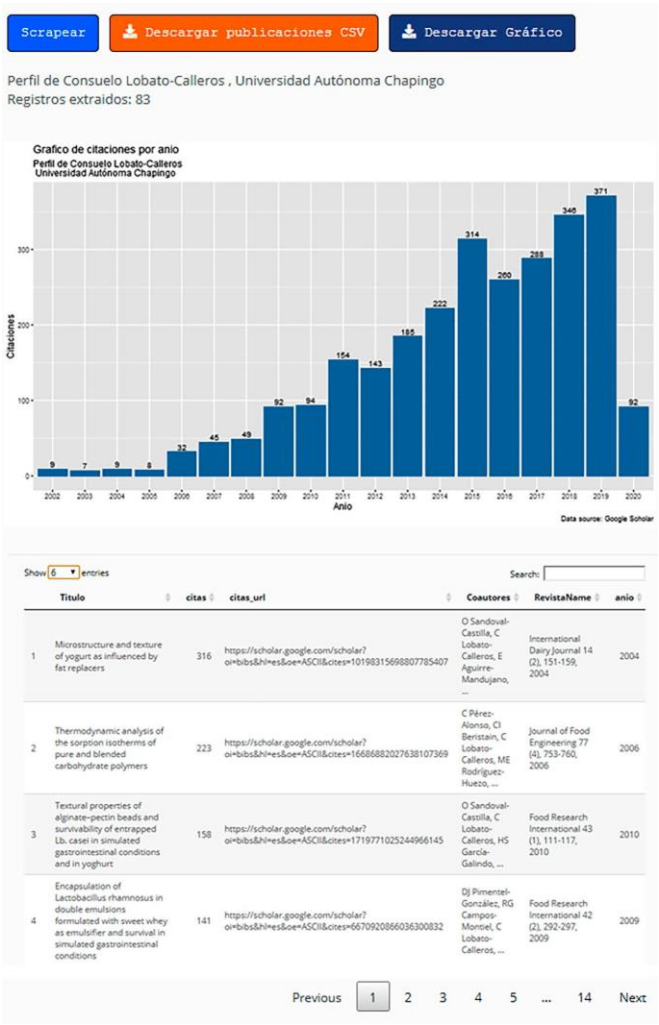


Fig. 6. Interface de los resultados al extraer perfil individual de GS en aplicación web Shiny

Para validar el funcionamiento de ambos métodos se realizó la **prueba de extracción de datos del perfil individual** en GS. La Tabla I, muestra los resultados con los datos, nombre e institución del perfil, ID en Google del perfil, #Pu que muestra el número de publicaciones extraídas en cada perfil y las columnas Shiny y Rstudio que muestran el tiempo en segundo de la extracción de datos en cada método. Los resultados muestran que en ambos métodos se logró extraer la totalidad de las publicaciones por perfil. La aplicación web Shiny muestra un tiempo promedio de extracción de 28 segundos por 172 publicaciones extraídas, inferior a los 36 segundos promedios usados por el algoritmo en Rstudio.

Al utilizar la aplicación se contabilizó el tiempo de ejecución utilizando la función de tiempo t = proc.time() de R, almacenando en la variable "t" la duración del proceso, al terminar este se utilizó la función proc.time() – t, para restar el tiempo transcurrido desde su ejecución. En el método del algoritmo en Rstudio se contabilizó el tiempo de extracción

con un cronómetro, desde el inicio de su ejecución hasta mostrar los resultados para así medir el tiempo completo que le tomaba a quien ejecutaba el proceso.

TABLA I
COMPARACIÓN DE TIEMPO DE PERFILES EXTRAIDOS UTILIZANDO APLICACIÓN WEB EN SHINY VS ALGORITMO EN RSTUDIO (SEGUNDOS)

Nombre / Universidad	ID GS	#Pu	Shiny	Rstudio
Martinha Pereira Inst. Politécnico do Cavado e do Ave	Xw1WpZEAAAAJ	486	80	86
Paul Anderson George Fox University	5g0Bm20AAAAJ	345	55	63
Victoria Castro Rojas Univ.Alberto Hurtado	rpd7uCwAAAAJ	204	35	42
Raul Ramos-Pollan Universidad de Antioquia	QObKt9IAAAAJ	163	25	35
Suzana Maria Ratusznei Inst. Mauá de Tecnologia	hWc9IV0AAAAJ	150	25	32
José Otero Parra Univ. Europea Miguel de Cervantes	MviWnOQAAAAJ	98	16	27
Raul Becchio Univ.Nacional de Salta	50vUdEUAAAAJ	93	15	23
Consuelo Lobato-Calleros Univ. Autónoma Chapingo	7MRYMjUAAAAJ	83	13	22
Daniel Alejandro Barrio Univ.Nacional de Río Negro	hVc1_CgAAAAJ	61	10	18
Reinhardt Pinzon Adames Univ. Tecnológica de Panamá	1TlCxmUAAAAJ	41	7	16

En ambos métodos de prueba se utiliza el mismo algoritmo, solo que en diferentes entornos donde se pretende medir no solo el funcionamiento, sino la facilidad de uso de la herramienta. La diferencia de tiempo se debe a que, para mostrar la tabla de resultados, guardar los datos en CSV y generar el gráfico en el método del algoritmo en R, es necesario utilizar tres funciones adicionales propias del lenguaje R, lo cual hizo que el usuario consumiera entre 5-8 segundos adicionales.

En la **prueba de extracción de datos del perfil institucional en** GS, la Tabla II, muestra la columna Universidad, con el nombre de la institución, #Pe que indica el número de perfiles extraídos, #Pu que indica el número de publicaciones extraídas en cada perfil, las columnas Shiny y Rstudio muestran el tiempo en minutos de la extracción total del perfil institucional. El resultado muestra que en ambos métodos se logró extraer la totalidad de los perfiles como de sus publicaciones de cada institución. La Aplicación web Shiny también muestra un mejor tiempo de extracción en minutos, con un promedio de 19 minutos por cada 250 perfiles y 4400 publicaciones, mientras que con el algoritmo en Rstudio el tiempo fue de 21 minutos en promedio.

La diferencia de tiempo se debe a que en el método del algoritmo en Rstudio, es necesario primero ejecutar la función PubGS_research() para extraer los perfiles y la tabla de resultado de los perfiles debe ser utilizada en la siguiente función, PubGS_publications() para extraer las publicaciones de cada perfil. Para visualizar la tabla en ambos resultados se deben utilizar las funciones view(), la función write.csv() se utiliza para guardar los datos en CSV, pero se debe configurar la ruta en Rstudio donde se guardarán los datos, lo que también consume tiempo.

TABLA II
COMPARACIÓN DE TIEMPO DE PERFILES Y PUBLICACIONES INSTITUCIONALES EXTRAIDOS UTILIZANDO APLICACIÓN WEB EN SHINY VS ALGORITMO EN RSTUDIO (TIEMPO EN MINUTOS)

Universidad	#Pe	#Pu	Shiny	Rstudio
World Maritime University	814	13258	57	64
Universidad del Atlántico	189	3924	15	17
Universidad Tecnologica de Panama	188	2680	14	16
Instituto Mauá de Tecnologia	38	1028	4	5
George Fox University	30	1156	4	5

Al utilizar la Aplicación web en Shiny para extraer perfiles y publicaciones de una institución, en la figura 7, se muestra la interface web con los resultados en una tabla dinámica que permite visualizar los datos antes de guardar, también se habilitan dos botones, uno de color naranja para guardar las publicaciones en CSV y el otro de color verde para guardar los perfiles completos también en CSV.

En las pruebas realizadas se intentaron extraer los perfiles y publicaciones de 3 universidades: Universidad de Costa Rica, Universidad de la República en Uruguay y la Universidad de México, con más de 800 perfiles. El tiempo de extracción fue mayor al resto de las pruebas debido a la disminución de la velocidad de internet, por lo que el servidor de Google Scholar bloqueo las múltiples peticiones de extracción enviando un error de acceso 403.

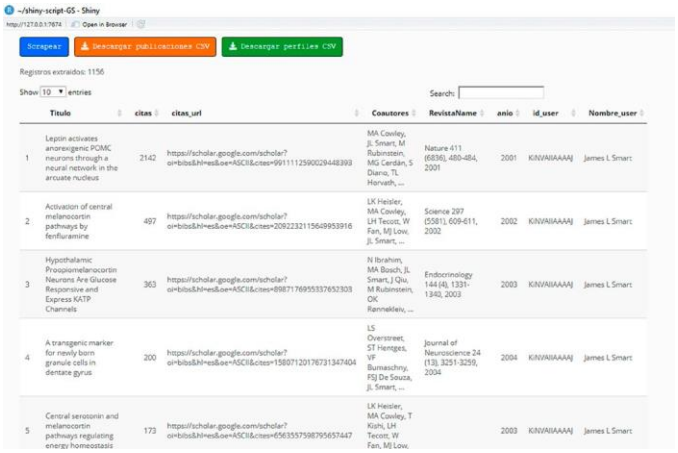


Fig. 7 Interface de los resultados al extraer perfil institucional de GS en aplicación web Shiny

Para la prueba de **extraer el perfil individual / institucional en GS** en la figura 8 muestra las seis tareas realizadas por los 10 usuarios sin conocimiento de R, solo realizó la extracción de un perfil individual e institucional.

En la tarea 1, relacionada con extraer el perfil individual, los 10 usuarios completaron la tarea con ambos métodos.

En la tarea 2 para extraer el perfil institucional, los 10 usuarios la completaron utilizando web Shiny, con el algoritmo en Rstudio solo 9 completaron la tarea. El usuario que no completó la tarea fue porque utilizó la URL del perfil institucional en la función de perfil individual.

En la tarea 3 para extraer las publicaciones del perfil institucional, los 10 usuarios la completaron utilizando web Shiny, con el algoritmo en Rstudio, 8 completaron la tarea. Los que no la completaron no recordaron que debían utilizar la tabla de resultados para extraer las publicaciones, por lo que volvieron a utilizar la URL del perfil institucional.

En la tarea 4 para visualizar y filtrar los datos de la tabla de las publicaciones del perfil institucional, los 10 usuarios la completaron utilizando web Shiny, pero, en el algoritmo en Rstudio, solo 1 completo la tarea. Los que no completaron la tarea lograron visualizar la tabla con la función view(), pero no recordaron que era posible filtrar los datos desde la vista de tablas en Rstudio.

En la tarea 5 para guardar los datos de la tabla de las publicaciones del perfil institucional, los 10 usuarios la completaron utilizando web Shiny, pero, en el algoritmo en Rstudio, solo el 50% completo la tarea. Los que no completaron la tarea no recordaron utilizar la función write.csv() y los que recordaron utilizarla, no identificaron que datos debían guardar de los perfiles y publicaciones.

En la tarea 6 para visualizar el gráfico en el perfil individual, los 10 usuarios la completaron utilizando web Shiny, pero, en el algoritmo en Rstudio, solo 2 completaron la tarea utilizando la función graficoIndividual(), pero no pudieron guardar el gráfico. Los que no completaron la tarea no recordaron que tabla de resultados utilizar del gráfico de las tres tablas extraídas, publicaciones individuales, perfiles y publicaciones institucionales.

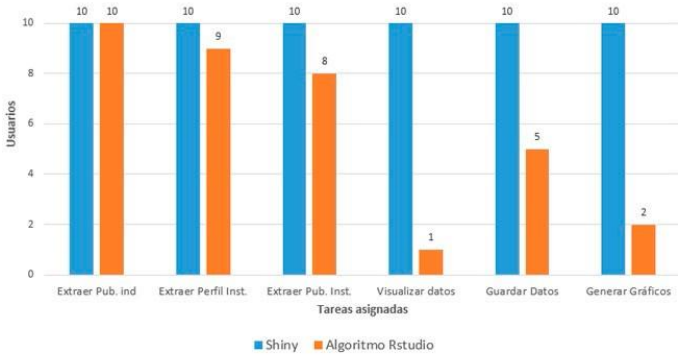


Fig. 8. Gráfico de las 6 tarea utilizando el algoritmo en R y la aplicación web en Shiny para extraer datos en GS

IV. Conclusiones

Los resultados mostrados en este trabajo indican que las funciones de webscraping creadas con código del lenguaje R se puedan integrar en un ambiente web utilizando el paquete Shiny sin desmejorar su funcionamiento, eficacia y tiempo de procesamiento. La aplicación web permitió mejorar el uso de este recurso, generando en el usuario una mejor interacción y facilidad de uso.

La implementación de la aplicación en Shiny de extracción de datos en Google Sholar según las pruebas realizadas en la extracción individual e institucional, mostraron que entre más simple sea la interface disminuye el tiempo en segundos o minutos al extraer, visualizar y descarga la tabla y gráficos de resultados.

En las pruebas realizadas a los 10 usuarios, en las 6 tareas utilizando la aplicación web Shiny, todas fueron ejecutadas sin dificultad, pero utilizando el algoritmo en Rstudio solo 1 tarea pudo ser ejecutada. El mayor problema de la no conclusión de las tareas fue que los usuarios debían recordar no solo las funciones generadas en el algoritmo, sino debía conocer funciones en R para visualizar, guardar y filtrar los datos, por lo que hacer que un usuario novato en R utilice el algoritmo en RStudio maximiza el tiempo de extracción y aumenta los errores propios del desconocimiento del lenguaje.

Nuestra experiencia en el uso del paquete Shiny para crear aplicaciones web, nos permite decir que minimiza el tiempo de programación de una herramienta, que no solo puede integrar tablas y gráficos, sino análisis estadísticos y predictivos que potencian la creación de algoritmos en este lenguaje y un mayor alcance en el uso de estas aplicaciones en un entorno web amigable.

Dentro de las limitaciones tanto de la aplicación como del algoritmo, es que, en la extracción de los perfiles a nivel institucional, si la velocidad de internet es inferior a la descrita en la metodología o el número de perfiles institucionales es superior a 800, Google Scholar bloque la extracción, por lo

que es necesario esperar varias horas para que GS de autorización para volver a extraer.

Esta herramienta se ha convertido en un recurso útil para nuestra institución debido a que dos de los indicadores del Plan de desarrollo institucional estaba relacionado con el número de perfiles por unidad y el numero de publicaciones, proceso que nos demoraba 1 día en realizar la extracción y ahora nos toma menos de 15 minutos, el cual creemos también puede ser de beneficio para otras instituciones. Este año realizamos un libro sobre perfiles de investigadores de Panamá con 860 perfiles de 48 instituciones.

TRABAJOS FUTUROS

Actualmente la aplicación solo se puede ejecutar en local, pero se puede descargar el código completo desde GitHub para hacer uso de esta, como también para mejorar su código https://github.com/dannymuPTY/shinyGS.

Se contempla mejorar las opciones de la aplicación creando gráficos de los perfiles con más citaciones y publicaciones cuando se realice la extracción del perfil institucional. Algunos perfiles institucionales contienen hasta 1000 perfiles individuales, por lo que se contempla crear la opción para extraer un limité de perfiles por institución en caso de que así lo requiera el usuario. Dentro de las limitaciones se pretende utilizar otro paquete en R para indicar a GS que el algoritmo no es dañino, esto a través de un mensaje de CAPTCHA que genera Google y que el usuario pudiera darle click para proseguir la extracción sin bloqueo.

Aunque la aplicación en Shiny ha resultado con mayor aceptación para los usuarios, se realizará un paquete en R con este algoritmo para que pueda ser descargado por quienes tengan un conocimiento más avanzado del lenguaje.

REFERENCIAS

[1] J. Abbate, "Internet : Su Evolución Y Sus Desafíos," 2005.

[2] J. Mingers, J. R. O'Hanley, and M. Okunola, "Using Google Scholar institutional level data to evaluate the quality of university research," *Scientometrics*, vol. 113, no. 3, pp. 1627–1643, 2017.

[3] J. R. Sánchez Carballido, "Perspectivas de la información en Internet: ciberdemocracia, redes sociales y web semántica," *Zer-Revista Estud. Comun.*, vol. 13; n.° 25, pp. 61–81, 2011.

[4] A.-W. Harzing, "Google Scholar Profiles: the good, the bad, and the better," 2018. [Online]. Available: https://harzing.com/blog/2018/11/google-scholar-citation-profiles-the-good-the-bad-and-the-better.

[5] L. Gil, "Google Scholar: el buscador académico con mayor impacto," *Soc. Media en Investig.*, 2015.

[6] D. Torres-Salinas, R. Ruiz-Pérez, and E. Delgado-López-Cózar, "<i>Google Scholar</i> como herramienta para la evaluación científica," *El Prof. la Inf.*, vol. 18, no. 5, pp. 501–510, 2009.

[7] R. Dávalos-Sotelo, "Una forma de evaluar el impacto de la investigación científica," *Madera y Bosques*, vol. 21, pp. 7–16, 2015.

[8] R. Mugnaini and L. Strehl, "Recuperação e impacto da produção científica na era google: uma análise comparativa entre o google acadêmico e a web of science 10.5007/1518-2924.2008v13nesp1p92," *Encontros Bibli Rev. eletrônica Bibliotecon. e ciência da informação*, vol. 13, no. 1, pp. 92–105, 2008.

[9] SerpApi, "Google Search API," 2020. [Online]. Available: https://serpapi.com/google-scholar-api.

[10] Harzing, "Publish or Perish," 2016. [Online]. Available: https://harzing.com/resources/publish-or-perish.

[11] A. Sharma and P. Gupta, "Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining," *Int. J. Adv. Res. ...*, vol. 1, no. 8, 2012.

[12] D. Murillo, D. Saavedra, and E. Quintero, "Extracción de datos de perfiles en Google Scholar utilizando un algoritmo en el lenguaje R para hacer minería de datos," *I+D Tecnológico*, vol. 14, no. 1, pp. 94–104, 2018.

[13] S. Munzert, R. Christian, and P. Meißner, *Automated Data Collection with R A Practical Guide to*, John Wiley. 2015.

[14] R. Ibar-Alonso, C. Cosculluela-Martínez, E. Sánchez-Muñoz, and F. Rodríguez-Lázaro, "La enseñanza con Shiny," in *XV Encuentro Internacional Anales de ASEPUMA nº 27*, 2015, pp. 1–22.

[15] S. Shi, C. Liu, Y. Shen, C. Yuan, and Y. Huang, "AutoRM: An effective approach for automatic Web data record mining," *Knowledge-Based Syst.*, vol. 89, pp. 314–331, 2015.

[16] J. M. Carrilo Corpas, "Aplicación web interactiva en 'Shiny' para el análisis espacio-temporal de riesgos de mortalidad por cáncer," 2017.

[17] G. B. Bosoni and F. R. Bruzzone, "Uso de RStudio para Estadística Univariada en Ciencias Sociales," no. June, 2018.

[18] D. Murillo and D. Saavedra, "Web Scraping de los Perfiles y Publicaciones de una Afiliación en Google Scholar utilizando Aplicaciones Web e implementando un Algoritmo en R," 2017, pp. 8–15.

[19] J. Casino Durán, "Aplicación web creada con shiny para el análisis de la variabilidad de las materias primas en la fabricación de piensos," 2017.

[20] J. Wojciechowski, A. M. Hopkins, and R. N. Upton, "Interactive pharmacometric applications using R and the Shiny package," *CPT Pharmacometrics Syst. Pharmacol.*, vol. 4, no. 3, pp. 146–159, 2015.

[21] C. Fryer and P. Guill, "Interactive Applications for Modeling and Analysis with Shiny," 2015.

[22] M. J. Peralta González, I. Maylín, F. Guzmán, I. Orlando, and G. Chaviano Ii, "Criterios, clasificaciones y tendencias de los indicadores bibliométricos en la evaluación de la ciencia Criteria, classifications and tendencies of bibliometric indicators in the evaluation of the science," *Rev. Cuba. Inf. en Ciencias la Salud*, vol. 26, no. 3, pp. 290–309, 2015.

[23] A. Barceló, M. Grimalt, and J. Binimelis, "Análisis bibliométrico de los estudios geográficos de la caza en España (1978-2015)," *Boletín la Asoc. Geógrafos Españoles*, vol. 74, pp. 301–332, 2017.

[24] E. Orduña-Malea, "Aplicaciones métricas de Google Scholar para la evaluación del impacto científico," *Actas las 4ª Jornadas Intercamb. y Reflexión acerca la Investig. en Bibl.*, vol. 4, no. April, pp. 29–30, 2015.

[25] Webometrics, "TRANSPARENT RANKING: Top Universities by Citations in Top Google Scholar profiles," 2020. [Online]. Available: https://www.webometrics.info/en/transparent.