DOI: http://doi.org/10.31048/1852.4826.v16.n2.41094

RMA
Dossier

Raspando la Arqueología: Una Aproximación Metodológica desde el *Web Scraping* y *Text Mining*

Scraping Archaeology: A Methodological Approach from the Web Scraping and Text Mining

Humberto Aguilar

Instituto de Antropología de Córdoba – CONICET, Museo de Antropología, FFyH – UNC. E-mail: humberto.aguilar01@gmail.com

Resumen

A medida que la cantidad de información disponible en la web aumenta, también lo hace la tarea de localizarla y analizarla, por lo cual realizar esta tarea de forma manual puede ser costosa en función al tiempo y esfuerzo invertido. Aunque los buscadores y los motores de bases de datos pueden ayudar a encontrar la información requerida, en infraestructuras digitales grandes donde los resultados de búsqueda se cuentan por millares – o más– se precisan de nuevas herramientas para obtener el contenido buscado de manera efectiva. Este trabajo propone la aplicación de Web Scraping y Text Mining como insumos metodológicos para poder compilar y procesar grandes volúmenes de datos en infraestructuras digitales de una forma más automatizada. La automatización de ambos procesos aporta una gran ventaja al analizar corpus textuales de miles de registros lo cual simplifica de manera significativa la obtención de diferentes tipos de datos, facilitando el trabajo considerablemente. Se espera que esta contribución permita ampliar las posibilidades de la comunidad arqueológica en clave de una metodología novedosa para la obtención y el manejo de datos estructurados y no estructurados que pueden ser integrados a las investigaciones de la comunidad arqueológica en general.

Palabras clave: R; Web scraping; Text mining; Análisis de datos; Arqueología digital.

Abstract

As the amount of information available on the web increases, so does the task of locating and analysing it, and performing this task manually can be costly in terms of time and effort. Although search engines and database engines can help to find the required information, in large digital infrastructures where search results are in the thousands - or more - new tools are needed to effectively retrieve the searched content. This paper proposes the application of Web Scraping and Text Mining as methodological inputs to be able to compile and process large volumes of data in digital infrastructures in a more automated way. The automation of both processes provides a great advantage in analysing textual corpora of thousands of records, which significantly simplifies the collection of different types of data, facilitating the work considerably. It is hoped that this contribution will expand the possibilities of the archaeological community in terms of a novel methodology for the collection and handling of structured and unstructured data that can be integrated into the research of the wider archaeological community.

Keywords: R; Web scraping; Text mining; Data analytics; Digital Archaeology.

Introducción

Una parte sustancial de la labor científica arqueológica consiste en generar y socializar información relevante, lo que incluye la búsqueda de cualquier tipo de dato que contribuya a esos fines. Si bien tradicionalmente nuestra práctica suele estar asociada a la búsqueda de información que permita esclarecer, con el mayor detalle posible, los modos y formas de vida de los individuos y sociedades pasadas a través del estudio de su cultura material, ha sido con el advenimiento de las tecnologías digitales que ha cambiado la forma en la que los arqueólogos experimentamos otros modos en los cuales producimos y compartimos nuestros hallazgos. Es en torno a estas transformaciones que Richards (2009: 29) sostiene que hemos sido testigos del desarrollo de una serie de sistemas para estandarizar el registro de nuestros datos, pero no de un sistema único que todos utilicen, por lo cual existe una gama heterogénea de herramientas y formas para almacenar la información relevada. A pesar de ello, los datos que uno requiere no siempre se encuentran disponibles de manera ordenada o en otros casos se presentan en grandes volúmenes, lo cual dificulta su búsqueda, extracción y gestión de tal manera

Editado por A. Izeta y R. Cattáneo

Recibido 02-05-2023. Recibido con correcciones 25-06-2023. Aceptado 17-07-2023

DOI: http://doi.org/10.31048/1852.4826.v16.n2.41094

que es necesario implementar nuevas herramientas que contribuyan a refinar y ampliar la información que se encuentra disponible.

Una parte considerable de la información requerida es asequible a través de internet y se encuentra almacenada en grandes infraestructuras digitales que contienen datos de todo tipo y tamaño, algunas de ellas son estrictamente científicas y específicas como el Repositorio Digital Institucional Suquia, Digital Index of North American Archaeology (DINAA) o ARIADNEplus las cuales se encuentran avocadas al almacenamiento y sistematización de datos arqueológicos, mientras que otras conviven con nuestra cotidianeidad y son ampliamente conocidas por el público en general como lo son las redes sociales. Si bien muchas de estas herramientas permiten realizar una búsqueda basada en filtros para refinar la información solicitada y hacer una búsqueda más ágil, en muchos casos esta tarea se ve dificultada debido al gran caudal de información disponible, especialmente en sitios con un enorme tráfico de usuarios. Por lo tanto, nos encontramos ante un problema que supone una doble dificultad marcada por la falta de precisión y el exceso de tiempo necesario para obtener información útil lo cual amerita herramientas eficientes para el procesamiento y administración de datos. En este trabajo, nos referiremos a dos de ellas: el *Web Scraping* y *Text Mining* ("raspado web" y "minería de textos" en español) y, a través de diferentes escenarios empíricos que involucran a la arqueología como disciplina y práctica, demostraremos cómo éstas funcionan y son útiles en la obtención y el posterior análisis cuantitativo y cualitativo de los datos. Este enfoque computacional, para nuestro caso, abarca todos los pasos del proceso investigativo: la recolección de la información, su transformación desde un formato no estructurado a uno estructurado (tabular), la limpieza y normalización del corpus, la datificación, el análisis y la visualización de los resultados (Laitano y Nieto, 2022: 156).

Para el primer caso de aplicación que aquí se plantea nos interesa indagar acerca de cómo diferentes usuarios tanto arqueólogos profesionales, aficionados o público en general interactúan mediante 280 caracteres en Twitter[1] mediante el uso de *hashtags* específicos que vinculan a la arqueología como disciplina y al mismo tiempo como es definida en el marco de una infraestructura digital masiva que para este estudio reúne las opiniones y comentarios de la comunidad anglófona e hispanohablante para discutir el estado y el relato construido en esta red social en torno a nuestra disciplina, de tal modo que siguiendo a Izeta y Cattáneo (2018: 2) nos interesa indagar si lo digital es algo que estamos reflexionando o si es algo sobre lo que debamos reflexionar pensando en función de cuál es el papel que juega lo digital cuando se encuentra mediado entre arqueólogos y no arqueólogos y sus reproducciones dentro de una red social. Mientras que para el segundo caso de aplicación nos interesa a través de una aproximación cuantitativa obtener información acerca al quehacer de los propios arqueólogos en comparativa con sus colegas mediante un trabajo de *web scraping* basado en sus producciones científicas catalogadas en sus perfiles personales del portal digital de CONICET tomando como eje las producciones de investigadores y becarios, las revistas que eligen a la hora de publicar y el idioma elegido, lo cual supone un ejercicio de "captura de datos" (*data capture* en la fuente original) en términos de Richardson (2019) más eficiente que realizado de forma manual.

Si bien estas metodologías son poco convencionales en la arqueología, ambas son herramientas basadas en el registro y análisis de datos, tarea que los arqueólogos han adoptado desde el origen de la disciplina, y que han encontrado soporte en las tecnologías digitales al menos desde la década de 1970 (Ali et al., 2022). De tal manera este producto encuentra un marco en la Arqueología Digital la cual parte como aquella disciplina que explora las relaciones básicas que los arqueólogos tienen con dicha tecnología en la evaluación del impacto que tienen sobre las distintas formas en las que ésta se realiza como tal (Daly y Evans, 2006) que en este caso versa sobre casos de aplicación que involucran a la arqueología en diferentes dimensiones de su espectro en el mundo de lo digital y que dicho sea de paso esta línea de investigación a nivel teórico no ha tenido un análisis teórico pormenorizado (Richardson, 2013).

Sobre los Datos

El término "dato" es uno de los más utilizados en el mundo dentro del vocabulario científico, pero, a pesar de que como expresión forma parte de nuestro léxico habitual, es tan polisémico que parece difícil entenderlo independientemente de su contexto (Bordignon y Maisonobe, 2022). En particular, la mayoría de los especialistas comparten la idea de que los datos no tienen un significado estable: su importancia y utilidad varían a lo largo del proceso de investigación (Bordignon y Maisonobe, 2022). Por ejemplo, los datos que se recolectan en la fase de exploración pueden ser muy diferentes de los datos que se recopilan en la etapa de análisis.

Para el caso que nos atañe, referimos al uso de un tipo de dato particular que es el que nos han provisto las herramientas utilizadas, los datos no estructurados. Estos no se ajustan a un formato o estructura predefinida, lo que dificulta su procesamiento y análisis por medios convencionales. Según Feldman y Sanger (2006), los datos no estructurados se refieren a cualquier tipo de información que no esté organizada en una tabla o

[1] El 23 de julio de 2023, posterior a la fecha de aprobación del manuscrito, Twitter pasó a llamarse X. De manera operativa e instrumental se seguirá utilizando la primera designación dado a que este estudio y sus fuentes fueron utilizadas y consultadas en momentos previos a este cambio de marca.

DOI: http://doi.org/10.31048/1852.4826.v16.n2.41094

base de datos con campos claramente definidos y que pueden incluir texto, imágenes, audio, video y cualquier otro dato que no esté organizado en una estructura lógica y coherente y que pueden ser encontrados en una amplia variedad de infraestructuras digitales, incluyendo sitios web, correos electrónicos, mensajes de texto, documentos y redes sociales.

La complejidad de los datos no estructurados se debe a que no se presentan en un formato uniforme y coherente, lo que dificulta su análisis y comprensión. Por lo tanto, su análisis requiere herramientas y técnicas especiales, como el procesamiento del lenguaje natural, la minería de datos y la inteligencia artificial para extraer información valiosa y significativa para poder y decodificar las interacciones observadas en línea que están incrustadas en la vida cotidiana y así comprender el significado, el uso y el valor aplicado a la información arqueológica e histórica en la multitud de prácticas sociales que encontramos en internet (Richardson, 2019).

R Project, *Web Scraping* y *Text Mining*

La mayoría de los investigadores utilizan computadoras como una herramienta esencial en su trabajo cotidiano, si bien el uso de estos dispositivos permiten un mayor almacenamiento y agilidad a la hora de trabajar con la información recabada a menudo se suele prestar escasa atención a la hora de cómo organizar los archivos que allí se encuentran contenidos, lo que dificulta la reproducción de los resultados y el intercambio de todo nuestro proceso de análisis con otros colegas y con nosotros mismos, por lo cual R Project y todo su conjunto de funcionalidades son una herramienta útil para el manejo eficiente de datos (Marwick et al., 2018). Por un lado, nos ofrece un modelo analítico computacional (Laitano y Nieto, 2022) mientras que, por el otro, nos permite avanzar hacia el propósito de nuestro trabajo en otras facetas donde la arqueología, como concepto y praxis, puede ser sometida a otro tipo de análisis donde los protagonistas son los usuarios y la comunidad arqueológica en general.

Según R Core Team (2020) R es un entorno y lenguaje de programación libre, abierto y gratuito que proporciona una amplia variedad de herramientas estadísticas y gráficas que permiten a sus usuarios definir sus propias funciones al ser un lenguaje de programación interpretado. Forma parte del proyecto GNU (*General Public License*) lo cual quiere decir que el *software* puede ser distribuido, copiado y editado libremente. R funciona a través de paquetes y librerías creados por su propia comunidad y almacenados en la CRAN (*Comprehensive R Archive Network*) por lo por cual es altamente extensible y personalizable, lo que permite a los usuarios crear soluciones a medida para sus necesidades específicas. En nuestro caso, hemos utilizado *Web Scraping* y *Text Mining* en clave metodológica sobre los casos de aplicación mencionados.

El método *Web Scraping* permite rastrear y extraer datos no estructurados o poco estructurados a un formato estructurado mediante robots o bots que automatizan todo el procedimiento sin que se requiera una acción manual más pormenorizada (Mártinez et al., 2019; Hernández et al., 2015). Este proceso se realiza en dos etapas: una donde se realiza una consulta de datos hacia un sitio, los cuales se guardan de manera local, y otra que implica el análisis de estos datos para obtener información relevante (Hernandez et al., 2015). De tal forma que "scrapear" una plataforma digital permite una obtención de datos en términos de tiempo, más eficiente que de forma manual y cuyos objetos varían según los lineamientos de la investigación de los autores, si bien los estudios vinculados a esta temática son escasos pueden citarse algunos ejemplos: Demir y colaboradores (2023) aplicaron esta metodología para buscar imágenes de patrimonio cultural almacenadas en internet cuyo objetivo era identificar los objetos más populares en varios sitios antiguos del sur de Turquía de manera automatizada, mientras que en otra investigación se utilizó para recopilar imágenes de diferentes fuentes relacionadas con el comercio de restos humanos en cuentas asociadas a vendedores en Instagram (Graham et al., 2020).

Como las redes sociales almacenan y hacen circular grandes volúmenes de información que impiden que los datos puedan ser analizados eficientemente de forma manual, el *Web Scraping* es un insumo metodológico apropiado para acceder a éstos, a la vez que se traduce en una mayor confiabilidad y, como veremos, disminuye notablemente los sesgos que puedan desviar la interpretación, por lo tanto nos permite plantear nuestros estudios y sus resultados de manera más acertada (Arcila-Calderón et al., 2016).

El *Text Mining* refiere al proceso de explorar y descubrir patrones y relaciones dentro de grandes conjuntos de datos de texto (Feldman y Sanger, 2006; Feldman y Dagan, 1995). Esto implica el uso de técnicas de procesamiento del lenguaje natural (*Natural Language Processing*) y minería de datos para analizar grandes cantidades de texto no estructurado y extraer su contenido textual en una estructura organizada y en un formato manejable que permita una interpretación crítica. Una de las aplicaciones más comunes del *Text Mining* es la extracción de información específica de grandes cantidades de texto: como nombres, fechas, números, ubicaciones, entre otros datos. Al aplicar técnicas de procesamiento del lenguaje natural y minería de datos, es posible analizar grandes cantidades de información textual, extraer patrones y relaciones entre diferentes categorías de datos. En la investigación arqueológica puede ser un insumo útil debido a que los arqueólogos generan una gran cantidad de información en forma de

DOI: http://doi.org/10.31048/1852.4826.v16.n2.41094

informes y literatura gris que a menudo no se publica o no se encuentra accesible facilmente, el *text mining* permite extraer información acerca de estos documentos, lo que facilita la búsqueda y la identificación de patrones en la literatura arqueológica, además permite integrar la información arqueológica con otras bases de datos y realizar búsquedas bibliográficas de manera más eficiente, esto ayuda a los investigadores a localizar textos relevantes y a realizar comparaciones y análisis de documentos de manera más rápida y precisa, otra ventaja del *text mining* en arqueología es que puede ayudar en la identificación de términos clave, como lugares, fechas y otros nombres, en los informes arqueológicos, lo cual facilita la clasificación y organización de los documentos (Richards et al., 2015). En resumen, ofrece una forma eficiente de extraer información de grandes volúmenes de texto, facilitando la búsqueda, el análisis y la integración de la información arqueológica.

Twitter, Usuarios y Arqueología

Las redes sociales permiten la interacción y la comunicación al facilitar la conexión entre personas, superando las barreras de tiempo y distancia y brindando respuestas inmediatas (Grzegorczyk y Salerno, 2022: 35). Para el primer caso de aplicación se utilizó *Twitter* la cual es una red social orientada al microblogueo, es decir una plataforma que permite publicar cadenas de texto de corta longitud, condensadas en un máximo de 280 caracteres en cuentas ordinarias y hasta 4000 adquiriendo los servicios de Twitter Blue[2] (Twitter Blue, 2023) y que permite articular conversaciones globales en torno a *hashtags* que no necesariamente se publican sincrónicamente (Grzegorczyk y Salerno, 2022). En términos de Van Dijck (2016: 74), Twitter se presenta como una caja de resonancia de conversaciones aleatorias, es decir, un soporte online para opiniones de masas donde, ante la mirada del público, se generan emociones colectivas, nacen y mueren tendencias y pululan flujos de datos con el objetivo de promover ciertos usos y usuarios por encima de otros. Esto nos lleva a pensar que el análisis de los tuits de los usuarios permite mostrar una aproximación al abanico y la polarización de ideas de cualquier comunidad sobre un tema en particular.

Para hacer una extracción de datos en Twitter es necesario interactuar con su API (*Application Program Interfaces* en inglés) es decir un conjunto de protocolos y funciones que median las interacciones entre los usuarios y las aplicaciones web. El uso de APIs nos permite manipular y organizar datos de la plataforma con la que estamos trabajando. La API de Twitter es pública y permite a los usuarios registrados interactuar con ella, aunque de forma limitada: una cuenta gratuita sólo permite recuperar hasta 1000 tuits por consulta[3], pero existen también cuentas de desarrollador pagas para acceder a un mayor caudal de datos. Para este trabajo se optó por su versión gratuita a través del paquete *rtweet* el cual se enlaza con nuestro usuario en la plataforma de *Twitter Developer Platform* (un conjunto de herramientas y recursos que Twitter ofrece a los desarrolladores para crear aplicaciones y servicios que interactúan con la plataforma de Twitter) para acceder a su API y extraer la información que necesitamos (Kearney, 2019).

Durante un período de 30 días en diferentes intervalos según la frecuencia de las palabras buscadas[4], se recuperaron tuits provenientes de los *hashtags* #Arqueología y #Archaeology con el objeto de obtener una aproximación hacia los patrones de comportamiento e interacción por parte de aquellos usuarios que manifestaron interés por la arqueología y cuyas cuentas no necesariamente se asocian con arqueólogos profesionales, organismos educativos y de investigación, comúnmente entendidos como parte del *Twitter for scholary networking* o *Academic Twitter* (Allés Torrent et al., 2020) por lo cual este análisis permite contar con una red de voces más heterogénea que permite darle forma al relato arqueológico en la red social.

Para el primer caso se recuperaron 1364 tuits originales (es decir filtrando los retuits) utilizando el *hashtag* #Arqueología, lo cual nos habla de poco más de 400 tuits por consulta y, al mismo tiempo, la de la poca incidencia del término, teniendo en cuenta que el límite de tuits que nos posibilita extraer la API es mucho mayor. Notamos que también la extracción de los datos que nos ofrece el paquete no fue *case sensitive*, es decir, que la palabra extraída no distinguía entre mayúsculas y minúsculas, por lo cual obtuvimos un volumen de datos mayor al contar con múltiples representaciones del tipeo del *hashtag*. Como muestra la Figura 1 el volumen de publicaciones con el *hashtag* #Arqueología fue homogéneo, no alcanzando más de 15 menciones en intervalos de 3 horas por día, a excepción de una anomalía de tuits durante el 26 enero, que incrementó su frecuencia. Calvo y Aruguete (2020) señalan a la presencia de operadores *trolls* y *bots* que afectan a la legibilidad de las redes, la legibilidad se refiere al proceso de verificar el comportamiento de los usuarios y se evidencia a través de sus interacciones, sin embargo este fenómeno, particularmente, no se correspondió con una tendencia a nivel global, que podría haber incrementado su visibilidad sino que se debió a la reiteración de un mismo tuit por parte de una sola cuenta, tal vez automatizado por un *bot* o realizado manualmente, que logró triplicar el volumen del *hashtag* y, en consecuencia, su viralización en la red

[2] Actualmente llamado X Premium

[3] Tras los cambios en la plataforma la interacción con la API de la misma aún sigue siendo gratuita, sin embargo han habido cambios, para este caso ahora solo permite obtener 100 tuits por consulta

[4] A pesar de que la librería rvest de R studio solo permite la extracción de hasta 1000 tuits por consulta, el número total de tuits recolectados no presenta valores de la muestra mayores debido a que los hashtags seleccionados presentaban una baja frecuencia de aparición en la plataforma. Por esta razón, fue necesario realizar las consultas en intervalos de 5 a 10 días para evitar la repetición de tuits y aumentar la cobertura temporal del estudio.

DOI: http://doi.org/10.31048/1852.4826.v16.n2.41094

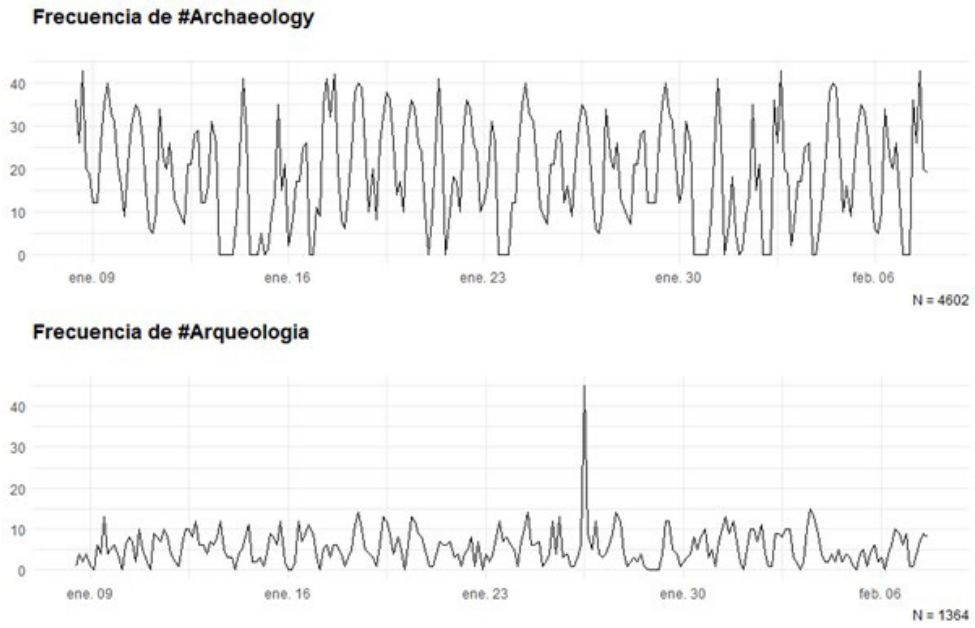

Figura 1. A) Frecuencia de *hashtags* durante un período de 30 días en intervalos de 3 horas. A) #Archaeology y B) #Arqueología

Figure 1. *A) Frequency of hashtags over a 30-day period at 3-hour intervals.. A) #Archaeology and B) #Arqueología.*

social. De tal modo que la visibilidad de cualquier tipo de publicación si puede ser susceptible de ser modificada en términos cuantitativos y de su significado según cual sea el contenido del discurso que sea viralizado, además si bien Twitter es una plataforma enorme con millones de cuentas y un caudal de información que se actualiza constantemente, la comunidad de usuarios que refieren a la arqueología es considerablemente baja si tomamos como antecedente el hecho de que una sola cuenta puede ser capaz de alterar su flujo de información.

Para el otro caso, mediante el *hashtag* #Archaeology se obtuvieron más de 4000 tuits que tampoco obedecieron a los parámetros del patrón *case sensitive*. Esto ubica al inglés como lengua preponderante en la representación general de nuestra disciplina en esta red social. Como puede verse en la Figura 1, el flujo de interacciones entre los usuarios que utilizaron la mención muestra mayor cantidad de tuits por intervalos con frecuencias homogéneas que en el caso de los tuits en español y, además, a primera vista no parecen haberse visto afectadas por otros agentes como la anomalía anteriormente mencionada.

También fue posible obtener una nube de palabras (Figura 2) mediante la aplicación de un algoritmo basado en el minado de textos para reflejar las 50 palabras más utilizadas en ambos idiomas, a través de los tuits que contenían la mención que buscamos con el objeto de tener una representación más detallada de aquellas palabras que eran más recurrentes Twitter y que de alguna manera también nos permiten obtener una imagen de la representación del léxico y la elección de palabras que tienen los usuarios para referirse a la arqueología y que contribuyen a modelar su relato digital a través de una búsqueda orientada a objetos específicos (Ávido y Vitores, 2018).

El proceso consiste, en primer lugar, en aplicar filtros a través de líneas de código específicas para eliminar links, emojis y otros signos no textuales e ir depurando, así, la base de texto. Luego es preciso "tokenizar" todo el conjunto de texto que tenemos o, dicho de otro modo, dividir el texto en unidades más pequeñas llamadas *tokens* que corresponden a palabras. Finalmente, nuestra lista de *tokens* es sometida nuevamente a un último filtro para obtener un gráfico con una representación de palabras más armónica, para ello se eliminan las *stop words* (a través de una base de datos propia) que son palabras recurrentes del lenguaje como los artículos, conjunciones, pronombres y preposiciones que no demuestran las inferencias, valoraciones, conceptos, y acciones vinculadas a los hashtags y cuya presencia impediría el análisis sustancial de los datos.

En general, tanto en inglés como en español se vio una representación gráfica de los sustantivos, adjetivos y verbos que conforman la nube vinculada hacia aspectos del quehacer y del uso del lenguaje concordantes con el oficio de la arqueología, aludiendo en su mayoría a cronologías, descubrimientos, hallazgos o la cultura material. A pesar de ello, es notable destacar que, en menor medida, en la nube de palabras de los *tokens* en español aparecen asociaciones que nada tienen que ver con una visión cientificista hacía lo que entendemos por arqueología, de tal manera que hay un puñado que refieren a lo que Schadla-Hall (2004) define como Arqueología Alternativa, es decir aquella que no está de acuerdo con los hechos generalmente aceptados que usan los arqueólogos para reconstruir el pasado y suele estar ligada a temas como las pseudociencias, fenómenos geológicos o astronómicos entre otros. Por ejemplo, observamos una vinculación con las palabras "marte", "intelegentelos" e "intelegentela", y a través del raspado

DOI: http://doi.org/10.31048/1852.4826.v16.n2.41094

Figura 2. Nube de palabras compuesta por los tokens de los tweets que contenían el *hashtag* #Arqueología y #Archaeology. A) En español y B) en inglés

Figure 2. *Word cloud composed of tokens of tweets containing the hashtag #Archaeology and #Archaeology. A) In Spanish and B) in English*

de tuits, verificamos que se trataban de publicaciones que aludían al espectro de una arqueología no científica. A pesar de ser sólo tres palabras, formaban parte de las 50 más comunes en un conjunto previamente depurado de poco más de 6000 *tokens*. Este dato no es menor si tomamos como punto de referencia a Hernando (2002) que, desde la Arqueología de la Identidad, considera a la disciplina no solamente como la cultura material, ni los aspectos concretos y visibles de las culturas del pasado, sino también a la reflexión general sobre el modo en que los seres humanos adquieren una imagen del mundo. Por lo cual este dato puede ser útil, entre otras cosas, para indagar acerca de cómo la arqueología puede ser moldeada por los usuarios (arqueólogos, no arqueólogos y cuentas académicas y científicas) dentro de un espacio en particular como lo es *Twitter* por un período de tiempo determinado y que para este caso de aplicación encuentra la difusión de mensajes que ligan a la arqueología con términos anticientificistas que son susceptibles de cobrar relevancia, más aun teniendo en cuenta de que la visibilidad de la disciplina en términos cuantitativos como hemos visto para ambos idiomas, es más bien acotada, prueba de ello es que todas consultas realizadas para la extracción de tuits en ningún momento alcanzaron los 1000 tuits, es decir, el límite máximo de recolección de información que nos brinda la API en su versión gratuita.

Raspando al IDACOR: ¿Qué publican sus miembros?

Según el portal web[5] del Instituto de Antropología de Córdoba (IDACOR), las Ciencias Antropológicas en Córdoba han tenido un largo desarrollo institucional con un gran incremento de recursos humanos profesionalizados en los últimos años. Al mismo tiempo, en conjunto con el Museo de Antropologías, se desarrollan tareas de investigación, docencia y extensión entre más de 100 personas, algunas de ellas vinculadas al CONICET. Teniendo en cuenta esta información y, haciendo especial hincapié en la diversidad de enfoques y líneas de investigación de sus integrantes, nos pareció adecuado aplicar nuevamente el *Web Scraping* en los perfiles de CONICET de 81 de sus miembros (32 investigadores y 49 becarios, tanto doctorales como posdoctorales) para obtener información acerca de dónde publican sus artículos[6].

Para ello se utilizó el paquete *rvest*, el cual permite obtener información HTML y XML de forma programática

[5] Información disponible en https://idacor.conicet.gov.ar/

[6] Fueron descartados aquellos miembros sin vinculación con CONICET por no contar con un repositorio que registre sus producciones científicas de forma ordenada.

DOI: http://doi.org/10.31048/1852.4826.v16.n2.41094

Figura 3. Representación del total de publicaciones por disciplina académica entre becarios e investigadores. A) Miembros becarios por disciplina, B) Miembros investigadores por disciplina, C) publicaciones de becarios por disciplina y D) publicaciones de investigadores por disciplina.

Figure 3. *Representation of total publications by academic discipline among fellows and researchers. A) Fellow members by discipline, B) Researcher members by discipline, C) Fellow publications by discipline and D) Researcher publications by discipline.*

y automatizada. Con este paquete es posible realizar una variedad de tareas, desde la extracción de datos simples como nombres y direcciones de una página web hasta la descarga de archivos y obtención de datos de páginas dinámicas. Gracias a ello, fue posible recabar información relevante cuya recopilación manual hubiese sido considerablemente más lenta e ineficiente (Wickham, 2016).

Los datos obtenidos a marzo de 2023 (Figura 3) muestran las publicaciones en la sección de artículos de los perfiles de personales de CONICET de los miembros del IDACOR, no incluyéndose aquellas producciones pertenecientes a libros y capítulos de libros que también forman parte de la información académica de cada autor. Es importante tener en cuenta que la sección de artículos del repositorio de CONICET recopila no sólo las publicaciones más "tradicionales" en revistas académicas, sino también otros tipos de escritos que forman parte del contenido de una revista como obituarios, traducciones, prólogos o introducciones. En este estudio, se tomaron en cuenta todos los datos obtenidos en la sección "artículos" sin realizar una discriminación entre ellos. Es importante tener en cuenta que esta metodología puede llevar a un sesgo en la cuantificación de la producción científica de un autor, pero se ha aplicado para ser fiel a los datos obtenidos en el repositorio.

Debido a la pluralidad de enfoques y la interdisciplinariedad dinámica de todos sus miembros, resulta complicado derribar barreras entre "disciplinas" (en términos de Wallerstein, 1999) por ello, se optó por dividir sus producciones según las áreas de conocimiento o comúnmente llamadas "comisiones" que componen al CONICET[7], lo cual dio a conocer que, al menos siguiendo este esquema de división de las ciencias arbitrario, el IDACOR reúne a diferentes científicos de siete áreas del conocimiento (Figura 3) que no solamente tienen vinculación con la antropología, sino también que hay una amplia amalgama de profesionales de otras áreas del conocimiento y que inclusive abarca disciplinas que no se encuentran encasilladas dentro de las Ciencias Sociales y Humanas cuya nomenclatura se encuentra representada por la asociación de las letras "KS" en adición a un carácter numérico. Este espacio se encuentra conformado por profesionales de las comisiones de Ambiente y Sustentabilidad (KA7), Biología (KB2), Ciencias de la Tierra, del Agua y de la Atmósfera (KE1), Historia

[7] Más información al respecto disponible en https://www.conicet.gov.ar/conicet-descripcion/

DOI: http://doi.org/10.31048/1852.4826.v16.n2.41094

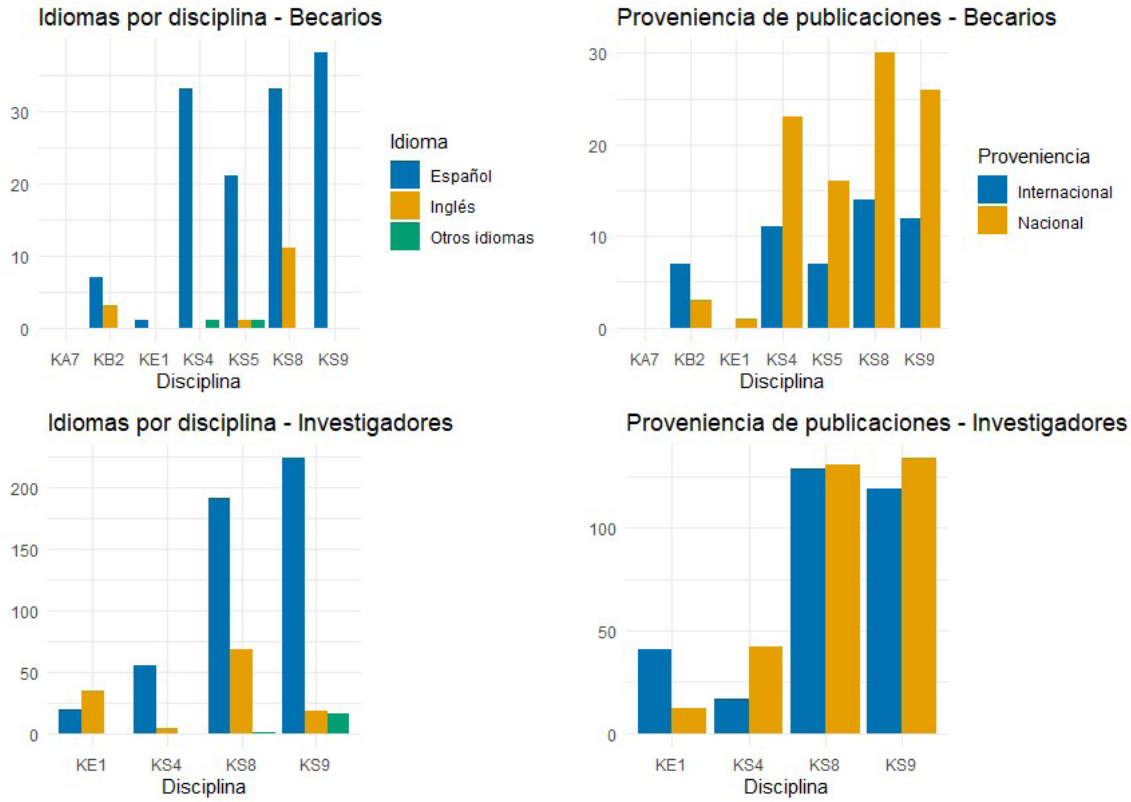

Figura 4. Representación del total de publicaciones por disciplina académica entre becarios e investigadores por idiomas y proveniencia. A) Idiomas de publicaciones en becarios, B) idiomas de publicaciones en investigadores, C) proveniencia de las publicaciones de becarios y D) proveniencia de las publicaciones de investigadores.

Figure 4. *Representation of the total number of publications by academic discipline among fellows and researchers by language and origin. A) Languages of publications by fellows, B) languages of publications by researchers, C) provenance of publications by fellows and D) provenance of publications by researchers.*

y Geografía (KS4), Sociología, Comunicación Social y Demografía (KS5), Arqueología y Antropología Biológica (KS8) y Ciencias Antropológicas (KS9).

Si bien el propósito de este trabajo no consiste en discutir aspectos vinculados al contenido de las publicaciones de sus miembros y de sus respectivas comisiones, a través de parámetros como la "calidad científica" y la "calidad editorial" (Rozemblum et al., 2015; Martinovich et. Al., 2014), ni analizar las estrategias que éstos adoptan para lograr un lugar en el mercado académico (Royero, 2007), es interesante exponer en términos cuantitativos el volumen de publicaciones de aquellos pertenecientes a la comisión KS8 en comparación con sus otros colegas para aproximarnos preliminarmente a los hábitos de publicación (en términos de Spengler y Kligmann, 2022) de los investigadores y becarios.

En la figura 4, podemos comparar la producción de nuestra área de conocimiento con otras, a fin de dar cuenta de las similitudes y diferencias que existen entre las formas de publicar y los canales que utilizan los arqueólogos con respecto a sus pares de otras áreas y haciendo una discriminación entre becarios e investigadores tomando como parámetros el idioma elegido y el origen de la revista al que pertenece la publicación.

En función de las formas en las que se publican las investigaciones en todas las ciencias; en líneas generales se encuentran en relación estrecha con los aspectos propios de cada comunidad científica como los canales, lenguas y círculos de validación o medición del impacto y evaluación científica, y, por el otro, existe un aparato de control, bases de datos, rankings y plataformas en línea que le dan visibilidad a la publicación científica (Isasi Belasco y del Rio Riande, 2022; Beigel, 2012) lo cual se ha traducido en la configuración y estandarización de un modo particular a la hora de publicar, donde generalmente el inglés prima como *lingua franca* en desmedro de las lenguas de origen de los investigadores. En el caso de América Latina y particularmente en las Ciencias Sociales y Humanas el panorama cambia considerablemente dado que en estas disciplinas las publicaciones mayormente suelen ser en español o portugués y están orientadas a revistas locales, dándole más visibilidad a los portales regionales, donde la organización nacional y regional

DOI: http://doi.org/10.31048/1852.4826.v16.n2.41094

de cada una de las disciplinas que la componen es extremadamente dominante en la estructura de publicación y, en consecuencia, también en el alcance geográfico de la investigación y el conocimiento (Beigel, 2014; Kristiansen, 2012).

En el caso de aquellas publicaciones de los miembros agrupados en la comisión KS8 del IDACOR (Figura 4) es notablemente visible la primacía del español como la lengua más utilizada para publicar en las producciones de becarios e investigadores, sin embargo si comparamos los números con otras comisiones, especialmente aquellas pertenecientes a las Ciencias Sociales y Humanas, es evidente que existe un patrón que le da un uso más pormenorizado a la lengua inglesa (esto se acentúa aún más en las áreas del conocimiento que no pertenecen al orden de Ciencias Sociales y Humanas).

Ahora bien, si consideramos el origen de las revistas en las cuales sus miembros publican, vemos que los becarios de la comisión KS8 tienen una tendencia a optar por revistas nacionales y que, salvando a la comisión KB2, el patrón en bastante similar en resto de la disciplinas sociales y humanas donde la recurrencia de las revistas nacionales prácticamente dobla a las de origen internacional. Mientras que, en las producciones de los investigadores de la comisión KS8, la estadística es prácticamente igual: no hay una diferencia significativa en cuanto la opción por publicar en revistas internacionales y nacionales, y, en comparación con las otras comisiones, tiene un patrón similar a la de KS9, mientras que KE1 se destaca por una mayor elección de revistas internacionales.

Conclusión y comentarios finales

Las actividades realizadas para este trabajo tuvieron el objeto de vincular nuevas herramientas que pueden ser formar parte de la investigación arqueológica, si bien mencionamos que, en términos generales, los arqueólogos tendemos a indagar hacia atrás en el tiempo para entender los modos de vida de individuos y grupos pasados, a través de estos casos de aplicación pudimos obtener datos primarios de otro orden que permitieron aproximarnos a entender cómo metodologías pertenecientes a la ciencia de datos pueden involucrarse con la arqueología en un estudio relevante y, al mismo tiempo, abrir el abanico de posibilidades que nos ofrece la obtención de datos estructurados y no estructurados mediada por otros vehículos digitales que permiten simplificar y acelerar la extracción de información en diversos formatos, especialmente en entornos digitales con grandes volúmenes de datos que pueden resultar difíciles de localizar. Estas técnicas proporcionan un insumo adecuado al permitir la recopilación automatizada y estructurada de información, lo que agiliza la identificación y obtención de datos relevantes para la investigación.

La intención de este manuscrito fue la de presentar el uso de nuevas herramientas (al menos para la arqueología) aplicados a casos de aplicación específicos y exponer los resultados obtenidos, si bien esta información permite realizar un análisis más pormenorizado de nuestros datos, aquí solo nos hemos limitado a dar cuenta de sus potencialidades para algunos escenarios que consideramos relevantes, sin dejar de pensar que otros espacios que pueden ser abordados mediante un análisis que incorpore los recursos utilizados que implican a una pormenorización más exhaustiva de varios otros datos que fueron obtenidos a lo largo de estos procesos. En el caso específico aplicado a Twitter, los resultados obtenidos pueden ser aún más detallados y precisos. Por ejemplo, es posible analizar el volumen y la repercusión de los tweets más retuiteados, identificar las cuentas más activas en términos de publicaciones y determinar cuáles son las publicaciones que se vuelven más virales, así como sus menciones más populares, los cual nos permite se pueden tener una visión más acabada de las interacciones entre los usuarios y comprender mejor cómo se relacionan entre sí. En otro ámbito de aplicación, la información extraída mediante estas técnicas puede ser de gran utilidad para abordar estrategias científicas en el ámbito académico. Con ello es posible realizar un filtrado más preciso de los espacios editoriales donde los investigadores deciden publicar, teniendo en cuenta tanto el idioma como los rankings estadísticos que miden el impacto de una revista. Además, mediante el análisis de redes de citas o autorías puede profundizarse en las relaciones entre publicaciones y autores, lo que permite comprender mejor el panorama de investigación y la relevancia que tiene el idioma a la hora de elegir escribir y enviar un manuscrito.

Aquí solo hemos aplicado técnicas de *Web Scraping* y *Text Mining* en Twitter para obtener información acerca de cómo los usuarios de la plataforma interactúan con la arqueología mediante el uso de palabras vinculadas a nuestra ciencia, pensando que lo digital se encuentra mediado entre arqueólogos y no arqueólogos (Izeta y Cattáneo, 2018) como se mencionó *ut supra*. Con ello, fue posible ofrecer un paneo acotado acerca de cómo fue circulación de las opiniones de los usuarios a través de dos *hashtags* donde pudo constatarse una mayor preponderancia y masividad en el inglés que en español pero, en líneas generales, el léxico utilizado por las cuentas en la red social tenía cierta vinculación con palabras que son frecuentes en el vocabulario arqueológico, las cuales no están exentas de variaciones, pues -como se constató en el caso de los tuits en español- fue posible identificar que existe un flujo de publicaciones ligadas a las pseudociencias que expresan una visión anticientificista de lo que es la arqueología y que tienen la capacidad de llegar a otros usuarios.

Por otro lado, también se utilizó *Web Scraping* mediante

otra librería de R para un caso de aplicación distinto donde se recabaron datos acerca de las producciones científicas de los miembros del IDACOR, tanto en becarios como en investigadores, y, si bien se hizo hincapié en aquellos pertenecientes a la comisión KS8 de CONICET. Asimismo, se obtuvo información del resto del cuerpo de esta unidad ejecutora para comparar los modos, formas y preferencias que tienen sus integrantes para publicar en revistas científicas. Nuevamente, fue posible constatar que, a diferencia del lenguaje en redes sociales, el español es el canal idiomático más utilizado para difusión científica. Sin embargo, es notable destacar que, a diferencia de otras comisiones de las Ciencias Sociales y Humanas, existe en la comisión KS8 un uso del inglés mucho más acentuado que en las otras, al mismo tiempo que la opción por publicar en revistas nacionales e internacionales es una variable que igualmente utilizada en el caso de los investigadores, mientras que, entre los becarios, hay una tendencia más marcada en recurrir a revistas nacionales.

Con todo ello este aporte con su introducción, esperamos abrir el abanico metodológico que la comunidad arqueológica tiene a disposición e invitar a la reflexión de cómo puede construirse su práctica desde un espacio que aún no ha sido abordado por tanto en estudios regionales como globales. Esto nos ofrece muchas posibilidades para los interesados en el campo de la arqueología, por supuesto, todas las plataformas y aplicaciones de redes sociales presentarán desafíos técnicos, sociales y culturales para el investigador, que diferirán según la capacidad de acceder a los datos de la plataforma, el tipo de actividad o fenómeno social bajo consideración y el aparato de interacción social provisto (Richardson, 2019: 153). Si bien los desafíos que enfrenta la comunidad académica que desea acceder a los datos invita a que deben involucrarse con habilidades tecnológicas y herramientas con las que recopilar datos, si bien son complejas cada vez son más accesibles e inclusive más dóciles a pesar de su mencionada complejidad (Richardson, 2019).

Agradecimientos

Este trabajo se encuadra dentro del "Proyecto integral de investigación, preservación y transferencia del patrimonio, Instituto de Antropología de Córdoba, UNC-CONICET. Proyecto PUE 22920160100024CO" y del proyecto "ARIADNEplus financiado por la Comisión Europea (H2020 Programme, contract no. H2020-INFRAIA-2018-1-823914)". Quiero hacer llegar mi gratitud a quienes construyen día a día un espacio más prolífico para la arqueología digital en todas sus latitudes, a los colegas que de forma desinteresada compartieron información no publicada. A la gran comunidad de R y otros lenguajes que respondió gran parte de mis inconvenientes. Cualquier error u omisión es responsabilidad del autor.

Bibliografía

Ali, R. H., Kashefi, A. K., Gorman, A. C., Walsh, J. St. P., y Linstead, E. J. (2022). Automated identification of astronauts on board the International Space Station: A case study in space archaeology. *Acta Astronautica*, 200, 262-269. https://doi.org/10.1016/j.actaastro.2022.08.017

Allés Torrent, S., del Rio Riande, G., De León, R., Fila, M., Hernández, N., Bonnell, J., y Song, D. (2020). Narrativas digitales de la COVID-19 en Twitter: de los datos a la interpretación. *Publicaciones de la Asociación Argentina de Humanidades Digitales*, 1. https://doi.org/10.24215/27187470e002

Arcila-Calderón C., Barbosa-Caro E. y Cabezuelo-Lorenzo F. (2016): Técnicas Big Data: análisis de textos a gran escala para la investigación científica y periodística. *El profesional de la información* 25 (4), 623-631.

Ávido, D., y Vitores, M. (2018). Lectura distante y visualización de textos en arqueología y disciplinas afines. *Trabajo presentado en el III Congreso Internacional de la Asociación de Humanidades Digitales (AAHD)*. https://n2t.net/ark:/13683/pzBp/DDe

Beigel F. (2012). David y Goliath. El sistema académico mundial y las perspectivas del conocimiento producido en la periferia. *Pensamiento Universitario* 15.

Beigel F. (2014). Publishing from the Periphery: Structural Heterogeneity and Segmented Circuits. The Evaluation of Scientific Publications for Tenure in Argentina's CONICET. *Current Sociology*, 62 (5), 743-765. https://doi.org/10.1177/0011392114533977

Bordignon, F. y Maisonobe, M. (2022). Researchers and their data: A study based on the use of the word *data* in scholarly articles. *Quantitative Science Studies*, 3(4), 1156-1178. https://doi.org/10.1162/qss_a_00220

Calvo E. y Aruguete N. (2020). *Fake news, trolls y otros encantos: Como funcionan (para bien y para mal) las redes sociales*. Siglo XXI Editores, Buenos Aires.

Daly P. y Evans T.L. (2006). Introduction: archaeological theory and digital pasts. En: T.L. Evans y Daly P (Eds.), *Digital Archaeology: bridging method and theory* (3-7). Abingdon: Routeledge.

Demïr, N., Boyoğlu, C. S., y Kayikci, D. (2023). A web scrapping and AI approach for archeologists to analyze the ancient cities. *Cultural Heritage and Science*, 4(1), 1-8. https://doi.org/10.58598/cuhes.1213426

Feldman R. y Dagan I. (1995). Knowledge Discovery in Textual Databases (KDT). *KDD* 95, 112-117.

Feldman R. y Sanger J. (2006). *The Text Mining Handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Graham, S., Huffer, D., y Blackadar, J. (2020). Towards a Digital Sensorial Archaeology as an Experiment in Distant Viewing of the Trade in Human Remains on Instagram. *Heritage*, 3(2), 208-227. https://doi.org/10.3390/heritage3020013

Grzegorczyk, M., y Salerno, V. (2022). Un análisis a través de las redes sociales y noticias periodísticas sobre el detectorismo de metales en Argentina. *Revista de Arqueología Histórica Argentina y Latinoamericana*, 16(1). http://www.doi.org/10.55695/rdahayl16.01.01

Hernández A., Gómez Vásquez E., Berdejo Rincón C., Montero Gacía J., Calderón Maldonado A. e Ibarra Orozco R. (2015). Metodologías para análisis político utilizando web scraping. *Resarch in Computing Science*. 95, 113-121.

Hernando A. (2002). *Arqueología de la identidad*. Akal: Madrid.

Isasi Velasco J. y del Rio Riande G. (2022). ¿En qué lengua citamos cuando escribimos sobre Humanidades Digitales?. *Revista de Humanidades Digitales* 7, 127-143. https://doi.org/10.5944/rhd.vol.7.2022

Izeta A.D. y Cattáneo R. (2018). ¿Es posible una arqueología digital en Argentina? Un acercamiento desde la práctica. *Humanidades Digitales: Construcciones locales en contextos globales*. Asociación Argentina de Humanidades Digitales: Buenos Aires. https://n2t.net/ark:/13683/ey3x/gwo

Kearney M. W. (2019). rtweet: Collecting and analyzing Twitter data. *Journal of Open Source Software*, 4(42). 1829. doi:10.21105/joss.01829

Kristiansen K. (2012). Archaeological communities and languages. En R. Skeates, C. McDavid y J. Carman (Eds.), *The Oxford Handbook of Public Archaeology* (461-467). Oxford University Press.

Laitano G. y Nieto A. (2022). La conflictividad social en los barrios de Mar del Plata (2016-2020): un acercamiento computacional. En: G. Laitano y A. Nieto (Eds.), *La conflictividad social en la historia reciente. Enfoques cuantitativos desde lo local a lo regional* (153-228). Buenos Aires.

Martínez R., Rodríguez R., Vera P. y Parkinson C. (2019). Análisis de técnicas de raspado de datos en la web – Aplicado al portal del estado nacional argentino. *XXV Congreso Argentino de Ciencias de la Computación* (457-466). Río Cuarto.

Martinovich V., Arakaki J. y Spinelli H. (2014). Diez años de Salud Colectiva: una aproximación a las reglas del juego del campo editorial científico. *Salud Colectiva* 10 (1). https://doi.org/10.18294/sc.2014.205

Marwick B., Boettiger C. y Mullen, L. (2018). Packaging data analytical work reproducibly using R (and friends). *The American Statistician*, 72(1), 80-88. https://doi.org/10.1080/00031305.2017.1375986

R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing: Vienna, Austria. https://www.R-project.org/

Richards, D. J., Tudhope, D., y Vlachidis, A. (2015). Text Mining in Archaeology: Extracting Information from Archaeological Reports. En J. A. Barcelo y I. Bogdanovic (Eds.), *Mathematics and Archaeology* (pp. 240-254). CRC Press. https://doi.org/10.1201/b18530-17

Richards J.D. (2009). From anarchy to good practice: the evolution of standards in archaeological computing. *Archeologia e Calcolatori*, 20, 27-35.

Richardson L. (2019). Using social media as a source for understanding public perceptions of archaeology: research challenges and methodological pitfalls. *Journal of Computer Applications in Archaeology*, 2(1), 151-162. https://doi.org/10.5334/jcaa.39

Richardson L. (2013). A Digital Public Archaeology? *Papers from the Institute of Archaeology*, 23(1), 10, 1-12. http://doi.org/10.5334/pia.431

Royero J.M. (2007). Las redes de investigación y desarrollo (I+D) en América Latina. *Revista de Universidad y Sociedad del Conocimiento* 3 (2). http://dx.doi.org/10.7238/rusc.v3i2.280

Rozemblun C., Unzurrungaza C., Banzato G. y Pucacco

C. (2015). Calidad editorial y calidad científica en los parámetros para inclusión de revistas científicas en bases de datos en Acceso Abierto y comerciales. *Palabra Clave* 4 (2).

Schadla-Hall T. (2004). The comforts of unreason: the importance and relevance of alternative archaeology. En: N. Merriman (Ed.), *Public archaeology* (269-285). Routledge.

Spengler, G. A., & Kligmann, D. M. (2022). Primeras aproximaciones al estudio de los hábitos de publicación de los arqueólogos argentinos. *Revista Iberoamericana de Ciencia, Tecnología y Sociedad*, 17(49), 91-125. http://ojs.revistacts.net/index.php/CTS/article/view/263

Twitter Blue. [@Twitter Blue] (8 de febrero de 2023). *need more than 280 characters to express yourself?.* [Tweet]. Twitter. https://twitter.com/TwitterBlue/status/1623411400545632256

Van Dijck J. (2016). *La cultura de la conectividad: Una historia crítica de las redes sociales*. Siglo XXI Editores: Buenos Aires.

Wallerstein I. (1999). *Impensar las Ciencias Sociales. Límites de los paradigmas decimonónicos*. Siglo XXI Editores: México.

Wickham H. (2016). *Package rvest*. https://cran.r-project.org/web/packages/rvest/rvest.pdf