En Del Rio Riande, Gimena, *La cultura de los datos: Actas del II Congreso Internacional de la AAHD*. La Plata y Rosario (Argentina): UNLP-FAHCE y UNR.

Lectura distante y visualización de textos en Arqueología. Ensayo preliminar.

Avido, Daniela Noemi y Vitores, Marcelo.

Cita:

Avido, Daniela Noemi y Vitores, Marcelo (2019). *Lectura distante y visualización de textos en Arqueología. Ensayo preliminar. En Del Rio Riande, Gimena La cultura de los datos: Actas del II Congreso Internacional de la AAHD. La Plata y Rosario (Argentina): UNLP-FAHCE y UNR.*

Dirección estable: https://www.aacademica.org/danavido/29

ARK: https://n2t.net/ark:/13683/pzBp/zUm

Lectura distante y visualización de textos en Arqueología. Ensayo preliminar

Daniela N. Ávido[1] **y** *Marcelo Vitores*[2]

Resumen

Este trabajo explora el potencial de las herramientas de análisis y visualización textual digital, utilizando un limitado corpus de trabajos académicos de arqueología en publicaciones periódicas argentinas. Inicialmente, el objetivo fue explorar los textos con herramientas de la lectura distante para observar el grado de similitud entre autores y, en lo posible, distinguir cronológicamente estilos de escritura. Más tarde, el interés giró hacia el relevamiento de las características de las digitalizaciones existentes, para evaluar si cumplían con los requisitos necesarios para que los textos pudieran ser procesados automáticamente. Para el análisis se utilizó el servidor web de *Voyant Tools* y el acervo disponible en bibliotecas digitales abiertas.

Introducción

La Arqueología puede definirse muy resumidamente como una forma de conocer el pasado humano a través de los restos materiales. Si bien esto determina que en los estudios arqueológicos exista un predominio de los objetos sobre los textos (por la ubicuidad de los primeros y la extrema restricción espacio-temporal de los segundos), el uso de fuentes escritas también se constituye recurrentemente como una línea de evidencia comparativa o como

[1] Universidad de Buenos Aires. danavido@gmail.com

[2] Universidad Nacional de Luján, ProArHEP. Universidad de Buenos Aires. marcelovitores@yahoo.com.ar

fuente para generar analogías y modelos con los que interpretar el registro arqueológico. Por otra parte, son cada vez más frecuentes las publicaciones cuyo material de estudio lo constituyen exclusivamente antecedentes bibliográficos, sobre todo al encararse la historiografía de la propia disciplina. En todos los casos, parte del enfoque puede y suele incluir alguna forma de aproximación bibliométrica, aunque sólo sea un rudimentario conteo.

En su uso como línea de evidencia comparativa, el tratamiento de los textos no es radicalmente distinto del que se haría, por ejemplo, en una investigación histórica, excepto en los objetivos y contenidos enfocados. Existe una copiosa producción sobre la relación entre estas fuentes de datos y su entrecruzamiento, parte de la cual recae en el campo de la Arqueología histórica y de la Etnohistoria aplicada a la Arqueología (Johnson, 2000). En algunos casos lo que se cuantifican son menciones o ausencias de elementos en los textos, a fin de representar un fenómeno a lo largo del tiempo y el espacio (Moreno e Izeta, 1999; Vitores, 2015).

El estudio bibliográfico de antecedentes, por otra parte, ocurre en cualquier rama del conocimiento, constituyendo el caso que concierne a este trabajo. En la Arqueología argentina muchos trabajos han encarado el estudio de tendencias siguiendo una selección de documentos o publicaciones periódicas a lo largo del tiempo, en temas disimiles como: la historia de la disciplina en un período particular (Bonnin y Laguens, 1984-1985), el devenir de una subdisciplina como la Arqueometría (Vidal, 2009), la influencia de los investigadores de una perspectiva teórica según los entramados de citas (Scheinsohn, 2009), el seguimiento de conceptos como el de patrimonio (Pupio y Salerno, 2014), la orientación de las investigaciones de una institución académica (Kligmann y Ramundo, 2014, Kligmann et al., 2016) o de las publicaciones en una revista científica particular (Kligmann y Spengler, 2016), la composición y características de las revistas que incluyen temáticas arqueológicas (Spengler y Kligmann, 2017) o el peso de variables y enfoques como los ambientales (Grana y Fernández, 2018). Tanto por su extensión y calidad como por su temprana aparición, cabe destacar entre todos el trabajo de Jorge Fernández al delinear su historia de la arqueología argentina, la cual cimentó con una detallada bibliografía. Entre otros aspectos, su exposición del tema incorporaba formas de visualización de la bibliografía cuantificada (Fernández, 1982, p. 173). Hoy, sin embargo, resulta inabarcable una continuación de esta

labor con las mismas técnicas manuales, dada la proliferación de las investigaciones y publicaciones en la materia. Esto se observa en los otros trabajos mencionados que necesariamente acotan la tarea a un subtema, período o publicación particular.

Es en este punto donde nos preguntamos ¿qué aplicación puede tener la lectura distante en Arqueología?

Sucintamente, el concepto de lectura distante refiere al análisis informático de los textos, sin mediar una lectura humana directa[3]. Para comprender las posibilidades que esto provee, consideremos que, así como en la fotografía los objetivos angulares permiten una visión más amplia, aunque menos detallada de lo observado, la lectura distante permite abarcar un volumen mayor de textos, para analizarlos cuantitativamente yendo más allá del significado de sus contenidos gracias a la automatización de los procesos.

Dado que la lectura distante permite abarcar muchos textos de manera rápida, provee la posibilidad de realizar búsquedas de términos clave (sin limitarse al abstract y keywords), permite observar similitudes entre textos y llevar a cabo búsquedas orientadas a conceptos específicos (sobre todo si no son centrales a los artículos), por ello, consideramos que podría resultar una herramienta útil para análisis de antecedentes como los ejemplificados.

Para evaluar el potencial de la lectura distante en el análisis de textos de arqueología en diferentes etapas de una investigación, diseñamos un ensayo en tres etapas:

1. Relevamiento de revistas
2. Selección de artículos y confección de la muestra
3. Procesamiento

Relevamiento de revistas

El trabajo consistió en una serie de fases en las que se registraron distintos elementos, avanzando desde lo general hacia lo particular[4]. En primer lugar, se realizó un relevamiento de publicaciones periódicas nacionales que

[3] Para una definición, véase Lacalle y Vilar (2018).

[4] La base de datos de revistas, la tabla de índices y la tabla de artículos, así como el corpus y la lista de stopwords pueden descargarse en: https://zenodo.org/record/2566623.

incluyeran artículos o notas de arqueología. Las revistas relevadas podían ser vigentes o estar discontinuadas, ser digitales, digitalizadas o únicamente impresas. La tabla 1 detalla los atributos registrados para cada revista.

atributo	contenido
nombre	nombre de la revista
ISSN	nº de issn
editor	última institución o grupo editor
lugar	ciudad de publicación
inicio	año de inicio de publicación
continúa	continuidad de la publicación
fin	año de finalización de publicación. Si no hay más datos, puede que se indique como “fin” la fecha del último artículo de arqueología que sabemos que se publicó allí.
primero digital	año del primer volumen digitalizado
último digital	año del último volumen digitalizado
especificidad	el alcance de la revista en cuanto a campo disciplinar. El orden jerárquico es Arqueología>Cs. Antropológicas>Cs. Sociales>Ciencias
acceso	tipo de acceso a las publicaciones digitales y digitalizadas
url	página web o enlace de repositorios
navegacion	dificultad para navegar la página web
clicks resumen	cantidad de clicks para llegar al abstract –o título– de un artículo (desde la página inicial de la revista)
clicks texto	cantidad de clicks para llegar al contenido de un artículo (desde la página inicial de la revista)
n°art. total	cantidad de total artículos publicados en la revista
n°art. arqueo	cantidad de artículos de arqueología publicados en la revista
n°art. accedidos	artículos incluidos en la muestra analizada
imagen/texto/HTML	formato de almacenamiento del artículo
comentarios	otra información relevante
bibliografía	fuente desde donde se obtuvo la información

Tabla 1. Atributos registrados

El objetivo para esta primera etapa era hacer una evaluación general del estado de los materiales bibliográficos que se desearían usar, respondiendo cuestiones como: ¿tienen todos los números subidos?, ¿es fácil navegar

los contenidos?, ¿es digital o está digitalizado?, en el caso de escaneo ¿es suficientemente bueno para realizar el reconocimiento óptico de caracteres (OCR)? Otros aspectos fácilmente observables a partir de la esta primera base de datos son los geográficos, como la diversidad de ciudades desde donde se editan las revistas, con una marcada predominancia de Buenos Aires (figura 1). No obstante, cabe recordar que el presente ensayo es una prueba piloto y, por ende, los datos preliminares son sumamente incompletos. Mientras se avanza en la recopilación, el lector interesado en la caracterización de las publicaciones periódicas arqueológicas puede consultar otras obras (Spengler y Kligmann, 2017).

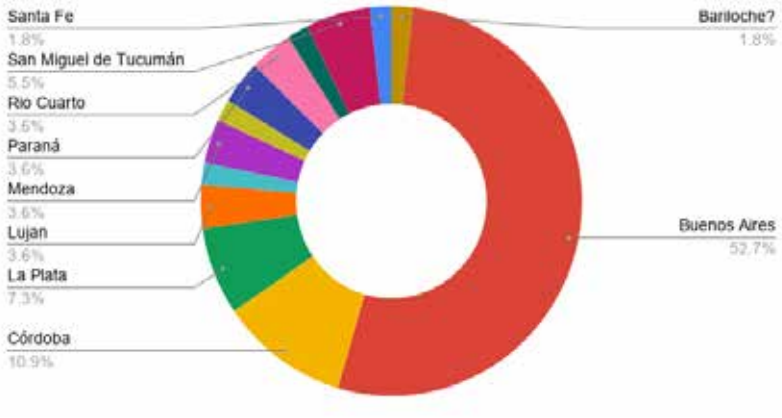

Figura 1. Frecuencia relativa de las ciudades de edición de las revistas

Selección de artículos y confección de la muestra

En la siguiente etapa, con el objetivo de aplicar lectura distante, seleccionamos una muestra de revistas electrónicas que tuvieran todos sus números digitalizados y en acceso abierto. Las revistas seleccionadas fueron: *Cuadernos del Instituto Nacional de Antropología y Pensamiento Latinoamericano-Series Especiales*, *Intersecciones en Antropología*, *Revista de Antropología del Museo de Entre Ríos*, y *Tefros*. El producto en esta segunda etapa de relevamiento fue la construcción de dos nuevas tablas[5]. La primera recopila

[5] Estas tablas también forman parte del dataset mencionado en la nota 3.

los enlaces a cada uno de los índices de los volúmenes (tocs) publicados en cada revista y la segunda registra cada uno de los artículos o notas de todos los volúmenes considerados en la primera.

La tabla que contenía los índices incluyó los siguientes campos: título de la revista, año, volumen/número, cantidad total de artículos y/o notas publicados, cantidad total de otros elementos publicados (editoriales, reseñas, comentarios), cantidad de artículos y/o notas de arqueología, enlace al índice y, finalmente, indicación de números temáticos. La tabla con los enlaces a los artículos incluidos en el corpus, cuyo criterio de selección fue la temática arqueológica, evitando también las notas editoriales, las reseñas y los artículos o notas de otros temas, contiene los siguientes campos: revista, volumen/número, título del artículo o nota, y enlace.

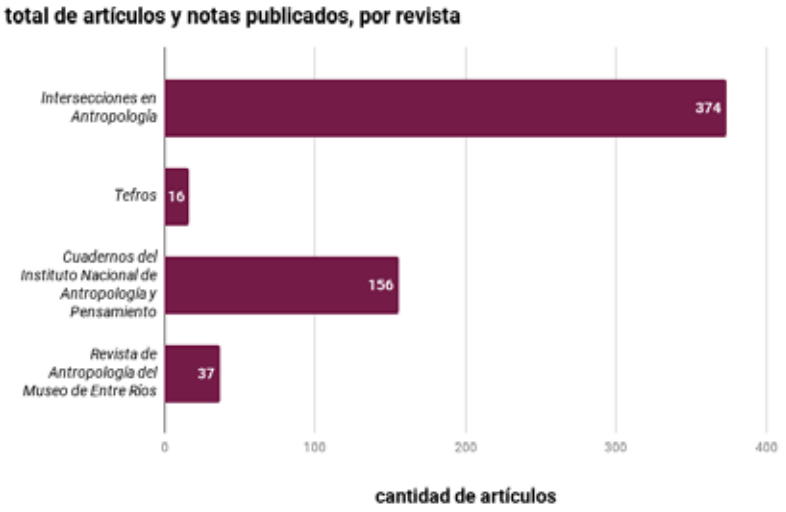

Figura 2. Artículos y notas de arqueología publicados en cada revista

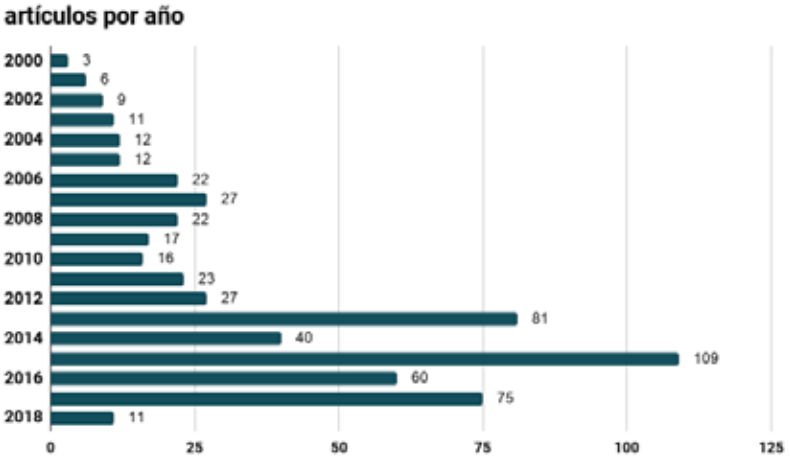

Figura 3. Cantidad de artículos y notas publicados por año, sin considerar la revista

Como se puede apreciar en las figuras 2 y 3, la revista *Intersecciones en Antropología* tiene un mayor peso en la muestra, esto se debe al rango temporal abarcado. Además, se evalúa que, independientemente de la trayectoria de cada revista, en los últimos cinco años hubo un marcado crecimiento en la cantidad de artículos, lo cual podría deberse a una disminución en los costos de publicación ya que las revistas que conforman la muestra analizada se editan electrónicamente. El valor bajo para 2018 responde a la fecha en que se realizó el relevamiento, cuando todavía no habían sido publicados los números correspondientes a ese año.

Procesamiento

Los artículos y notas seleccionados para analizar en este trabajo se encontraban en diversos formatos, algunos disponían el texto completo en HTML mientras otros requerían descargar un PDF. Una de las diferencias que permite este último formato, es el agregado de elementos propios del maquetado (encabezados, pie de página, etc.). Se realizaron ensayos parciales con diferentes formatos, copiando los archivos en una ubicación física o proveyendo los enlaces de acceso a *Voyant Tools* (Sinclair y Rockwell, 2016). Para homogeneizar los elementos recopilados, se optó por descargar masivamente

los textos con extensión .pdf (la que habitualmente proveen casi todas las publicaciones). Por otro lado, para hacer factible el procesamiento de tal número de textos, se transformaron los archivos a texto simple (.txt). Atendiendo al objetivo de explorar tendencias temporales, se los renombró anteponiendo el año. Para estas subtareas también se apeló a diferentes aplicaciones de acceso libre (Anthony 2017; TGRMN Software, 2008-2016).

Resultados

Los archivos de texto del corpus fueron cargados en *Voyant tools*, que luego nos permitió realizar algunas consultas y visualizar el texto cuantitativamente. En la figura 4 se puede observar una vista general de los paneles de *Voyant tools*[6].

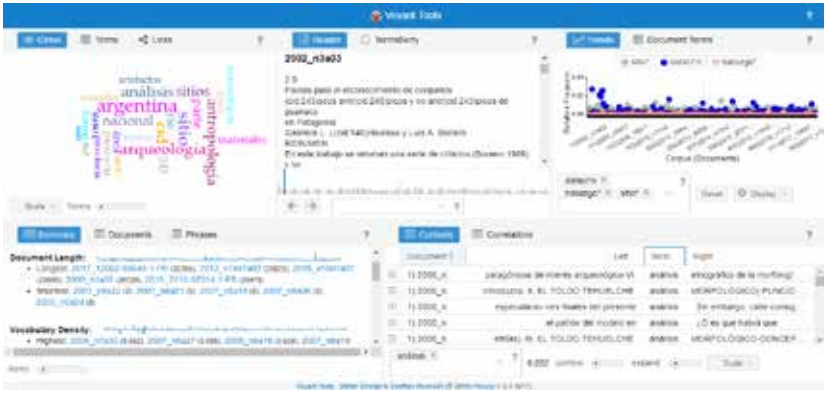

Figura 4. Captura de pantalla de *Voyant Tools*

La herramienta *Cirrus*, que conforma nubes de palabras con aquellas más frecuentes en el corpus, junto a la herramienta *Terms*, que muestra una tabla con la frecuencia de cada palabra, permitió observar las diez más frecuentes en la primera versión: sitio (n=9127), argentina (n=8753), cid (n=7046), análisis (n=6652), sitios (n=6167), antropología (n=6113), parte (n=5139), nacional (n=5095), río (n=5042). Editando sucesivamente la lista de *stopwords* se eliminaron las palabras no informativas (figura 5).

[6] También se puede acceder online al corpus en Voyant Tools a través de: https://bit.ly/2J91t78.

Figura 5. Cirrus con las palabras más frecuentes, y ejemplo antes y después de editar la lista de *stopwords*

Otras herramientas de la aplicación nos permitieron abordar palabras específicas para observar su variación a lo largo de los textos y del corpus, como *Trends* (tendencias) y *Links* (enlaces). En la figura 6 se ilustran dos ejemplos de la herramienta *Trends*, donde se puede ver la variación en la frecuencia de las palabras *isótopos* y *comunidad*.

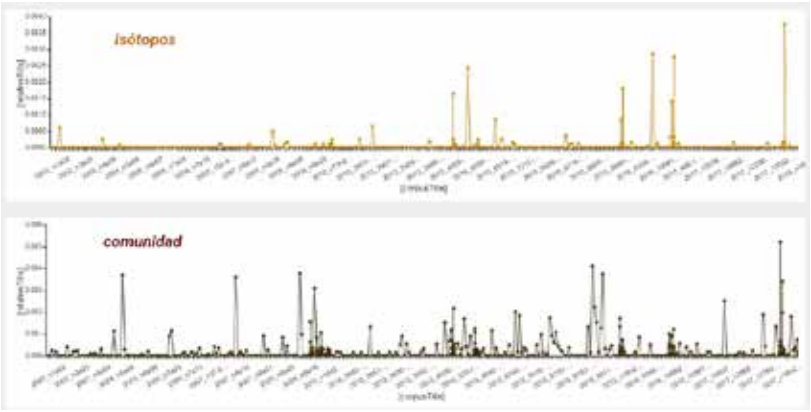

Figura 6. Herramienta *Trends*

En la figura 7 puede verse el uso de la herramienta *Links*, que muestra cuánto se relacionan entre sí distintos términos, a partir de palabras específicas como *patagonia*, *pampa*, *noroeste* y *nordeste*.

Figura 7. Herramienta *Links*

Los ensayos con la muestra piloto nos permitieron obtener algunas observaciones. En primer lugar, la necesidad de estandarizar y desarrollar las listas de *stopwords*[7] según los objetivos del análisis (rastrear un tema, subtema) y la naturaleza del corpus (ej.: artículos académicos, fuentes históricas, registros documental o arqueológico, etc.). La revisión de los términos más frecuentes evidenció el género de los textos; en este caso, la publicación académica. Bibliografías, epígrafes, encabezados y otros elementos textuales son compartidos por toda la muestra y no guardan interés analítico, por lo que debimos adicionar sus vocablos más recurrentes a la lista de *stopwords* genéricas para el castellano. Posteriormente, adicionamos los términos más genéricos y menos informativos que refieren a la disciplina (el primer ejemplo podría ser la misma palabra *arqueología*).

Estandarizar y compartir las listas de *stopwords* habilita la posibilidad de comparación entre los trabajos de diferentes investigadores. Asimismo, es deseable que dichas listas funcionen de forma aditiva, permitiendo conjugar varias, según los intereses del análisis. Por ejemplo, en nuestro caso encontraríamos útil combinar una lista genérica para castellano y para inglés, con una para publicaciones académicas y otra más para la arqueología, de modo de poder enfocar cuestiones

[7] Stopwords refiere a una lista de palabras que se evita incluir en el análisis del corpus. Accesible desde: https://voyant-tools.org/docs/#!/guide/stopwords.

al interior de la disciplina. Pero, hipotéticamente, una investigación más amplia sobre las ciencias sociales podría obviar las *stopwords* de arqueología, con el fin de comparar las referencias a esta y otras disciplinas, y así sucesivamente.

El sencillo procedimiento de revisar los términos frecuentes expuso el grado en que es necesario realizar una limpieza de los textos. La transformación de los archivos de un tipo de contenedor a otro (por ejemplo, de extensión .pdf a .txt) produjo algunas aberraciones como ser caracteres mal codificados, sobre todo vocales con acento gráfico que fueron cambiadas por una cadena de código (por ejemplo, "cid:243" = ó; "cid:146" = '; "cid:237" = í; etc.), rompiendo la palabra que los incorporaba. Esto lleva a la subrepresentación de los términos afectados, por lo que es necesario limpiar los textos para obtener resultados confiables. Este tipo de tarea es de trabajo intensivo, por lo que deberán sopesarse las alternativas para lograrlo con eficacia.

Consideraciones finales

En el presente trabajo buscamos explorar la aplicación de herramientas con una curva de aprendizaje rápida, que potencialmente nos facilite a los arqueólogos ampliar la capacidad de análisis al historiar la disciplina. Encontramos en la lectura distante, mediante el software *Voyant Tools*, una vía sencilla para profundizar el camino que muchas investigaciones han tomado al implementar alguna forma de medición bibliométrica.

En esta primera instancia el objetivo fue precisar los pasos de la implementación, por lo que se utilizó una muestra piloto a fines de ensayar el manejo del software y los procedimientos. Por su escasa representatividad y profundidad temporal no se pudieron apreciar cambios o tendencias que nos habiliten a una discusión de los contenidos, pero se lograron diversas observaciones para proseguir un trabajo que será más extenso y que sin duda requerirá desarrollarse con una metodología colaborativa.

En tanto que la lectura humana directa propone muchas veces la interpretación de conceptos y focos de atención explícitos, las tendencias de frecuencia y asociación mediante la lectura distante quizás puedan remitirnos a la recurrencia discursiva de los términos, independientemente del objetivo de cada texto. Su devenir a lo largo del tiempo puede ayudar a percibir cómo se tornan más familiares (y tácitos) ciertos conceptos o recortes temáticos. Por lo tanto, los resultados que se obtengan en próximas etapas serán de interés

para comparar con trabajos bibliográficos como los citados, en especial por las diferencias que el método propone. Al fin y al cabo, de lo que se trata es de "hacer comparaciones significativas" (Froehlich, 2015, p. 13).

Agradecimientos

A Gimena del Rio Riande por los talleres organizados junto a sus colegas del Laboratorio de Humanidades Digitales de CAICYT-CONICET, gracias a los cuales conocimos la existencia de estas herramientas. A la Universidad Nacional de Luján por brindar las condiciones para la realización del trabajo. A todas las instituciones que han decidido (y sostienen) publicar bajo licencias de acceso abierto. A los desarrolladores de softwares libres y amigables para usuarios inexpertos.

Referencias bibliográficas

Anthony, L. (2017). *AntFileConverter (Versión 1.2.1)* [Software de escritorio]. Tokio: Waseda University. Recuperado de https://bit.ly/2JbFyMA el 02/08/2018.

Anthony, L. (2018). *AntConc (Versión 3.5.7)* [Software de escritorio]. Tokio: Waseda University. Recuperado de https://bit.ly/1MeMh0f el 02/08/2018.

Bonnin, M., y Laguens, A. G. (1984-1985). Acerca de la Arqueología argentina en los últimos 20 años a través de las citas bibliográficas en las revistas Relaciones y Anales de Arqueología y Etnología. *Relaciones de la Sociedad Argentina de Antropología, 16,* 7-25. Recuperado de https://bit.ly/2vLffDa el 02/08/2018.

Fernández, J. (1982). *Historia de la arqueología argentina.* Mendoza: Asociación Cuyana de Antropología.

Froehlich, H. (2015). Análisis de corpus con AntConc. *The Programming Historian*. Recuperado de https://bit.ly/2GUw2XS el 02/08/2018.

Grana, L., y Fernández, M. (2018). El enfoque ambiental en la Arqueología argentina: análisis sobre su desarrollo en la disciplina a través de los trabajos publicados en la revista Relaciones. *Relaciones de la Sociedad Argentina de Antropología*, *43*(2): 261-286. Recuperado de https://bit.ly/2T0Xs45 el 02/08/2018.

Johnson, M. (2000). *Teoría arqueológica. Una introducción.* Ariel Historia. Ariel: Barcelona.

Kligmann, D. M., y Ramundo, P. S. (2014). ¿Qué nos cuentan las actas de defensa de las tesis de licenciatura en Ciencias Antropológicas (orientación Arqueológica) de la Facultad de Filosofía y Letras de la Universidad de Buenos Aires? *Relaciones de la Sociedad Argentina de Antropología*, *39*(1), 245-276. Recuperado de https://bit.ly/2PQdcXO el 02/08/2018.

Kligmann, D. M., y Spengler, G. (2016). Análisis histórico de una publicación científica especializada: pasado, presente y futuro de la revista Arqueología a 25 años de su creación. *Arqueología*, *22*(1), 15-60. Recuperado de https://bit.ly/2JlZ7QF el 02/08/2018.

Kligmann, D. M., Spengler, G. y Starrópoli, L. (2016). La arqueología argentina en los últimos 35 años: Reconstrucción de su historia disciplinar a través de las tesis de licenciatura y de doctorado de la FFyL de la UBA y de los trabajos con referato publicados en la revista Arqueología (FFyL, UBA). En *Actas del XIX Congreso Nacional de Arqueología Argentina*. (pp. 2427-2433). San Miguel de Tucumán: Facultad de Ciencias Naturales de la Universidad Nacional de Tucumán.

Lacalle, J. M. y Vilar, M. (2018). Una lectura distante de la investigación actual en Letras en Argentina. En G. del Rio Riande, G. Calarco, G. Striker y R. De León (Ed.), *Humanidades Digitales: Construcciones locales en contextos globales. Actas del I Congreso Internacional de la Asociación Argentina de Humanidades Digitales*. Buenos Aires: Facultad de Filosofía y Letras Universidad de Buenos Aires. Recuperado de: https://www.aacademica.org/aahd.congreso/18 el 27/05/2018.

Moreno, J. E. e Izeta, A. D. (1999). Estacionalidad y subsistencia indígena en Patagonia central según los viajeros de los siglos XVI-XVII. En J. B. Belardi, P. M. Fernández, R. A. Goñi, A. G. Guráieb, A. G. y M. De Nigris (Ed.), *Soplando en el Viento. Actas III Jornadas de Arqueología de la Patagonia* (pp. 477-490). Neuquén: Instituto Nacional de Antropología y Pensamiento Latinoamericano-Universidad del Comahue.

Pupio, A. y Salerno, V. M. (2014). El concepto de patrimonio en el campo de la arqueología argentina. Análisis de los trabajos presentados en los congresos nacionales de arqueología (1970-2010). *Intersecciones en Antropología*, *15*(1), 115-129. Recuperado de http://ref.scielo.org/qkfcf7 el 27/05/2018.

Ramundo, P. S. (2010). La historia contemporánea de la arqueología argentina, analizada a través de sus congresos nacionales. En *Actas de las IV Jornadas de Historia de la Ciencia Argentina* (pp. 255-266). Buenos Aires: Grupo Argentino de Historia de la Ciencia.

Scheinsohn, V. G. (2009). Evolución en la periferia. El caso de la arqueología evolutiva en la Argentina. En G. López y M. Cardillo (Ed.), *Arqueología y evolución. Teoría, metodología y casos de estudio* (pp. 73-86). Buenos Aires: Sb editorial.

Sinclair, S y Rockwell, G. (2016). *Voyant Tools* [Software de aplicación online]. Recuperado de http://voyant-tools.org/ el 29/06/2018.

Spengler, G. y Kligmann, D. M. (2017). Historia de la arqueología argentina a través del análisis de las revistas científicas nacionales. En *Actas de las XVI Jornadas Interescuelas de la Departamento de Historia de la Facultad de Humanidades de la UNMdP*. Mar del Plata: UNMdP.

Tgrmn Software. (2008-2016). *Bulk Rename Utility* [Software de escritorio]. Recuperado de http://bulkrenameutility.co.uk el 02/08/2018.

Vidal, A. S. (2009). La arqueometría americana en la actualidad: un pequeño paso para el investigador, un gran salto para la disciplina. En *Arqueometría Latinoamericana: 2do Congreso Argentino y 1er Latinoamericano* (pp. 15-24). Buenos Aires: Comisión Nacional de Energía Atómica.

Vitores, M. (2015). De ollas y fuentes en la etnohistoria patagónica. *Runa 36*(1), 29-49. Recuperado de https://bit.ly/2VYxHmO el 02/08/2018.