Desafíos teórico-metodológicos en el análisis de una problemática emergente: El caso de la Big Data en las Ciencias Sociales

Lucas Federico Sánchez (FaHCE-UNLP)

icherudim@gmail.com

Introducción

El tema de investigación se inserta en el marco de una tesina de grado y una beca EVC-CIN. Dicho trabajo se propone estudiar los distintos tipos de estudio y análisis que existen sobre el fenómeno de la Big Data dentro del campo de las Ciencias Sociales. Luego de la pandemia del COVID 19, pudimos observar a gran escala como el uso de las tecnologías digitales aumentó por parte de distintos actores sociales como Estados-Nación, empresas, ONG's, centros de investigación, universidades, entre otros. Este aumento del uso de las tecnologías digitales trajo consigo el crecimiento del uso y producción de Big Data por la sociedad en general. De esta forma, la Big Data parece como una nueva materia prima, que muchos de estos actores sociales utilizan para lograr maximizar su capacidad de inserción y producción de bienes y servicios para la sociedad. Partiendo de esta base, el objetivo propuesto por esta investigación es sistematizar y analizar las diferentes interpretaciones que existen sobre este fenómeno dentro de las múltiples disciplinas que integran las Ciencias Sociales. Esto se hará a partir de la selección de un corpus bibliográfico a través de un abordaje mixto usando técnicas cuantitativas y cualitativas de análisis. Para llevar adelante la investigación, se optó por combinar técnicas cuantitativas y cualitativas para la selección de una bibliografía representativa de la problemática elegida para examinar. Esto nos trajo algunos desafíos: ¿de qué manera seleccionamos revistas? ¿con qué criterio evaluar su representatividad? y ¿de qué forma nos damos cuenta de que son representativas del campo?

En un principio intentamos relevar el universo de revistas que publican la mayor cantidad de artículos sobre "Ciencias Sociales". De esta forma se evidenció que no todas las editoriales tienen criterios de búsqueda de artículos parecidos, y en muchos casos al combinar o comparar criterios, la búsqueda se vuelve ininteligible. Esto permitió que durante el proceso de investigación aparezcan preguntas como ¿de qué forma producimos una muestra que entonces sea representativa del universo que estamos buscando? ¿Qué técnicas de selección usar que sean representativas de ese universo? La solución a dichas preguntas fue utilizar y buscar "Journals" que integren y publiquen sobre la temática y basarnos en el criterio de impacto de cada una para seleccionar las que traten sobre el tema. Luego de la selección de dichas revistas, el análisis cuantitativo trajo consigo otro desafío referido a las técnicas de análisis. El análisis de los datos recolectados sobre la bibliografía trajo interrogantes sobre la

utilización de diversas herramientas para analizarlos y sacar conclusiones de cuanta bibliografía disponible había, dónde, de qué forma, qué autores participan, palabras clave, entre otras. Para esto surgieron preguntas sobre ¿qué herramientas de análisis cuantitativo podrían permitirnos realizar una mejor lectura de la información recolectada? y otras sobre la utilización correcta de dichas herramientas como Atlas.ti, Hojas de Google, Excel, etc. El análisis cuantitativo de una temática emergente trae consigo diversos retos que abarcan desde la selección de editoriales/revistas hasta el uso de técnicas correspondientes para su análisis.

Como conclusiones de esta ponencia se expondrán los resultados obtenidos del análisis de datos cuantitativos, las limitaciones durante el proceso de recolección de datos, el carácter central del análisis de dicha temática y las características generales de un proceso de investigación sobre dicha temática.

Este trabajo se presenta como un avance dentro de lo que fue el desarrollo de una beca EVC-CIN y de una tesina de grado de la carrera Licenciatura en Sociología dictada por la Facultad de Humanidades y Ciencias de la Educación de la Universidad Nacional de La Plata.

Primera Etapa: Selección de revistas y servidores de acceso abierto

Durante el inicio del proceso de investigación, al comenzar a analizar la problemática seleccionada “Big Data” logramos destacar algo sumamente importante que marcaría los objetivos de todo el proceso. Cuando nos preguntamos bien de qué forma se problematizaría la Big Data, nos dimos cuenta que dentro de las Ciencias Sociales en general, no existe un criterio univoco. La Big Data se presenta dentro de estas ciencias como un criterio polisémico, es un concepto, una definición que no se encuentra homogeneizada por todos los campos de las Ciencias Sociales.

Dicha definición depende enteramente del campo de estudios al que hace referencia, podemos obtener algunas definiciones que hacen referencia a las capacidades revolucionarias que trae para lo tecnológico, otras a las posibilidades de generar modelos más ricos para el análisis de datos, otras a que son datos de dificil acceso, otras a lo económico, a la creación de distintos tipos de carreras de grado, entre otras muchas más. Este aspecto polisémico nos permite observar que la Big Data, no es una noción que se encuentre estabilizada y por lo tanto se puede caracterizar como una definición heterogénea que admite múltiples significados distintos dentro del campo de las Ciencias Sociales.

Al lograr identificar a través de unos primeros acercamientos a esta noción como heterogénea y polisémica nos realizamos la pregunta siguiente que fue la que organiza esta todo el proceso

de selección cuantitativa y cualitativa que es ¿Cómo constituyen desde las Ciencias Sociales a la Big Data?

Con esta pregunta, lo que buscábamos durante el proceso de investigación y construcción de la tesina de grado es realizar un estado del arte sobre cómo está constituido el campo de estudio sobre la Big Data dentro de las Ciencias Sociales. Cuando hablamos del estado del arte retomando a Vargas, Galeano y Muñoz (2015) podemos decir que es una metodología de investigación cualitativa que tiene un carácter fundamentalmente critico-interpretativo que revisa como está el estado actual de un conocimiento producido. El estado del arte se puede caracterizar porque pretende y busca dar cuenta del estado actual de un conocimiento sobre un determinado concepto.

Continuando con la pregunta que nos planteamos, para poder responder a la misma, en primer lugar, nos decidimos por seleccionar un corpus bibliográfico que permitiera e intente ser representativo del campo que estudia a la Big Data dentro de las Ciencias Sociales. Esta cuestión nos trajo algunas problemáticas diferentes ya que ¿Qué criterios debíamos tener para poder seleccionar revistas/editoriales que fueran representativas del universo buscado?

Para poder seleccionar dicha pregunta, nos propusimos elegir un numero de revistas o Journals provenientes de países centrales y algunos servidores de acceso abierto que sean representativos de nuestra región y para eso, seleccionamos las editoriales y servidores que integraban la mayor cantidad de textos publicados sobre Ciencias Sociales en general de ambas regiones. Luego de este primer paso, teniendo en cuenta la magnitud de la bibliografía que íbamos a recibir, decidimos poder hacer un recorte de las disciplinas que íbamos a tomar y nos decantamos por 6, sociología, geografía, historia, estudios culturales, economía y humanidades.

En base a las disciplinas que seleccionamos, buscamos cuales eran las editoriales que integraban la mayor cantidad de publicaciones y teniendo ese dato en cuenta, elegimos de los países del primer mundo las editoriales Wiley, Springer, Elseiver y SAGE, y provenientes de nuestra región seleccionamos los servidores de acceso abierto de CLACSO, La Referencia y el SNRD (Sistema Nacional de Repositorios Digitales). Una vez seleccionadas estas revistas y servidores comenzaron a surgir limitaciones que obligaron a tomar ciertas decisiones que modificaron el acceso al corpus bibliográfico que fuimos construyendo.

En primer lugar, las editoriales de los países del primer mundo no manejaban criterios similares para poder elegir Journals o revistas que problematicen o estudien la temática, en muchos casos dentro de las Ciencias Sociales no aparecían las mismas disciplinas científicas y en otros sí, pero otras aparecían junto con otras temáticas de estudio. Esto generó que

tengamos que ajustar cuales eran las formas en las que íbamos a llegar a esas revistas o servidores de acceso abierto en nuestro caso.

Por lo tanto, para poder llegar a estas revistas representativas optamos por elegir las revistas que involucren la temática y integren la mayor cantidad de textos sobre la temática “Big Data”. Las revistas seleccionadas fueron “Big Data & Society”, “Journal of Big Data” y “Association for Information Science and Tecnology”, estas revistas nos permiten saber cuál es la cantidad exacta de artículos publicados de Ciencias Sociales sobre Big Data.

Segunda Etapa: Procesamiento de bibliografía

Luego de la selección de las revistas y los servidores de acceso abierto nombrados, comenzamos la tarea de recolectar todo el material bibliográfico disponible dentro de estos sitios web y de datos que acompañaban dichos artículos. Este material bibliográfico recolectado iba tener el rol de ser la materia prima que luego funcione como base para la construcción de un corpus representativo de la problemática que buscamos analizar.

Para realizar la recolección de los artículos necesarios utilizamos la herramienta “Zotero”, dicha herramienta digital, lo que permite es poder guardar en nuestra computadora todos los archivos que seleccionemos en un link y los almacena en nuestra PC. Estos archivos que almacena se guardan junto con, los y las autores/as, el título, resumen, titulo corto, DOI, lugar y fecha de publicación, “tags” o palabras clave, entre otras cosas.

Esta recolección a través del programa Zotero permitió que se recolecten en total entre las 6 disciplinas seleccionadas unos 3167 artículos con –en la mayoría de los casos- gran parte de la información que acompañaba dichos artículos, para de esta forma, realizar una compilación que intentara ser lo más completa posible. De estos 3167 artículos podemos decir que 876 pertenecen a la disciplina “Economía”, 197 a “Sociología”, 757 pertenecen a “Humanidades”, 724 pertenecen a “Historia”, 187 a “Geografía” y 426 a “Estudios Culturales”. Teniendo esto en cuenta, si observamos los gráficos siguientes, podemos ver cómo podemos ver cómo era la división dentro de cada revista y servidor.

Artículos de la revista ASSIS&T

- Economía
- Historía
- Geografía
- Estudios Culturales
- Humanidades
- Sociología

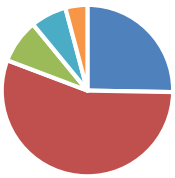
Artículos del servidor CLACSO

- Economía
- Historía
- Geografía
- Estudios Culturales
- Humanidades
- Sociología

Artículos de la revista Big Data & Society

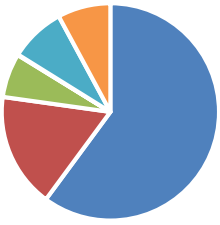
- Economía
- Historía
- Geografía
- Estudios Culturales
- Humanidades
- Sociología

Artículos de la revista Journal of Big Data



- Economía
- Historía
- Geografía
- Estudios Culturales
- Humanidades
- Sociología

Artículos del servidor La Referencia



- Economía
- Historía
- Geografía
- Estudios Culturales
- Humanidades
- Sociología

Artículos del SNRD

- Economía
- Historía
- Geografía
- Estudios Culturales
- Humanidades
- Sociología

Esta técnica de recolección con Zotero si bien en las revistas provenientes de la región de los países centrales funcionó recolectando todos los datos, en los servidores de acceso abierto de nuestra región no fue así. En CLACSO, el SNRD y La Referencia solo recolectó la PDF, el título, subtitulo y los autores, dejando por fuera toda la otra información que estaba subida en los sitios.

Un primer aspecto a destacar de esta recolección es que, los servidores de acceso abierto de nuestra región tienen una cantidad mucho menor de artículos publicados sobre Big Data, que los buscadores privados provenientes de la otra región. En total, los servidores de acceso abierto tienen de estos 3167 artículos, unos 1151 artículos en donde la mayoría provienen de La Referencia. Mientras que el resto de artículos provienen de los servidores privados y la mayoría de artículos son publicados en Big Data & Society.

Un segundo aspecto, no menor de toda esta recolección, es que, en la mayoría de casos observados, los artículos publicados no hacían hincapié en la disciplina a la que pertenecían sino, por el contrario, hacían referencian exclusivamente al área o temática de estudio a la que pertenecían, ya que muchos de los artículos tendían a repetirse 2, 3 o 4 en las disciplinas tenidas en cuenta. Esto nos permitía ver que había publicaciones que se involucraban en más de una disciplina y debido a esto, para poder constituir un corpus que sea más representativo de la muestra tomada nos decidimos por la construcción de áreas de estudio. En dichas áreas podíamos integrar material que si bien se caracterizara por tener un rol interdisciplinar se enfocaría en ejes teóricos que problematizaran el objeto desde una misma mirada teórica.

Tercera Sección: Análisis de palabras clave

En el proceso, se recolectaron en total unas 5317 palabras clave o tags de los artículos que provenían de las revistas de editoriales privadas. Sobre los servidores de acceso abierto intentamos realizar la misma tarea a través de esta herramienta y fue algo que no pudimos realizar, ya que como dijimos la herramienta no las recolectaba por un propio límite de los buscadores y, por lo tanto, esta recolección tiene que ser manual para esos buscadores.

Para analizar estas palabras clave, en primer lugar y con los criterios que veníamos utilizando de disciplinas –y para poder de alguna forma “afinar” la búsqueda- aplicamos el uso de la herramienta digital Atlas.ti. Con el uso de esta herramienta se intentaría ver cuáles eran los términos y tags que se utilizaban con mayor frecuencia dentro de esta problemática emergente. Por lo tanto, con el fin de detectar los términos, se optó por las herramientas disponibles del software Atlas.Ti (ver. 9) Lista de palabras y Nube de palabras.

Previo al análisis en Atlas.Ti, se construyó la matriz de datos y se ordenaron las etiquetas (tags) por medio de una planilla excel de modo que las filas representen las revistas y las columnas el número de etiquetas que varió entre un mínimo de 5 y un máximo de 11. Asimismo, para el procesamiento de los datos, se editó y refinó la lista de exclusión de palabras, así como también se tildó la opción de No distinguir mayúsculas y minúsculas.

En el caso de la Lista de palabras, se optó por mantener un umbral bajo (5 palabras) y un orden descendente de acuerdo al porcentaje de aparición del término en la matriz. En total, se contabilizan 31 términos relevantes que tienen una mediana de 7 apariciones en un rango que varía entre 5 y 87. En el caso de la Nube de palabras también se optó por un umbral bajo (6 palabras). A continuación, los resultados.

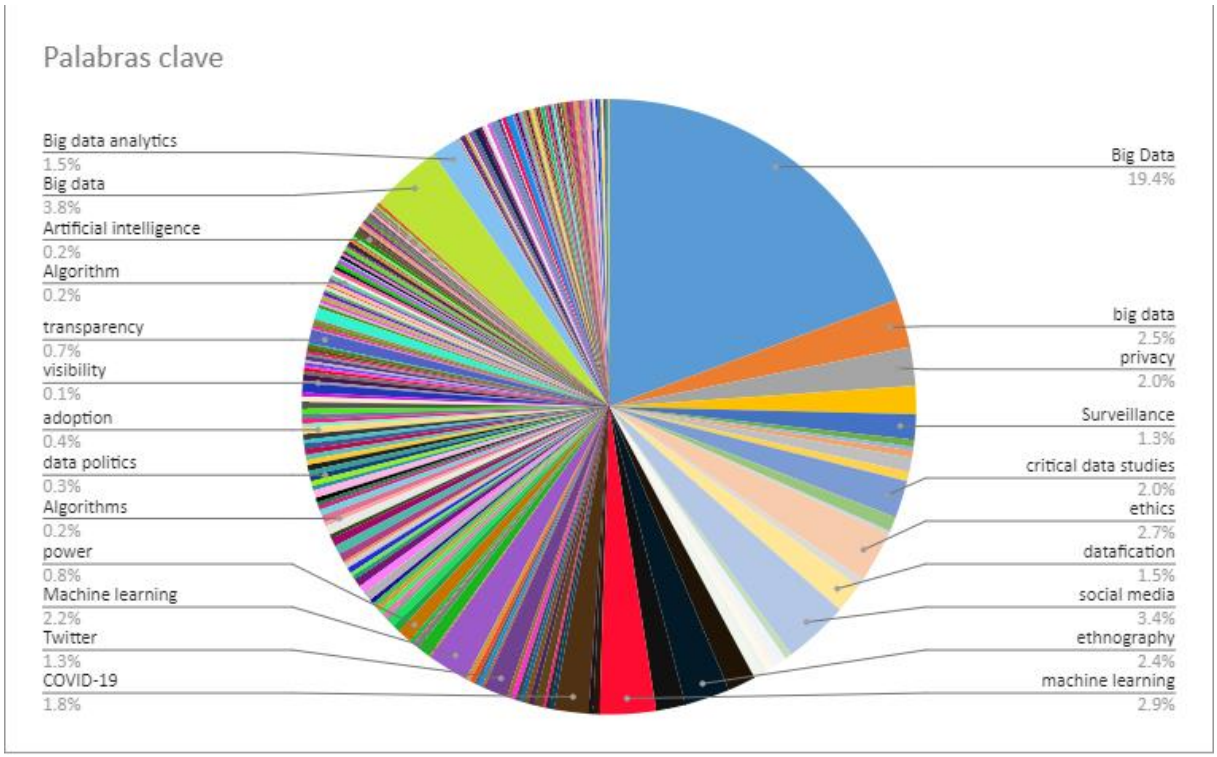
Como es posible apreciar, el principal obstáculo de esta técnica -así como el desconocimiento acerca de su posible solución- es el hecho de que el software es capaz de organizar términos (en nuestro caso etiquetas) de una sola palabra, pero no es capaz de presentarlos adecuadamente en caso de que tenga más de una palabra. Por ejemplo, el término Big Data no aparece completo, sino por separado como Big y Data o como otros términos que se caracterizan por tener más de una palabra.

Una posible vía de solución podría ser unir estos términos compuestos por más de una palabra en la planilla Excel previa. Sin embargo, hasta donde nuestro conocimiento en el software lo permite, se trataría de una tarea manual, la cual sería muy costoso de hacer tratándose de matrices de datos muy grandes.

Esta primera prueba con el programa Atlas.ti para poder ver como se comportaba con las palabras clave nos hizo dar cuenta que, para poder realizar bien una lectura de esto, se tendría que pasar todas las palabras a Excel o Hojas de Docs (herramienta de Google). Esto nos permitiría poder separar más que entre palabras separar entre conceptos, porque una de las ventajas que tiene la recolección con Zotero, es que permite exportar toda la cantidad de datos que extraemos de los artículos, en un archivo .csv y dentro de ese archivo, tenemos todos los datos de los artículos separados en columnas, cada columna con su propia definición.

Para poder completar el análisis de las palabras clave que pudimos obtuvimos, tuvimos que primero, exportar el .csv de Zotero, trasladar todas esas palabras a un Excel, luego ponerlas todas en una columna, esto es una tarea que se puede realizar manualmente. Y luego a través del uso de un gráfico dinámico simple, podemos observar cual es el comportamiento de las palabras clave que encontramos.

En el siguiente gráfico construido a través de Excel, podemos observar cómo queda compuesto el universo de palabras clave recolectadas de las revistas que se encuentran en los países del primer mundo:



Algo que se puede observar de primera mano observando el gráfico, es que el único concepto que tiene una relevancia importante para los trabajos relacionados al tema es el propio término “Big Data” y que, en los demás casos, los conceptos en s u mayoría no aparecen más del 0.5% del total de veces de las búsquedas realizadas. Luego, si apartamos la mirada de dicho

concepto, podemos observar que de los 5317 tags que obtuvimos, hay pocos que tengan un peso lo suficientemente fuerte como para catalogarse como relevante dentro de la problemática.

El comportamiento que se puede observar de las palabras clave obtenidas nos obligó a tomar una decisión pensando en ¿Cómo hacer para que tal magnitud de palabras clave tenga sentido o adquiera algún tipo de sentido? Ya que, de esta forma, no parezca que tengan algún sentido claro aparte de la aparición del término “Big Data”. Si nuestro objetivo con las palabras clave era encontrar algunas que aparezcan como relevantes para poder esclarecer futuras búsquedas de bibliografía sobre la temática, nos dimos cuenta que si anteriormente, la bibliografía respondía mayoritariamente a criterios que eran trasversales a algunas de las distintas disciplinas que estábamos usando y debíamos formar o tematizar la problemática en áreas, lo que debíamos hacer era también tematizar este universo de palabras clave.

Si bien, sostuvimos el criterio de recolección a través de las disciplinas, optamos por construir –como se dijo antes- temáticas de estudio para lograr agrupar los artículos dentro de determinadas áreas. Estas áreas a su vez están construidas también utilizando las palabras clave que recolectamos anteriormente a través de Zotero a su vez, estas palabras clave también funcionarían retroalimentando la búsqueda previa para saber que palabras clave se utilizan en caso de querer realizar búsquedas dentro de determinadas áreas o temáticas de estudio.

En parte, esto aparece como un ajuste dentro del propio trabajo, ya que a través de esta recolección fue como nos dimos cuenta que las disciplinas, como es que pensamos en un primer momento, no construyen en sí mismas divisiones internas dentro de esta problemática, como por otro lado, si lo hacen las áreas temáticas, algo que se logra evidenciar con las palabras clave o tags.

Por lo tanto, para proceder con el análisis y la tematización de dichas palabras clave la respuesta que pudimos encontrar y que aún se encuentra en desarrollo es ir revisando todo el universo de palabras clave mientras se va realizando la lectura de algunos de los textos que fuimos recolectando cuando realizamos el primer paso de la recolección del corpus bibliográfico en su totalidad. La selección de dicha bibliografía va a depender de 3 ejes fundamentales, la cantidad de citas que tengan los textos seleccionados, los criterios de impacto que manejen y la cantidad de veces que vayamos encontrando citas referidas a esos textos o esos autores mientras se avanza con la lectura del corpus bibliográfico.

Conclusiones

A través de este proceso de investigación pudimos comprender cuales son las características fundamentales de nuestro objeto de investigación y podemos dar cuenta y enumerar algunas conclusiones de esto.

Una de las primeras conclusiones es que, durante este proceso, aparecen algunas limitaciones técnicas. Estas limitaciones que fueron apareciendo fueron los diferentes criterios que cada buscador tiene dentro de sí mismo para la bibliografía de la que dispone. Esto puede llevar a que tengan que crear u optar por formas de búsqueda bibliográfica para cada repositorio o editorial acorde a los criterios que construyen internamente para almacenar la bibliografía. Luego, una segunda limitación de esta búsqueda es que no todos los sitios web te permiten poder recolectar a través de herramientas como Zotero las palabras clave y todo lo que se une al archivo, por lo tanto, para realizar un análisis más profundo de dichas áreas es necesario realizar trabajos manuales para llegar a esa información que queremos, modificando así la forma de acercarnos a dicha información.

Si observamos los datos recolectados sobre la bibliografía, podemos observar que los estudios sobre Big Data, tienen en muchos casos unos límites muy difusos para su publicación, límites que en muchos casos no terminan dependiendo del tema que está estudiando, si es puntualmente la Big Data o la Big Data dentro de un área o disciplina particular de estudio. Estos límites difusos en muchos casos también terminan siendo perjudicados por la falta de criterios claros por parte de las editoriales para agrupar los artículos publicados dentro de sus sitios web. Esto unido a la idea de que, en muchos casos, este concepto no se encuentra estabilizado teóricamente y tiene múltiples variables refuerza la necesidad de poder generar áreas bien delimitadas de estudios, claras y que se pueda observar con claridad cuáles son los criterios que logran que dicho estudio integre un área o disciplina particular.

Una última conclusión posible de esto es que, el proceso de investigación se caracteriza fuertemente por tener un carácter flexible, que, en muchos casos, el proceso, los objetivos o las preguntas con las cuales problematizamos nuestro objeto de investigación tienen que ser modificadas para poder encontrar una respuesta coherente a lo que nos estamos preguntando. Retomando a Sautu (2005) podemos decir que todo proceso de investigación no se reduce a secuencias rígidas, sin importar el tipo de estrategia que nosotros podamos optar, las distintas etapas de un proceso de investigación se van armando de una forma circular, es decir, que cada parte y etapa del proceso de investigación vuelve sobre sí misma para poder reformularse e ir actuando sobre todo el procedimiento. Esto sobre lo que nos

habla Sautu, es algo que se puede observar durante –al menos en este caso- el proceso de investigación, donde luego de generar las preguntas para analizar la problemática, las mismas, debido a límites que excedían de alguna forma lo que podemos hacer lograron que tengamos que modificar la forma de acceder a esa información y que esto es algo propio de cada proceso de investigación, en donde, en muchos casos aparecen situaciones que exceden a los y las propios/as investigadores/as.

Situaciones como las nombradas durante este trabajo, no son situaciones que puedan ser sabidas de antemano, son cosas que se van manifestando de alguna forma durante el proceso de investigación y son propias también de él.

Bibliografía

Gómez Vargas, M., Galeano Higuita, C. y Jaramillo Muñoz, D. A. (julio-diciembre, 2015). El Estado del arte: una metodología de investigación.

Sautu, ***Ruth***. ***Todo es teoría***: objetivos y métodos de investigación. - la ed. - Buenos Aires: Lumiere,. 2005. 180p. 22x16cm. ISBN 950-9603-57-O.

Anexo

https://journals.sagepub.com/home/bds

https://journalofbigdata.springeropen.com/

https://www.asist.org/

https://www.clacso.org/

https://www.lareferencia.info/es/

https://repositoriosdigitales.mincyt.gob.ar/vufind/