TECHNO REVIEW | Vol. 13, No. 1, 2023 | ISSN 2695-9933
International Technology Science and Society Review / Revista Internacional de Tecnología Ciencia y Sociedad
https://doi.org/10.37467/revtechno.v12.3383

CIENCIAS SOCIALES Y *BIG DATA*
Plataformas y desafíos

Social Sciences and *Big Data*. Platforms and Challenges

RENZO ANTONIO SEMINARIO-CÓRDOVA
Universidad Cesar Vallejo, Perú

KEYWORDS

Big data
Humanities
Platforms
Research
Repositories
Social Sciences

ABSTRACT

The objective of this research was to explore and characterize the main big data repositories in the area of social sciences available in 2021. The research design was non-experimental, exploratory and descriptive. The population consisted of 110 big data located by the Google dataset search engine. The sample corresponded to the top 10 big data. The results indicated that the most important big data repositories and platforms are centralized by the private sector located in US companies, fundamentally.

PALABRAS CLAVE

Big data
Ciencias sociales
Humanidades
Investigación
Plataformas
Repositorios

RESUMEN

El objetivo de esta investigación fue explorar y caracterizar los principales repositorios de big data en el área de ciencias sociales disponibles en 2021. El diseño de la investigación fue no experimental, exploratoria y descriptiva. La población estuvo constituida por 110 big data localizados por el motor de búsqueda para conjuntos de datos (datasets) de Google. La muestra correspondió a los 10 principales big data. Los resultados indicaron que los repositorios y plataformas más importantes de big data se encuentran centralizados por el sector privado localizado en empresas de EE. UU., fundamentalmente.

Recibido: 11/ 08 / 2022
Aceptado: 19/ 10 / 2022

1. Introducción

La investigación en el área de ciencias sociales se encuentra transversalizada por la dinámica de las tecnologías de la comunicación e información. El conocimiento fluye a través de las redes sociales en una generación incesante de textos digitalizados que provienen de fuentes diversas, las mismas que comprenden prensa, individuos y organizaciones, generando millones de terabytes en datos no estructurados, contentivos de valiosa información en tiempo real. Tradicionalmente, el análisis de contenido, texto, narrativas y discurso era realizado por el investigador social de forma manual, lo que consumía ingente cantidad de recursos (Kobayashi *et al.*, 2017). En la actualidad, la minería de texto brinda al investigador social la opción de acceder a la *big data*, donde reposan datos con la potencialidad de transformarse en información valiosa, en tanto se apliquen técnicas de minería de datos donde el estudioso trace la ruta delimitando categorías, tiempo y espacio (Lee *et al.*, 2017; Portillo, 2016; Sheldon y Bryant, 2016). La replicación de la investigación en ciencias sociales, fundamentalmente en las áreas donde el análisis de contenido es esencial, es uno de los grandes aportes de la minería de datos (Humphreys y Wang, 2018), que, aunado a la fenomenología y la teoría social, posibilita la construcción del conocimiento fundamentado en análisis macro en ejercicio del principio recursivo de la complejidad.

La transdisciplinariedad, entonces, constituye una estrategia ineludible para el investigador social, en tanto los avances I+D generan saltos cualitativos en los métodos de investigación donde confluyen las ciencias de la computación, estadística, matemáticas y lingüística, entre otras disciplinas (Martínez, *et al.*, 2019; Meneses, 2018), que convergen en la aplicación de métodos metaheurísticos orientados a la generación de nuevo conocimiento e, incluso, la predictibilidad del proceso (Nambisan *et al.*, 2017), mediante el aprovechamiento sistemático de los datos no estructurados que se encuentran dentro de los repositorios de *big data* (George *et al.*, 2016; Antons y Breidbach, 2017).

Diversos estudios en la esfera de género, desigualdad, pobreza, política, salud etc., están siendo abordados desde los datos provenientes de los repositorios de *big data*, mediante el procesamiento de datos no estructurados provenientes de data geoespacial, indicadores de salud, narrativa digital, podcast, IoT, tarjetas de crédito etc. (UN Global Pulse, 2021; Lopes y Bailur, 2018; Paterson y Mc Donagh, 2018; Metcalf y Crawford, 2016).

La *big data* apareció como categoría a mediados de los años 90 en una presentación de Masley (1998) de Silicon Graphics Inc (Diebold, 2012); desde entonces el concepto ha evolucionado. Así, Laney (2001) mencionó las famosas 3 V que identificaban a la *big data*: Volumen, Valor y Velocidad; posteriormente, Gartner (2021) incorpora dos más: Variabilidad y Valor, de tal forma que, para el 2001, se resume la *big data* en la siguiente definición: activos de información de gran volumen, velocidad y variedad que exigen formas rentables e innovadoras de procesamiento de la información para mejorar la visión y la toma de decisiones (Gartner, 2021). Los avances tecnológicos en el área de almacenamiento y velocidades de cómputo, aunado a la reducción de costos, han posibilitado el almacenamiento y procesamiento de ingentes cantidades de información. La inteligencia artificial, los dispositivos IoT y miles de aplicaciones asociadas colectan incesantemente datos del entorno en tiempo real almacenando datos a gran escala y configurando plataformas *big data*. Esta, a su vez, puede clasificarse según su origen en cinco grandes grupos (Rana, 2020):

1. Datos de origen humano: todos aquello datos que han sido generados por la intervención humana directa, blogs, búsquedas en internet, redes sociales, mensajes en redes sociales, foros, etc.
2. Datos generados por máquinas: provenientes de la intervención de dispositivos para generar o medir datos, desde sensores de carreteras o dispositivos científicos hasta internet de las cosas.
3. Datos mediados por procesos: aquellos que se originan en una combinación de humanos y su interacción con equipos, como mensajes por celulares, registros médicos u odontológicos, uso de tarjetas de crédito, datos generados por instituciones gubernamentales y no gubernamentales.
4. Datos de origen mediático: todos aquellos datos provenientes de los *mass media*: videos, imágenes, audio, etc.
5. Datos de origen colectivo: aquellos generados de manera colectiva por los ciudadanos, los que pueden ser de distinta índole.

Esta clasificación no es rigurosa, y si se examina en detalle la data puede considerarse proveniente de una u otra fuente de manera indistinta. La clasificación corresponde a una panorámica de la variedad de datos que diariamente se generan en el mundo y que circulan a través de los canales de comunicación y de conexión del mundo globalizado e interconectado. Desde otra perspectiva, la *big data* se clasifica en dos grandes grupos: aquellos datos que poseen una forma de tipo «estructurada», es decir, poseen una estructura organizativa, y los datos no estructurados, que no poseen esa forma organizativa. El ejemplo de los datos generados por las redes sociales evidencia que la mayor parte de los datos generados son del tipo no estructurado. La empresa Raconteur (2021) estimó que para el 2025 la tendencia será que el 90 % del contenido de datos en Internet será no estructurado.

En ese contexto, la factibilidad de compendiar, clasificar y categorizar segmentos por conceptos o indicadores provenientes de la información generada por cientos, miles o millones de personas, agrupados por patrones, transforma sustancialmente la construcción de agendas públicas, organización social, democratización, equidad, ambiente y en síntesis el alcance de los objetivos del milenio. Las redes sociales y en general la virtualización de las acciones humanas mediante las TIC, transformaron la metodología de investigación social. Las redes como Faceebook, Twitter, Vkontakte, Snapchat, Instagram, y aplicaciones como WhatsApp, constituyen gigantes de la *big data*, derivado de las acciones e interacciones entre los usuarios (Ureña, 2019). No obstante, el acceso a los datos no es factible en todos los casos; algunas plataformas como Twitter han brindado el acceso mediante Application Programming Interface (API) (Digital Guide, 2020). En ese orden, existen grandes repositorios de *big data* disponibles de manera pública y gratuita o privada que ofrecen el servicio a costos diversos; donde los investigadores pueden obtener datos con los que se desarrollen estudios en torno a temáticas específicas.

2. Enfoques de *Big data* en ciencias sociales

Los enfoques de *Big data* han dado paso a nuevas herramientas metodológicas para investigar decisiones y comportamientos humanos más allá de lo que es posible con las formas tradicionales de análisis. No obstante, al igual que cualquier otro paradigma dentro de las ciencias sociales y del comportamiento, *Big data*, no es inmune a una serie de compensaciones típicas: (1) predicción versus explicación, perteneciente a los objetivos generales de la investigación; (2) inducción versus deducción, en cuanto al enfoque epistemológico; (3) grandeza versus representatividad en enfoques de muestreo; y (4) acceso a datos versus independencia científica, abordando las formas de uso de datos (Mahmoodi *et al.*, 2017).

En la actualidad, *Big data*, ha demostrado posee un gran potencial para el abordaje de cuestiones de larga data en las ciencias sociales (Cioffi-Revilla, 2010; George, *et al.*, 2016), y para abrir grandes oportunidades de investigación de experiencias y comportamientos humanos (Kosinski *et al.*, 2015; Snijders *et al.*, 2012). *Big data* ha demostrado, por ejemplo, predecir con éxito la personalidad (Quercia *et al.*, 2011). Por lo tanto, proporciona un amplio acceso a cantidades de datos sin precedentes, ofrece nuevos conocimientos sobre las emociones humanas, cogniciones, motivaciones, decisiones, preferencias, comportamientos e interacciones, y facilita el desarrollo basado en datos de nuevas ideas conceptuales en las ciencias sociales (Chen & Wojcik, 2016).

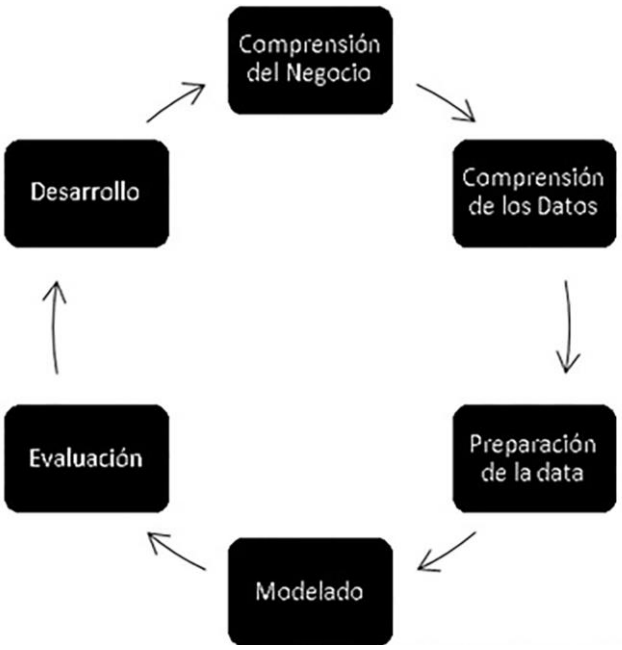
En ese mismo contexto, *Big data*, se anuncia como un nuevo y poderoso recurso para la investigación en ciencias sociales. El entusiasmo por los grandes datos surge del reconocimiento de las oportunidades que puede ofrecer para avanzar en nuestra comprensión del comportamiento humano y los fenómenos sociales de una manera que nunca antes había sido posible (Connelly *et al.*, 2016). Asi pues, en los próximos años, se espera que persista el creciente volumen de datos creados y recopilados en Internet (Kaisler *et al.*, 2013). Sin embargo, la mayor parte de *Big data* sigue siendo desestructurada. En ese sentido, las técnicas computacionales avanzadas están explotando el potencial de la tecnología para capturar y analizar cantidades tan grandes de datos de Internet de formas cada vez más potentes (Eynon, 2013). Esto está ofreciendo a las disciplinas humanísticas y de las ciencias sociales la posibilidad de hacer cuantificables muchos espacios sociales, para que puedan ser estudiados siguiendo un enfoque cuantitativo (Boyd, D., y Crawford, 2012). En realidad, la evolución de los métodos de investigación asistidos por computadora está cambiando la forma en que se realiza la investigación en ciencias sociales y el procesamiento de datos (Demchenko *et al.*, 2013).

3. Minería de datos

El término minería evoca el proceso de excavación y búsqueda intrincada. A su vez, la minería de datos comprende el diseño de métodos que posibiliten la recolección de datos, que exigen procesamiento desde la aparente «babel» ubicada en los repositorios hasta la estructuración de la información en notación numérica, que posibilita el ejercicio metaheurístico. En la minería de datos coexisten diversos enfoques y modelos específicos para la ejecución de estos proyectos. El adscribirse a una metodología o modelo en particular permite realizar las tareas de la minería de datos de manera sistemática y no trivial, o de manera intuitiva. Los enfoques corresponden a un conjunto de pasos para el desarrollo de actividades y tareas organizadas especificando la forma en que deben realizarse. El primer enfoque conocido fue el denominado KDD (knowledge discovery in database) de 1996; en el año 2000, fueron propuestos los modelos Semma, Catalyst y CRISP-DM. Semma es el acrónimo de las 5 fases fundamentales del proceso (SAS Institute, 1998): Sample (Muestreo), Explore (Explorar), Modify (modificar), Model (modelo) y Assess (Valorar). Semma se enfoca en los aspectos técnicos, y excluye las actividades relativas al análisis y entendimiento de lo que se está estudiando. Su origen está asociado a la utilización del *software* SAS, de la compañía SAS Enterprise Miner (SAS, 1998; SAS 2021). Dentro de las metodologías más aplicadas se encuentra la denominada Catalyst, conocida también por el acrónimo P3TQ (*product, place, price, time* y *quatity*), fundamentada en la proposición de dos modelos, el de negocios y el de minería de datos (Oussous, *et al.*, 2017). En el primero se sugiere una guía de pasos o fases para identificar un problema y los requerimientos de la organización involucrada; posteriormente, define actividades específicas para cada escenario.

En el modelo de minería, los pasos propuestos permiten la creación y ejecución de un proyecto a partir del modelo de negocios. El punto focal se encuentra en la cadena de valor de la organización, y su particular enfoque indica que luego de realizar una acción o actividad deben evaluarse los resultados y determinar cuál es el siguiente paso a seguir; es muy flexible y permite la incorporación progresiva de un conjunto de pasos a seguir dependiendo de las exigencias del problema. La metodología más ampliamente utilizada, CRISP-DM (Cross Industry Standard Process for Data Mining), fue ideada por el grupo de empresas SPSS, NCR y Daimler-Chrysler con el respaldo de la Unión Europea (Huber *et al.*, 2019). El enfoque CRISP-DM comprende el minado de datos desde un enfoque analítico, estructurado en seis fases flexibles en el desarrollo de las tareas de cada una, con la singularidad de que el desplazamiento es bidireccional según lo demande el aprendizaje del algoritmo (Figura 1).

Figura 1. CRISP-DM



Fuente: adaptado de Espinoza (2020)

Todos los enfoques comprenden etapas comunes que requieren agotase para alcanzar el objetivo del conocimiento incremental del conocimiento previo; algunos compendian en una misma fase varias tareas, otros introducen fases adicionales. En la primera fase, el investigador establece las categorías y límites de la investigación fundamentado en la información contextual subyacente. Comprende la definición del problema a investigar, solucionar, ajustar y delimitar el tipo de datos que se requiere para el abordaje (Antons *et al.*, 2018). En esta etapa se limita el objetivo, incluso en los casos de minería de datos exploratoria, por cuanto la ambigüedad en la teleología deriva en dispersión de datos y obtención de datos innecesarios u equívocos. Las señales obtenidas potencialmente adolecen de falsedad, por lo que demandan filtrado de los patrones, limitando la aceptación de los que son válidos y relevantes para la respuesta del objetivo. El investigador requiere conocer el tema, el contexto el proceso que genera los datos, la metodología empleada durante la recolección de la data almacenamiento, transformación, reporte y uso (Boullier, 2016); todo ello orientado a establecer una hoja de ruta hacia el logro del objetivo propuesto, considerando los factores que afectan la data, como calidad, cantidad, disponibilidad, ausencia de datos, etc., partiendo del principio de que la correlación entre los atributos de entrada y de salida no son garantía de causación.

En una segunda fase, el investigador adelanta el proceso de exploración y depuración, mediante herramientas simples del análisis exploratorio para comprender y exponer la estructura de los datos, distribución de valores, existencia de anomalías, frecuencia de ocurrencia, valores extremos y atípicos e interrelaciones entre los datos (George, *et al.*, 2016). En esta fase se incluyen controles apropiados para garantizar la seguridad y el nivel de calidad de los datos; mediante la eliminación de datos duplicados, cuarentena de datos atípicos que exceden los límites, estandarización y normalización de los datos, sustitución de datos perdidos, tratamiento de conversión en caso de que los datos lo requieran y diseño de estrategias de tratamiento para datos atípicos. Por último, en esta etapa se determinan los atributos que serán considerados y se ejecuta el muestreo y verificación de la data, lo cual se encuentra asociado a la técnica y el algoritmo que se decide implementar (Kobayashi, *et al.*, 2017). La metaheurística persigue evidenciar patrones ocultos mediante la exposición de relaciones entre los atributos. Dentro de las técnicas aplicadas en el análisis de texto en ciencias sociales, la más utilizada corresponde al recuento de frecuencia de palabras orientadas a medir atributos específicos. Este enfoque se basa en diccionarios preexistentes o no, fundamentados en el conocimiento previo, teorías e, incluso, la experiencia del investigador, quien podrá diseñar un diccionario específico que responda a las necesidades específicas del objetivo propuesto incluyendo opinión, emociones y sentimientos (Pennebaker *et al.*, 2015). Otras técnicas corresponden al análisis de contenido algorítmico supervisado y no supervisado. Las primeras corresponden a la asignación de conceptos limitados por el investigador; posteriormente, el algoritmo de clasificación genera etiquetas de metadatos, que posibilitan la ubicación de estos en la *big data*, incluso generan etiquetas nuevas basadas en el etiquetado anterior. Las técnicas no supervisadas o *clustering* se distinguen por constituir procesos de agrupamiento y prescinden de conocimiento previo, basados en la asociación automática de los conceptos (Antons *et al.*, 2018; Ingersoll *et al.*, 2013).

En una tercera fase se genera un modelo o representación abstracta de los datos y sus relaciones, al seleccionar el algoritmo que se aplicará, ya sea de clasificación, regresión, análisis de asociación, agrupamiento o detección de anomalías. Cada categoría dispone de unas pocas docenas de algoritmos, cada uno con un enfoque ligeramente distinto para resolver el problema. Las categorías de predicción, clasificación y regresión necesitan datos de conocimiento previo, para que el modelo que se construye "aprenda" de esos datos. Mientras que las categorías de agrupamiento y asociación son técnicas de tipo descriptivo, en estas no hay un grupo de datos que sirva de prueba o control, ya que no se desea predecir nada, pues ambas categorías poseen un paso de evaluación del modelo. En la categoría de detección de valores atípicos, si se conoce la data es posible detectar vías técnicas de predicción y si no se conoce mediante técnicas no supervisadas.

La cuarta fase corresponde a la aplicación del modelo, una vez superada la etapa de modelización y abordada las correcciones pertinentes. La última fase corresponde al conocimiento incremental, una vez suprimidos los patrones irrelevantes e identificada la información significativa.

3.1. Repositorios

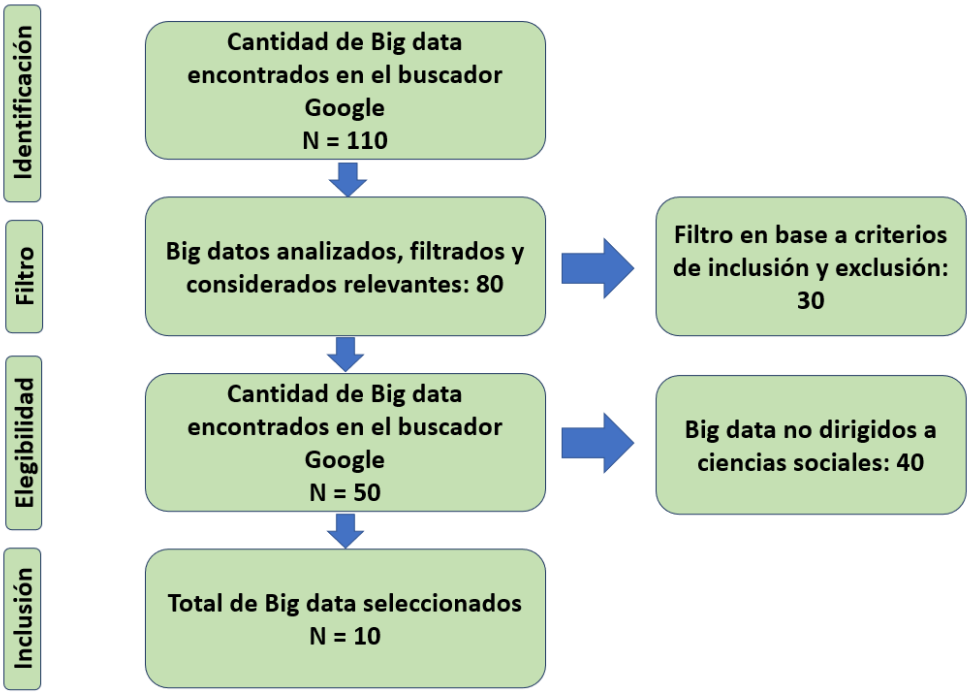
La minería de datos ha evidenciado el potencial como herramienta metodológica para la I+D en ciencias humanas y sociales. Multiplicidad de estudios en el ámbito de la administración (Angus, 2019; Moehrle *et al.*, 2017), sociología (Antons *et al.*, 2018) y otras áreas del conocimiento han evidenciado los alcances de esta herramienta, fundamentalmente en la replicabilidad de las investigaciones. Aunado a ello, en materia de comprobación de hipótesis, la metaheurística posee altos niveles de eficiencia en investigaciones documentales, donde la cuantificación, agrupación y asociación de conceptos y categorías es crítica (Antons *et al.*, 2018).Una buena historia de datos debe comenzar con unos buenos datos, es decir, que sean confiables, reflejen el fenómeno que expresan, consistentes y completos. La gran pregunta es: ¿dónde se encuentran? Existen varias iniciativas para promover el acceso público, gratuito y abierto a grandes bases de datos; entre las tres más importantes se encuentran Open Data Institute (2021); Web World Wide Foundation (2021), y Open Data for Development (2021).

En ese contexto, el objetivo de esta investigación fue explorar y caracterizar los principales repositorios de *big data* en el área de ciencias sociales disponibles en 2021.

4. Metodología

Investigación de diseño no experimental, exploratoria y descriptiva. La población estuvo constituida por 110 *big data* localizados por el motor de búsqueda para conjuntos de datos (*datasets*) de Google. La muestra corresponde a los 10 principales *big data*, contentivos de datos que potencialmente pueden minarse y brindar información para la investigación en el área de ciencias sociales. Las *big data* de la muestra contienen información de diferentes áreas del conocimiento, con la singularidad de constituir los más importantes repositorios en el área de las ciencias sociales. Las categorías utilizadas en la búsqueda fueron «ciencias sociales», «plataformas *big data*», «repositorios *big data*», «corporaciones *big data*». Una vez localizadas las plataformas se procedió a explorarlas, según el orden de relevancia indicado por el posicionamiento indicado por el motor de búsqueda y las recomendaciones de las plataformas y de revistas especializadas e indexadas en Scopus. El proceso realizado se resume en la figura 2.

Figura 2. Proceso de búsqueda de información

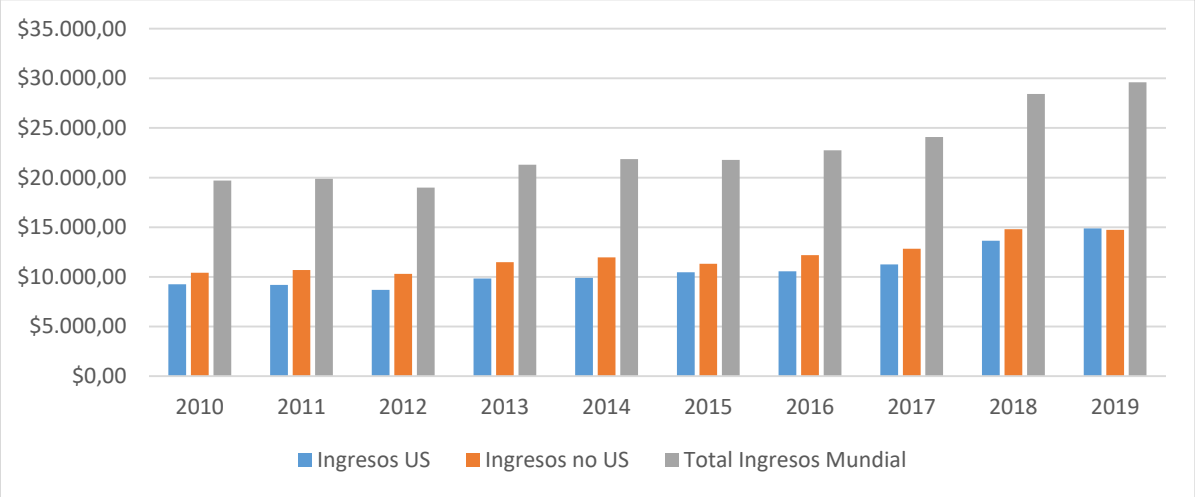


Fuente: adaptado de Seminario-Córdova & Paredes-Gutiérrez (2021)

5. Resultados

En el 2016, un reporte del Banco Mundial (2016) indicaba que la capacidad de almacenamiento en formato comprimido de los datos alcanzaba los 4.2 *zettabits*; un *zettabit* corresponde a unos 10 247 *bytes*, es decir, unos 10 000 terabytes por día. Dada la cantidad de datos que circulan en internet, caracterizados por la dispersión y descentralización, es difícil contar con estadísticas exactas. No obstante, al observar la evolución de los ingresos de las 50 más grandes corporaciones dedicadas al negocio de la *big data* al 2019, se distingue el crecimiento sostenido del sector en la última década:

Figura 3. Evolución de los Ingresos de las 50 más grandes corporaciones mundiales en *big data* en millones $.



Fuente: adaptado de Insights Association (2020)

Los ingresos reportados en la figura 2 se refieren a las 50 empresas privadas (Anexo 1) que brindan servicios como proveedores y repositorios de *big data*. Existen asimetrías en el nivel de ingreso entre las primeras 10 empresas que registraron ingresos anuales al 2019 de $ 24 730 000, que corresponde al 83.5 % del total de los ingresos de las 50 más grandes (Insights Association, 2020). Las diez más grandes incluyen:

1. Nielsen (Nueva York, NY). Empresa con un siglo de existencia especializada en el estudio de mercados, medios, consumo y perfiles del consumidor. Constituye uno de los más grandes repositorios de información socioeconómica del mundo.
2. IQVIA (Danbury, CT; Durham, NC). Con tan solo 5 años en el mercado ha logrado posicionarse como proveedor de servicios de información en uno de los sectores de mayor crecimiento en el uso de la *big data*, como es el sector sanitario. Recaba y procesa información referente al área en todas sus aristas.
3. Gartner Research (Stamford, CT). Con 50 años en el mercado del *marketing*, brinda servicios como proveedor de información por suscripción.
4. Kantar (Nueva York, NY). Empresa de consultoría con 30 años de experiencia como consultora de datos socioeconómicos, políticos y *marketing*.
5. Information Resources Inc. (Chicago, IL): Brinda servicios de predictibilidad de mercados desde hace más de 40 años. Ofrece estrategias y metodologías novedosas derivadas del conocimiento del cliente.
6. Ipsos (Nueva York, NY). Especializada en la recolección y procesamiento de información desde hace más de 40 años; ampliamente conocida en el medio debido a la introducción sistemática de innovaciones en sus investigaciones.
7. Westat (Rockville, MD). Especializada en encuestas e investigación de mercados, es conocida por su versatilidad en el procesamiento de información orientada al cliente.

8. The NPD Group (Port Washington, NY). Especialista en estudios prescriptivos y predictivos de los mercados. Estudia la dinámica de la distribución y el consumo.
9. comScore (Reston, VA). Considerada pionera en la aplicación de la minería de datos para el estudio de los mercados, brinda información sobre el comportamiento de consumidores.
10. GfK (Nueva York, NY). Con 90 años en el mercado, es reconocida por su perspectiva ecléctica en la recolección y procesamiento de datos.

Las mencionadas 10 empresas y las otras 40 que se agrupan en el informe de Insights Association (2020), constituyen importantes repositorios de *big data*, contentivas de valiosa información social, con la impronta que estos repositorios en su gran mayoría no son de acceso público.

En el ámbito de las plataformas de acceso gratuito, y como consecuencia del crecimiento exponencial de los datos que circulan en internet, corporaciones como Google diseñaron motores de búsqueda específicos para *datasets* (conjuntos de datos). El motor de búsqueda de Google (2021) posibilitó la localización de los más importantes repositorios de *big data* de acceso público, posteriormente fundamentados en la relevancia de la plataforma fueron seleccionados los diez principales, en materia de investigación en ciencias sociales.

Tabla 1. Diez principales *big data* y motores de búsqueda en el área de ciencias sociales

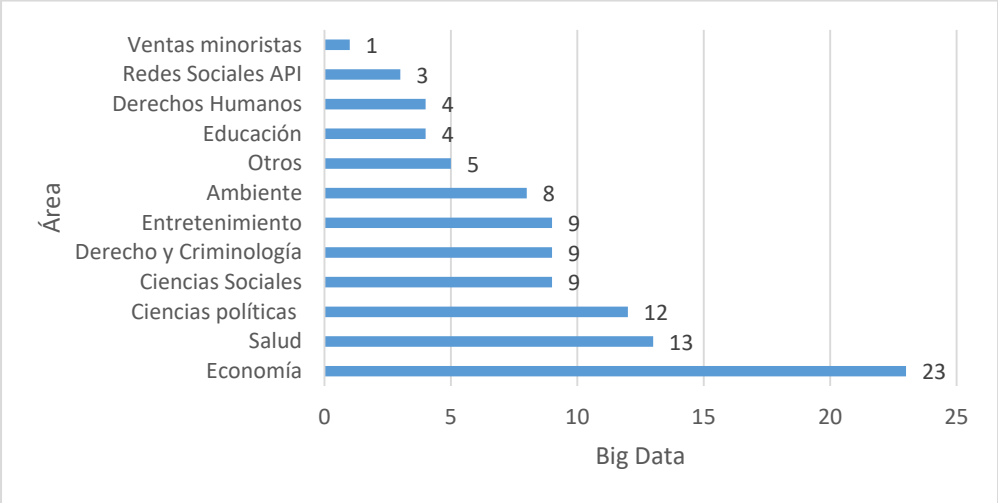
Plataforma	Descripción	Datos	Características
Google data search	Motor de búsqueda de Google referido a conjuntos de datos gratuitos disponibles en Internet	25 millones de *datasets*	Filtros por tipo de *datasets* requeridos, metadata de acuerdo con el standard schema.org
Data Portal	Una de las listas más amplia y numerosa de portales de datos abiertos del mundo. Elaborada por un grupo de expertos en datos abiertos de todo el mundo, entre los que se encuentran representantes de gobiernos locales, regionales y nacionales, organizaciones internacionales como el Banco Mundial y numerosas ONG.	519 portales de *datasets*	Motor de búsqueda y repositorio de enlaces
Elite Data Sciences	Provee lista curada de conjuntos de datos gratuitos para la ciencia de datos y el aprendizaje automático, organizados por su uso.	No disponible	Motor de búsqueda y repositorio de enlaces
Kaggle Dataset	Portal que permite a los usuarios encontrar y publicar conjuntos de datos, explorar y crear modelos en un entorno de ciencia de datos basado en la web; y realizar trabajo colaborativo trabajar con otros investigadores y poder participar en concursos para resolver problemas en ciencia de datos.	16 mil *datasets*	Repositorio de dataset, plataforma para concursos de solución de problemas de minería de datos
Pew Research Center	Portal del Instituto de Investigaciones e información sobre problemáticas, actitudes y tendencias que caracterizan a los	No disponible	Repositorio de *dataset*, posee motor de búsqueda local

	Estados Unidos y el mundo.		
Nature	Portal de la revista *Nature* contentiva de listado de bases de datos discriminados por temas.	No disponible	Repositorio de *dataset*, posee motor de búsqueda local y listas organizadas por temas de los repositorios
ICPSR	Consorcio Interuniversitario para la Investigación Política y Social (ICPSR), con más de 750 instituciones académicas y organizaciones de investigación, ofrece acceso a conjuntos de datos, su conservación y los métodos de análisis para la comunidad de investigadores en ciencias sociales, auspiciado por la Universidad de Michigan.	250.000 *datasets*	Repositorio de *dataset*, posee motor de búsqueda local y listas organizadas por temas de los repositorios
UNData:	Es el motor de búsqueda en Internet que recupera series de datos de bases de datos estadísticas proporcionadas por el sistema de las Naciones Unidas. Constituye un servicio de la División de Estadísticas de las Naciones Unidas (UNSD) desarrollado en colaboración con la Estadística de Suecia y la Agencia Sueca Internacional de Cooperación para el Desarrollo (ASID). UNdata	No disponible	Motor de búsqueda, permite buscar y descargar una variedad de recursos estadísticos que cubren las siguientes áreas: educación, empleo, energía, medio ambiente, alimentación y agricultura, salud, desarrollo humano, industria, tecnología de la información y comunicación, cuentas nacionales, población, refugiados, comercio y turismo.
Amazon Public Data Set	Portal, motor de búsqueda, repositorio y servicios de computación en la nube de la empresa Amazon y referido a conjuntos de datos gratuitos disponibles en Internet.	No disponible	Repositorio de *dataset*, posee motor de búsqueda local y listas organizadas por temas de los repositorios
Hong Kong Batipste University	Repositorio de enlaces de *datasets* de todo el mundo, auspiciado por la Universidad baptista de Hong Kong.	No disponible	Repositorio de enlaces, organizados por temas

Fuente: UN Data (2021); ICPSR Sharing data to advance Science (2021); Nature (2021); Pew Research Center (2021); Kaggle (2021); Elite Data Sciences (2021); Data Portal (2021); Canada Goberment (2021); Web World Wide Foundation (2021); Open Data Institute (2021); Insights Association (2020)

Al agrupar los 100 repositorios gratuitos de *big data* más importantes al año 2021 por sectores, se observa congruencia en relación con los datos arrojados por las más grandes corporaciones de la *big data*. El sector donde confluye la mayor cantidad de repositorios está asociado a la producción, economía y finanzas (Figura 4).

Figura 4. Repositorios de *big data* gratuitos por sector 2021.



Fuente: adaptado de Kilroy (2021)

6. Discusión

La data que brindan las corporaciones privadas de la *big data* comprende información amplia y relevante para los estudios en el ámbito de las ciencias sociales. Aun cuando lo dominante en las principales empresas y sus repositorios comprende el estudio de mercados, ello las induce a compilar ingente cantidad de datos demográficos, socioeconómicos, políticos y culturales que definen el perfil del consumidor expresados en las agrupaciones por clústeres, lo que incluye gran cantidad de datos valiosos en la investigación social que representan un universo por explorar, donde se encuentran respuestas para brechas abiertas en materia de investigación sobre la dinámica de los grupos sociales, acción social, proyectos, etc. De hecho, cuando una empresa adelanta los procesos de agrupación por clúster de las categorías que interesan al cliente, aparecen asociaciones que incluso no siendo del interés del estudio, constituyen datos que potencialmente pueden derivar en valiosa información científica.

El resultado en cuanto a las más grandes corporaciones de la *big data* señala que las empresas más importantes en la prestación de servicios de minería de datos se encuentran dentro del sector de las empresas encuestadoras y de análisis de mercados, dominado la presencia de empresas estadounidenses. En ese orden, Metzler *et al.* (2016), en una investigación referida a la investigación social utilizando *big data*, señalaron que los científicos sociales jerarquizaban como una de las más importantes limitaciones el acceso a los datos, centralizados por empresas privadas, la mayoría con sede en EE. UU. Este dominio de la perspectiva norteamericana sobre las categorías que construirán el conocimiento previo, que, posteriormente sustanciará el algoritmo orientado hacia la localización de la información, genera suspicacias entre los investigadores como Meneses (2018), quien cita el caso Propublica (Angwin *et al.*, 2016) como ejemplo de manipulación de algoritmos. Sin embargo, la replicabilidad de los estudios mediante la descripción metodológica exhaustiva de los procedimientos, categorías, límites tempo-espaciales y la metodología utilizada durante el proceso metaheurístico, constituyen el camino que el científico social necesita recorrer, para evadir tergiversaciones derivadas de la incorrecta o manipulada gestión de los datos y el diseño algorítmico.

El dominio de las grandes corporaciones de la *big data* sobre los datos recabados durante sus investigaciones, limita, en principio el acceso a repositorios, donde se encuentran millones de *terabytes* en datos no estructurados de carácter privado. Sin embargo, iniciativas API como la de Twitter y posteriormente la de Facebook, entre otras iniciativas de acceso abierto, constituyen avances en el acceso y procesamiento de datos con fines científicos y académicos. Lo fundamental para el investigador social es disciplina, ética, rigurosidad y trazabilidad en la denominada cadena de valor de los datos, tal como lo señalan Leonelli y Carrigan (2015), quienes, incluso, recomiendan la indagación integral sobre los métodos de recolección de datos, previo a su incorporación en la base de datos, lo que para el científico social comprende un paso previo, correspondiente a la validación del repositorio y plataforma *big data* de donde extraerá datos. En consecuencia, ello comprende la categorización de la plataforma y sus fuentes, distinguiendo subjetividades derivadas de los roles de los sujetos durante

la generación de contenidos destinados a las redes sociales; en relación con datos derivados de investigaciones donde la subjetividad en el contenido se minimiza por la naturaleza de la recolección, como en los casos de datos demográficos, distribución geográfica, preferencias fundamentada en consumo, como Netflix, Análisis de Redes Sociales (ARS), etc.

En ese sentido, el investigador social confronta un desafío transdisciplinario, en tanto el desarrollo de competencias en materia metaheurística constituye una demanda con un retraso de más de una década. No se trata de tergiversar la esencia del investigador social y los estudios cualitativos, para imponer una suerte de datocracia científica; se trata de convocar a la ciencias de la computación como ciencias conexas a la investigación social, donde el científico social aplica el principio moriano de recursividad (Morin, 2005), define el objetivo de la investigación, construye los diccionarios o amplía diccionarios prexistentes, realiza seguimiento al modelado del algoritmo e interpreta el resultado de los clústeres, fundamentado en la comprensión de lo que representan en el ámbito técnico, epistemológico, fenomenológico y semántico. Comprende la renovación de la reflexividad de lo social inscritos en la propuesta de Boullier (2016), quien aboga por indagar las fuentes y proveniencia de los datos como «huellas digitales», y cómo estas son transformadas en el proceso de recolección, de manera que responden al interés del actor, quien agencia la recolección. Boullier no deslegitima el método fundamentado en la virtualidad, por el contrario, convoca a la replicabilidad a la ciencia y la reflexividad inherentes al investigador social, mediante la indagación, categorización y clasificación de las fuentes, que posibiliten la trazabilidad de la data y validación del método aplicado en su recolección.

7. Conclusión

El objetivo de esta investigación fue explorar y caracterizar los principales repositorios de *big data* en el área de las ciencias sociales disponibles al 2021. Los resultados indican que los repositorios y plataformas más importantes de *big data* se encuentran centralizados por el sector privado localizado en empresas de EE. UU., fundamentalmente. Las plataformas de libre acceso constituyen importantes fuentes de información en las diversas ramas de las ciencias sociales, no obstante, los datos disponibles en su mayoría constituyen datos no estructurados, que demandan del ejercicio transdisciplinario de la metaheurística, lo que es común para las plataformas y repositorios privados de *big data*. En consecuencia, el investigador de lo social requiere el desarrollo de competencias que posibiliten el ejercicio dialógico para con los especialistas en ciencias de la computación, donde la aplicación del principio recursivo de la complejidad devenga en la construcción y aplicación de metodologías de investigación que respondan a los objetivos del investigador social, eludiendo las tergiversaciones derivadas de la datocracia y las mediaciones técnicas. En ese orden, la replicabilidad y reflexividad inherentes a las investigaciones sociales convocan al ejercicio de indagación, categorización y clasificación de las fuentes, en este caso, los repositorios y plataformas *big data*, que posibiliten la trazabilidad de la data y validación del método aplicado en su recolección.

Referencias

American Marketing Association (2022, 12 de febrero). *2020 Top 50 U.S. Market Research and Data Analytics Companies*. https://www.ama.org/marketing-news/2020-top-50/

Angus, R. (2019). Problemistic Search Distance and Entrepreneurial Performance. *Strategic Management Journal, 40*(12), 2011-2023. https://bit.ly/3ZHhGng

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). Machine Bias. *ProPublica*. https://goo.gl/8MAfhK

Antons, D. & Breidbach, C. (2017). *Big data*, Big Insights? Advancing Service Innovation and Design with Machine Learning. *Journal of Service Research, 21*(1), 17-39. https://bit.ly/3k4Hedr

Antons, D., Joshi, A. & Salge, T. (2018). Content, Contribution, and Knowledge Consumption: Uncovering Hidden Topic Structure and Rhetorical Signals in Scientific Texts. *Journal of Management, 45*(7). 3035-3076. https://bit.ly/3XrgcM0

Banco Mundial (2021, 12 de febrero). *Informe Annual 2016.* https://bit.ly/3WXWtDH

Boullier, D. (2016). *Big data challenges for the social sciences: from society and opinion to replications*. Cornel University. https://arxiv.org/abs/1607.05034

Boyd, D., & Crawford, K. (2012). CRITICAL QUESTIONS FOR *BIG DATA*. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Canada Goberment (2021). Open Data for Development. https://www.od4d.net/

Cioffi-Revilla, C. (2010). Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(3), 259–271. https://doi.org/10.1002/wics.95

Connelly, R., Playford, C., Gayle, V., & Dibben, C. (2016). The role of administrative data in the *big data* revolution in social science research. *Social Science Research, 59*, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Chen, E., & Wojcik, S. (2016). Supplemental Material for A Practical Guide to *Big data* Research in Psychology. *Psychological Methods, 21*(4), 458–474. https://bit.ly/3QOJkL7

Data Portal (2021). A Comprehensive List of Open Data Portals from Around the World. Data Portal. https://dataportals.org/search

Demchenko, Y., Grosso, P., de Laat, C., & Membrey, P. (2013). Addressing *big data* issues in Scientific Data Infrastructure. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 48–55. https://doi.org/10.1109/CTS.2013.6567203

Diebold, F. (2012). *On the Origin(s) and Development of the Term "Big data".* SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2152421

Digital Guide (2021, 12 de febrero). Application Programming Interface (API): cómo se comunican las aplicaciones. *Digital Guide*. https://bit.ly/3w49Q9n

Elite Data Sciences. (2021, 10 de febrero). Datasets for Data Science and Machine Learning. Data Sets. *Elite Data Sciences*. https://elitedatascience.com/datasets

Espinosa, J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería, investigación y tecnología*, *21*(1). https://bit.ly/3ka9RpV

Eynon, R. (2013). The rise of *Big data*: what does it mean for education, technology, and media research? Learning, *Media and Technology*, *38*(3), 237–240. https://bit.ly/3CHvMLk

Gartner (2021). *Gartner Glosary*. Gartner. https://gtnr.it/3CH6MUB

George, G., Osinga, E., Lavie, D. & Scott, B. (2016). *Big data* and Data Science Methods for Management Research. *Academy of Management Journal, 59*(5), 1493-1507. https://bit.ly/3IJtorl

Hong Kong Baptiste University & Library. (2021) *Data across countries.* https://bit.ly/3XkI9EU

Huber, S., Wiemer, H., Schneider, D. & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. *Procedia CIRP, 79,* 403-408. https://www.sciencedirect.com/science/article/pii/S2212827119302239

Humphreys, A. & Wang, R. (2017). Automated Text Analysis for Consumer Research. *Journal of Consumer Research, 44*(6), 1274-1306. https://bit.ly/3X6TEAk

ICPSR Sharing data to advance Science (2021). *Home*. https://www.icpsr.umich.edu/web/pages/

Ingersoll, G., Morton, T. & Farris, A. (2013). *Taming text: How find, organize, and manipulate it.* Manning Publications Co.

Insights Association (2020). Research & data analytics industry. *Top 50 Report US, 2020.*

Kaggle (2021). *Datasets*. https://www.kaggle.com/datasets

Kaisler, S., Armour, F., Espinosa, J., & Money, W. (2013). *Big data*: Issues and Challenges Moving Forward. *2013 46th Hawaii International Conference on System Sciences*, 995–1004. https://doi.org/10.1109/HICSS.2013.645

Kilroy, J. (2021). 100+ of the Best Free Data Sources for Your Next Project. *Colum Five.* https://www.columnfivemedia.com/100-best-free-data-sources-infographic

Kobayashi, V., Mol, S., Berkers, H., Kismihók, G. & Den Hartog, D. (2017). Text Classification for Organizational Researchers. *Organizational Research Methods, 21*(3), 766-799*.* https://journals.sagepub.com/doi/10.1177/1094428117719322

Kosinski, M., Matz, S., Gosling, S., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543–556. https://doi.org/10.1037/a0039210

Laney, D. (2001). 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note. *Scientific Research.* https://bit.ly/3ZxMVAO

Lee, J., Kim, C. & Shin, J. (2017). Technology opportunity discovery to R&D planning: Key technological performance analysis. Technological Forecasting and Social Change*, Elsevier, 119*(C), 53-6. https://www.sciencedirect.com/science/article/abs/pii/S0040162516305893

Leonelli, S. y Carrigan, M. (2015). Sabina Leonelli: What constitutes trustworthy data changes across time and space. *lse: Impact of Social Sciences Blog*. https://bit.ly/3GCChQG

Lopes, C. & Bailur, S. (2018). *Gender Equality and Big data*. UN Women. https://bit.ly/3iD1Toz

Mahmoodi, J., Leckelt, M., van Zalk, M., Geukes, K., & Back, M. (2017). *Big data* approaches in social and behavioral science: four key trade-offs and a call for integration. *Current Opinion in Behavioral Sciences*, *18*(59), 57–62. https://doi.org/10.1016/j.cobeha.2017.07.001

Martínez, F., Contreras, L., Ferri, C., Hernández, J., Kull, M., Lachiche, N. & Flach, P. (2019). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering, 1*(1). https://bit.ly/3kaaiR5

Masley, J. (1998). *Big data* and the Next Wave of InfraStress. *Computer Systems Laboratory Colloquium February 25, Silicon Valley.* https://web.stanford.edu/class/ee380/9798win/lect08.html

Mayer, V. & Kenneth, C. (2014). *Big data: A Revolution that will Transform how we Live, Work, and Think*. Houghton Mifflin Harcourt.

Meneses, M. (2018). Grandes datos, grandes desafíos para las ciencias sociales. *Revista Mexicana de Sociología*, *80*(2), 415-444.

Metcalf, J. & Crawford, K. (2016). Where are human subjects in *Big data* research? The emerging ethics divide. *Big data & Society*, *3*(1), 1-14. https://doi.org/10.1177/2053951716650211

Moehrle, M., Wustmans, M. & Gerken, J. (2017). How business methods accompany technological innovations - a case study using semantic patent analysis and a novel informetric measure. *R&D Management, 48*(3), 331–342. https://onlinelibrary.wiley.com/doi/10.1111/radm.12307

Nambisan, S., Lyytinen, K., Majchrzak, A. & Song, M. (2017). Digital innovation management: reinventing innovation management research in a digital worl. *MIS Quartely, 41*(1), 223-238. https://bit.ly/3XcUBqY

Nature (2021). *Scientific Data*. Nature. https://www.nature.com/sdata/policies/repositories

Open Data Institute (2021). We want a world where data works for everyone. https://theodi.org/

Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A. & Belfkih, S. (2017). *Big data* technologies: A survey. *Journal of King Saud University-Computer and Information Sciences, 30*(4), 431-448. https://www.sciencedirect.com/science/article/pii/S1319157817300034

Paterson, M. & Mc Donagh, M. (2018). Data Protection in an era of *Big data*: The challenges posed by big personal data. *Monash University Law Review*, *44*(1), 1-31. https://bit.ly/3k96jUI

Pennebaker, J., Boyd, R., Jordan, K. y Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015.* Texas University. https://bit.ly/3H16Tgq

Pew Research Center (2021). *Download Datasets*. https://www.pewresearch.org/download-datasets/

Portillo, J. (2016). Planos de realidad, identidad virtual y discurso en las redes sociales. *Logos (La Serena)*, *26*(1), 51-63. http://dx.doi.org/10.15443/RL2604

Pyle, D. (2003). *Business Modeling and Data Mining*. Morgan Kaufmann Publishers.

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. *2011 IEEE Third Int'l Conference on Privacy, Security, Risk*

and Trust and 2011 IEEE Third Int'l Conference on Social Computing, 180–185. https://doi.org/10.1109/PASSAT/SocialCom.2011.26

Raconteur (2021). *Content for business decision-makers.* https://www.raconteur.net/

Rana, A. (2020). Leveraging *Big data* to Advance Gender Equality. *EMCompass (86).* https://openknowledge.worldbank.org/handle/10986/34308

SAS Enterprise Miner. (2021). Reveal valuable insights with powerful data mining software. SAS Enterprise Miner. https://www.sas.com/en_th/software/enterprise-miner.html

SAS Institute. (1998). Data Mining and the Case for Sampling. https://bit.ly/3XsDi4W

Seminario-Córdova, R., & Paredes-Gutiérrez, P. (2021). Principales factores influyentes en el incremento de casos de violencia contra la mujer en Perú: contexto pandémico. *Social Innova Sciences, 2*(3), 17–35. https://socialinnovasciences.org/ojs/index.php/sis/article/view/61/74

Sheldon, P. & Bryant, K. (2016). Instagram: Motives for its use and relationship to narcissism and contextual age. *Computers in Human Behavior*, *58*, 89-97. https://bit.ly/3k5RgLv

Snijders, C., Matzat, U., & Reips, U.-D. (2012). Structural color and microstructure of ligament in bivalve shells of Cyclina sinesis. *International Journal of Internet Science, 7*(1), 1–5. https://bit.ly/3XlRrAI

UN Data (2021). A World of information. *UN Data*. https://data.un.org/

UN Global Pulse (2021). *Big data* and Artificial Intelligence. https://www.unglobalpulse.org/

Ureña, R. (2019). Autoridad algorítmica: ¿cómo empezar a pensar la protección de los derechos humanos en la era del "*big data*"? *Latin American Law Review*, *2*, 99-124. https://doi.org/10.29263/lar02.2019.05

Web World Wide Foundation (2021). Open Data Barometer. https://opendatabarometer.org/