KARPETA

Teknokultura. **Revista** de Cultura Digital y Movimientos Sociales

https://doi.org/10.5209/tekn.83875

EDICIONES COMPLUTENSE

El Big Data como metodología de investigación social: Propuestas, renuncias y dilemas desde la sociología

Héctor Puente Bienvenido[1]; Diego de Haro Gázquez[2], Sergio D'Antonio Maceiras[3]

Recibido: 22 de septiembre de 2022 / Aceptado: 28 de febrero de 2023 / Publicación en línea: 9 de marzo de 2023 / **OPR**

Resumen. En el ampliamente citado artículo 'The coming crisis of empirical sociology' anticipaban una inminente problemática a la que se enfrentaría la sociología empírica, en su incapacidad de incorporar en su práctica investigadora los datos 'transaccionales' propios del capitalismo cognitivo. La progresiva extensión de la intermediación tecnológica en los procesos de interacción social genera una importante barrera de carácter tecnocientífico, pero sobre todo económica y política, que relega a la sociología (y sus métodos) a la retaguardia de la investigación social y de mercado. En este artículo haremos una suerte de arqueología de la construcción del concepto de 'Big Data', de las principales alternativas metodológicas que se han ido proponiendo desde las mismas, y de las nuevas perspectivas que se abren a partir de la concentración de la información en un menor número de empresas.
Palabras clave: base de datos; capitalismo informacional; ciencia y sociedad; datos abiertos; métodos digitales.

[en] Big Data as a methodology for social research: Proposals, disclaimers, and dilemmas from sociology

Abstract. Since the widely cited 'The Coming Crisis of Empirical Sociology' foresaw an imminent problem that empirical sociology would face, in its inability to incorporate the 'transactional data' of cognitive capitalism into its research practice. This is due to the progressive extension of technological intermediation in the processes of social interaction, generating an important techno-scientific barrier, but above all an economic and political one, which relegates sociology (and its methods) to the rearguard of social and market research. In this article, we will make a sort of archaeology of the construction of the concept of 'Big Data', of the main methodological alternatives that have been proposed by the social sciences, and of the new perspectives that are opening up as a result of the concentration of information in a smaller number of companies.
Keywords: databases; digital methods; informational capitalism; open data; science and society.

1 Universidad Complutense de Madrid (España)
E-mail: hector.puente@ucm.es
ORCID: https://orcid.org/0000-0002-7441-1908

2 Universidad Complutense de Madrid (España)
E-mail: hgazquez@ucm.es
ORCID: https://orcid.org/0000-0003-1350-725X

3 Universidad Politécnica de Madrid (España)
E-mail: sergio.dantonio@upm.es
ORCID:https://orcid.org/0000-0001-8320-0902

Como citar: Puente-Bienvenido, Héctor, de Haro-Gázquez, Diego y D'Antonio-Maceiras, Sergio (2023). El Big Data como metodología de investigación social: Propuestas, renuncias y dilemas desde la sociología. *Teknokultura. Revista de Cultura Digital y Movimientos Sociales*, *20*(2), avance en línea, 1-15. https://doi.org/10.5209/tekn.83875

1. Introducción: Investigando en tiempos del capitalismo informacional

En el momento del surgimiento de la pandemia de la COVID-19 en marzo de 2020, el Instituto Nacional de Estadística de España (INE) hacía casi un año que había firmado respectivos acuerdos de colaboración con las tres principales compañías de comunicación del país —Movistar, Orange y Vodafone— para la realización de un primer estudio de movilidad ciudadana entre municipios al que se llamó EM-1 (INE, 2021), a partir de los datos geoposicionados de los teléfonos móviles de aquellas personas que fueran clientes de alguna de las operadoras mencionadas. Estos acuerdos se renovaron con el pretexto de realizar un seguimiento epidemiológico de la pandemia y de la necesidad de controlar la movilidad para poder realizar modelos y previsiones de contagio, en otros tres proyectos: EM-2 (del 16 de marzo al 20 de junio 2020), EM-3 (de junio a diciembre de 2020) y EM-4 (durante 2021) (INE, 2021).

Se trataba de proyectos de investigación relativamente innovadores en el ámbito de las instituciones públicas ya que, aunque encontramos con facilidad distintos precursores como el *Dutch Mobile Mobility Panel* (Thomas et al., 2018), en este caso el registro de la movilidad de los ciudadanos a partir de los datos geoposicionados de sus teléfonos se realiza a partir de una aplicación específica, que instala una pequeña muestra de los ciudadanos del país durante un periodo de tiempo determinado (tras previo consentimiento para participar en dichas investigaciones).

El caso de los proyectos EM del INE era radicalmente diferente por diversas razones.

- La primera de ellas es que en ningún momento había un consentimiento previo de las personas 'participantes' en este estudio (es decir, potencialmente se incluía a todos los ciudadanos que hubieran contratado su línea de móvil con alguna de las tres principales compañías del país). Según el documento técnico del primer estudio (INE, 2019), la muestra estaba constituida por el 80% de los ciudadanos con teléfonos móviles, que es la cuota de mercado que alcanza la suma de las tres principales operadoras del país. Tampoco existía en todos los casos, tal y como ha investigado la agencia de noticias Newtral, un procedimiento para poder desistir de la participación en dicho estudio, más allá de cancelar el contrato de prestación de servicios con dicha empresa, o activar el 'modo avión' (Rodrigo, 2019).
- La segunda gran diferencia, que está recogida de datos se realiza de forma inconsciente para las personas participantes. En investigaciones análogas

sobre movilidad intermunicipal la persona debe hacer, bien un ejercicio de recuerdo al rellenar el diario de movilidad del periodo de referencia, bien una acción consciente de activar la aplicación de su teléfono móvil que registrará sus movimientos, vía GPS, en el periodo de estudio acordado. En el caso de los proyectos del INE, sólo con posterioridad a la publicación de los datos queda claro cuál era la muestra de días que entraban dentro del campo del proyecto.

– La tercera diferencia respecto a otras investigaciones afines radica en los contratos «negociados sin publicidad» a causa de «motivos técnicos», tal y como aparece descrito en los diferentes anuncios de contratación realizados en el Boletín Oficial del Estado (BOE-B-2020-42528, BOE-B-2020-46213, BOE-B-2020-47419, BOE-B-2021-25336, BOE-B-2021-18242, etc.).

El motivo técnico es sencillo de entender: si deseamos hacer un estudio de movilidad a partir de una matriz de localizaciones geoposicionadas mediante las antenas de los teléfonos móviles de la ciudadanía y tres grandes empresas conforman un oligopolio que domina el 80% de la cuota de mercado, no existe otra alternativa que pagar el precio que quieran pedir por los *datasets* (precio que, conforme a los anuncios publicados en el BOE, asciende a unos 300.000 euros por cada una de las matrices y operadoras); o bien renunciar a la posibilidad de explotar esos datos procedentes de la trazabilidad digital (Venturini y Latour, 2009) para satisfacer nuestros objetivos de investigación.

No es casualidad que otro caso paradigmático, que además se suele utilizar como uno de los primeros ejemplos de utilización del Big Data con un 'interés público' (el Google Flu Trends, ver Mayer-Schönberger et al., 2013), sea en el marco de otra pandemia, la de la gripe A H1N1 de 2009. Sin embargo, no fue ningún instituto de investigación, sociedad científica, hospital o Ministerio de Sanidad quien realizara dicha investigación, sino seis ingenieros de Google (Ginsberg et al., 2009) quienes publicaron en la revista *Nature* un modelo realizado a partir de los datos de diferentes términos de búsqueda, capaz de trazar el avance de la enfermedad en los diferentes territorios de Estados Unidos con una granularidad asombrosa.

Tal y como Viktor Mayer-Schönberger et al. (2013, p. 7) explican, «otros ya habían intentado hacer esto con los términos de búsqueda de internet, pero nadie disponía de tantos datos, capacidad de procesarlos y *know-how* estadístico como Google». Algo que fue destacado, desde una perspectiva significativamente más crítica, por David Lazer et al. (2014, p. 1205):

> La replicabilidad es una preocupación creciente en el mundo académico. Los materiales que soportan [el estudio de Google] no llegan a alcanzar los estándares exigidos. […] Es imposible para Google hacer público su pleno arsenal de datos para los investigadores externos, e incluso puede que ni siquiera fuera éticamente aceptable debido a problemas de privacidad. No obstante, este problema no existiría con datos derivativos o agregados. Pero incluso para un investigador con acceso completo a todos los datos de Google, sería imposible replicar los análisis realizados. […] Twitter, Facebook, Google, e Internet en su conjunto se

encuentran siempre en constante cambio debido a las acciones de millones de ingenieros y consumidores.

En definitiva, creemos que las tres características mencionadas ilustran a la perfección las principales problemáticas a la hora de plantear una investigación a partir de lo que se ha dado en llamar Big Data. Estas son: 1) se trata de datos involuntarios, 2) configurados a partir de trazas digitales fundamentalmente inconscientes y de límites difusos, y 3) que son propiedad de las grandes empresas prestadoras de servicios de la Sociedad de la Información.

Si bien es cierto que, siguiendo a Tommaso Venturini (2012, p. 800), podríamos decir que «a través de la mediación digital, la trazabilidad y la agregabilidad se han convertido en prestaciones (*affordances*) intrínsecas de los fenómenos sociales», planteamos que el ritmo desenfrenado de concentración de capital característico de este momento histórico —especialmente exacerbado en el caso de las empresas de carácter tecnológico— lleva aparejado una generación artificial de escasez sobre la información que la convierte, irónicamente, en algo cada vez menos asequible, en manos de un número reducido de agentes y difícilmente asimilable a los estándares abiertos de la práctica investigadora. Todo ello hace que las diferentes propuestas metodológicas que se han ido construyendo desde las ciencias sociales partan necesariamente de una renuncia económica y política, en tanto en cuanto no aborden este dilema fundamental.

En este artículo pretendemos hacer un breve repaso de las diferentes definiciones construidas en torno al concepto de Big Data, así como un recorrido crítico por las principales propuestas metodológicas que se han realizado en el ámbito de la sociología desde esta perspectiva crítica.

2. ¿Qué es el Big Data?

En cuanto a su popularización más comercial y uso en la práctica cotidiana, las entidades bancarias, compañías de gestión de riesgos o aseguradoras, fueron pioneras en el uso del Big Data para optimizar sus servicios y actividades. Prevención de fraude, transacciones financieras, interacciones y registros de pago con tarjeta bancaria o análisis de riesgos, son algunos de los primeros campos donde se implementaron estrategias basadas en Big Data especialmente a partir de finales de los años 2000. De este modo, compañías como BNP Paribas o JP Morgan fueron las primeras entidades en incorporar dichos métodos para optimizar sus procesos de trabajo con datos transaccionales y maximizar su rendimiento y beneficio. A día de hoy, aunque las tecnologías, espacios y ámbitos donde se utiliza Big Data son ubicuos, sigue asociados principalmente a estos campos, junto a otros como comercio online, redes sociales, Internet de las cosas (IoT) o telecomunicaciones (Sánchez-Holgado, 2021).

¿Pero a qué nos referimos antes con datos transaccionales? Mike Savage y Roger Burrows (2007, 2009) los definieron como un tipo de información o dato generado como un subproducto digital de transacciones y prácticas rutinarias entre ciudadanos, consumidores, empresas u otro tipo de organizaciones públicas y privadas

(transacciones de compra, gestión de inventarios, fluctuaciones bursátiles, *trending topics* en redes sociales, flujos de interacción en navegación web, itinerarios y rutas seguidas en transporte, segmentación de anuncios y de contenidos audiovisuales en plataformas como Netflix, etc.). Alojados en grandes y complejos sistemas de bases de datos comerciales y gubernamentales, parece claro que los datos transaccionales constituyen una parte crucial de las infraestructuras de información del capitalismo tardío, que revelan profundas implicaciones sociológicas.

Entendiendo el Big Data como fenómeno sociotécnico, en los dos siguientes epígrafes vamos a abordar sus dimensiones y particularidades como objeto de estudio y como método propio de las Ciencias Sociales.

2.1. El Big Data como objeto de estudio

En origen, el término Big Data se remite a mediados de la década de 1990 y aparece por primera vez en un conjunto de trabajos de carácter industrial desarrollados por John Mashey como científico-jefe de Silicon Graphics. Inicialmente, él y su equipo utilizaron el término para referirse al manejo y análisis de conjuntos de datos masivos (Diebold, 2012), evidenciando una problemática clara de volumen de información (grandes cantidades de datos, que pueden generarse, por ejemplo, a partir de teléfonos móviles, estaciones meteorológicas, tarjetas de crédito, drones o fotografías). Respecto al ámbito académico, las primeras discusiones y referencias sobre el concepto aparecerán unos años más tarde, concretamente en las investigaciones iniciales de Sholom Weiss y Nitin Indurkhya (1998), en el ámbito de la informática y de Peter Christoffersen y Francis Diebold (2000) en los campos de la estadística y de la econometría.

Prontamente, el concepto de Big Data se complejizó, revelando nuevos matices y problemáticas definitorias. En 2001, Doug Laney, además de la cuestión del volumen (consistente en manejo de cantidades masivas de datos), añadiría otros dos rasgos definitorios al término de Big Data (conocidos como las 3 V's). Estos son: la velocidad (los datos son generados a tiempo real y se procesan y analizan con gran rapidez) y la variedad (problemáticas relacionadas con el formato, integración y estructuras de datos). Por ejemplo, respecto a la variedad, ya no se opera meramente con datos estructurados que pueden ser mostrados de manera clara en una tabla (con campos como el nombre, teléfono, email o ID), sino que también incorporamos el reto de manejar datos no estructurados como audios, imágenes, o datos provenientes de medios sociales como Twitter, Facebook o Instagram, entre otros. Y no solo eso, sino que la variedad y diversidad de datos también hace referencia a la fuente de estos. El Big Data no se refiere meramente a datos provenientes de interacciones o transacciones humanas, sino que, desde la perspectiva técnica, también incluye extracción y tratamiento de otros datos con orígenes y procesos muy diversos, como meteorológicos, genómicos o astrofísicos.

Desde esta conceptualización dominante e inicial de las 3 V's, otros autores y autoras han atribuido múltiples características adicionales al Big Data (Kitchin y McArdle, 2016), entre las que destacarían: i) exhaustividad y posibilidad de

explotación segmentada (Mayer-Schönberger et al., 2013), ii) relacionalidad o posibilidad de combinar diferentes conjuntos de datos (boyd y Crawford, 2012), iii) veracidad, sistematicidad y riguroso control de errores (Marr, 2014), iv) disponibilidad de profundización y de reutilización (Marr, 2014), v) variabilidad y mutabilidad (transaccionalidad, volatilidad y datos cuyo significado y propiedades pueden cambiar constantemente en relación con el contexto en el que se generan y alojan) (Kitchin y McArdle, 2016; McNulty, 2014). Así pues, el concepto de Big Data sigue a día de hoy en pleno proceso de replanteamiento y reconfiguración, dando cuenta de las múltiples aristas e implicaciones derivadas del entramado de redes científico-técnicas, sociopolíticas y económicas (Venturini, 2010; Venturini et al., 2017) en las que se renegocian sus características, potencialidades y límites.

2.2. El Big Data como método

A medida que la World Wide Web fue creciendo a finales de los 1990 y principios de los 2000 (pasando de docenas a millones de páginas), tareas que inicialmente eran desarrolladas por humanos (búsqueda y localización de información relevante) empezaron a requerir automatización. Se crearon los primeros rastreadores Web, muchos de ellos como iniciativas lideradas por universidades o compañías privadas (Yahoo, AltaVista, etc.). Durante estos años, surgieron proyectos como los buscadores Nutch o Google, cuya finalidad principal era reducir los tiempos de búsqueda a partir de la distribución de datos y cálculos en diferentes ordenadores a fin de que «se pudieran procesar múltiples tareas de manera simultánea» Este tipo de tecnologías de almacenamiento y procesamiento de datos de manera distribuida (redes de máquinas conectadas que ejecutan un mismo proceso paralelamente) constituyeron el germen de lo que sería el actual Hadoop, una de las tecnologías más comúnmente asociadas al Big Data. El proyecto de código abierto Hadoop desarrolla software que, en vez de utilizar un equipo o máquina muy potente para procesar y almacenar los datos, permite la creación de clústeres de hardware para analizar ingentes conjuntos de datos simultáneamente. En la actualidad, aunque Hadoop cuenta con varios módulos, es una tecnología en la que destacan dos grandes componentes, su sistema de almacenamiento (HDFS) y su sistema de procesamiento (MapReduce).

Se suele decir que el método del Big Data consiste en dividir y paralelizar, pero para realizar dichas tareas se requiere una infraestructura organizada en nodos y clústeres. De manera sencilla, podemos decir que un nodo es una máquina de almacenaje y procesamiento (como si fuera un ordenador personal) mientras que un clúster sería un conjunto de nodos que operan de manera coordinada. Los nodos, a su vez, se subdividen entre nodos maestros y trabajadores en función de las tareas que realizan. Los primeros dividen y distribuyen las tareas en procesos más pequeños y los segundos las ejecutan. En función de las necesidades, se podrán activar o apagar nodos y clústeres (cuanto mayor sea la necesidad de procesamiento y almacenamiento mayor será el número de estos, permitiendo cierta escalabilidad o procesamiento a medida). Este tipo de tareas de gestión y planificación de recursos

suelen ser realizadas con sistemas como MapReduce o, más recientemente, Yarn (Hadoop 2.0).

Desde el ámbito de las Ciencias Sociales, este tipo de métodos y procesos muchas veces resultan confusos y ciertamente opacos, existiendo una clara brecha de competencia técnica que opera como una caja negra (Latour, 2021). Sin embargo, el análisis sociológico no puede quedar limitado a la mera observación y análisis de los datos resultantes (consecuencias o impactos) sino que el entendimiento y crítica de los procesos (y de las tecnologías mediadoras) debe ser también un aspecto relevante en su estudio (revelar lo invisibilizado tecnológicamente para comprender las estructuras sociotécnicas y redes de poder) (Rogers, 2013).

3. Propuestas y aproximaciones desde las ciencias sociales

Aunque el almacenamiento y consecuente análisis de los datos transaccionales ha sido aprovechado por las empresas prestadoras de servicios digitales desde el principio, no sería hasta mucho más tarde cuando las ciencias sociales empezarían a usar esas grandes bases de datos transaccionales con propósitos de investigación social.

3.1 La etnografía virtual/digital y los primeros estudios de la cibercultura

Fue con la popularización de Internet como medio de interacción social cuando las ciencias sociales empezaron a utilizar herramientas que simultáneamente eran sustentadas y generadoras de grandes bases de datos transaccionales (motores de búsqueda, listas de correo, Usenet, nodos de IRC, MUDs, etc.)

Y, aunque podemos encontrar fundamentalmente a partir de los años noventa diferentes ensayos y propuestas teóricas sobre cambios sociales producidos a partir de la incorporación de dichas tecnologías en nuestras vidas (Rheingold, 1993; Turkle, 1995) todavía existía una separación textual entre lo virtual de aquellas prácticas frente a las experiencias reales (Gómez-Cruz, 2007). No fue hasta finales de los años noventa y principios de siglo cuando comenzamos a ver las primeras propuestas metodológicas que habilitaban Internet no sólo como un objeto de estudio, sino también como un campo, que además iba a requerir de un nuevo arsenal metodológico (Gómez-Cruz, 2020, 2007).

Steve Jones edita en 1999 el libro *Doing Internet research*, en el que se comienza a plantear, de manera específica al campo de las ciencias sociales, diferentes problemáticas sobre la investigación social en/sobre el ciberespacio y la aplicación o validez de las diferentes perspectivas metodológicas de las que disponían hasta el momento. Partiendo de algunos mapeos previos sobre potenciales objetos sociales de estudio a partir de Internet y agentes interesados en realizar dichas investigaciones (Rice, 1989), Jones afirma que

> Internet es un tipo diferente de cosa […] y en consecuencia requiere un cambio consciente en el foco. Y, en cualquier caso, esperaré que podamos seguir cambiando el foco de manera continua […] y no anclemos nuestra mirada de una

> manera o de otra, o fallaremos en captar la mutabilidad esencial de Internet (Jones, 1999, p. 241).

Por su parte, Christine Hine publica en el año 2000 su precursora propuesta (*Virtual ethnography*) en el que define simultáneamente Internet como cultura y como artefacto cultural, reflexionando sobre distintas problemáticas emergentes tras la posibilidad de realizar trabajo etnográfico desde/a partir/sobre el ciberespacio: autenticidad, identidad, interactividad, presencia constante/ubicua del etnógrafo, desdibujamiento del campo, etc. (Hine, 2000). Este germen metodológico, además, sigue siendo desarrollado en la actualidad (Hine, 2015; Kitchin y McArdle, 2016; Pink et al., 2015). No obstante, aunque en ambos planteamientos se parte de una utilización extensiva de tecnologías como el correo electrónico, la WWW, los motores de búsqueda, Usenet, los MUDs, etc., no vemos todavía una reflexión del potencial de utilización de dichos datos transaccionales de los que parten dichas herramientas, y cuyas investigaciones iban a colaborar a alimentar. Habrá que esperar algunos años más para encontrar las primeras reflexiones sobre el Big Data (aún sin llamarlo por este nombre).

3.2. La crisis que viene de la sociología empírica, la ciencia social del siglo XXI y el manifiesto de las ciencias sociales computacionales

En el año 2007 coincidió la publicación de dos textos que posteriormente se han convertido en esenciales para comprender la construcción del Big Data desde la sociología, tanto como amenaza apocalíptica como oportunidad histórica.

La primera de las perspectivas quedó brillantemente plasmada en el que pronto se convirtió en uno de los artículos más citados de la historia de la revista *Sociology*: '*The coming crisis of empirical sociology*' (Savage y Burrows, 2007). Este texto destaca la relevancia actual de las distintas disciplinas (desde el marketing hasta los *data scientist*s) que acometen regularmente investigaciones sobre ciertos aspectos de lo social, en los que la sociología empírica tradicional lleva a cabo un papel marginal —o en muchas ocasiones inexistente—, precisamente por su rezago en comprender las implicaciones que la existencia de las grandes bases de datos transaccionales conllevan y su déficit en innovaciones metodológicas para hacerles frente. Ante esta evidente incapacidad sociológica de incorporar los datos transaccionales en su práctica investigativa, los autores sugirieron tres posibles estrategias de respuesta: i) la necesidad de repensar los repertorios de la sociología empírica en una era del 'capitalismo cognitivo' (donde el conocimiento e información son privatizados y mercantilizados); ii) considerar centrales 'las políticas del método' y sus diversas implicaciones (entendiendo las herramientas de investigación social como un producto intrínseco de la organización capitalista); y iii) vincular críticamente narrativas, números e imágenes de manera que interactúen con los tipos de análisis transaccionales en auge.

Por otro lado, en ese mismo año, el también sociólogo Duncan Watts publicó una breve reflexión en la revista *Nature* a la que tituló 'Una ciencia del siglo XXI' (Watts, 2014), en la que se pregunta cómo los grandes avances técnicos derivados de Internet

y el procesamiento computacional pueden conducir también a una revolución en las Ciencias Sociales (ya que por primera vez se podían observar a tiempo real millones de interacciones con un nivel de refinamiento que permitía profundizar o segmentar hasta el nivel individual).

Las principales ideas que aquí se plasmaron posteriormente fueron convertidas en un manifiesto firmado por quince académicos de reconocido prestigio, en el año 2009, por una 'ciencia social computacional' (*Manifesto of computational social science*, Conte et al., 2012; Watts, 2014). En dicho texto continuista, se reivindica el potencial de la ciencia social computacional, como campo interdisciplinar, para dar respuesta a los problemas sociales más acuciantes del momento (desigualdad, salud pública, inestabilidad económica, cambios en la estructura poblacional, crimen organizado, etc.). De tal manera que, así como las Ciencias Físicas prueban sus modelos a través de experimentos con datos masivos, la ciencia social computacional debe servirse de una combinación sensata de conjuntos de datos, teorías y contraste experimental a gran escala (pudiendo así modelizar de manera fidedigna distintas problemáticas de la realidad social).

3.3. Los métodos digitales, el análisis de controversias y la analítica cultural. ¿Propuestas intermedias o renuncias implícitas?

A partir de dichas contribuciones fueron surgiendo otras alternativas desde las Ciencias Sociales ante la conceptualización hegemónica del Big Data. Entre ellas, algunas de las más relevantes han sido los 'métodos digitales' (Rogers, 2013), el 'análisis de controversias' (Marres, 2015; Venturini, 2010; Venturini y Munk, 2021) o las 'analítica cultural' (Manovich, 2020).

Con ciertas diferencias, los tres enfoques tratan de estudiar tendencias sociales o condiciones culturales a través datos extraídos de la web. En el caso de los métodos digitales (Rogers, 2013), se pone de relieve el uso de las problemáticas derivadas de importar al ámbito digital métodos preexistentes (como sería la encuesta online) y se aboga por el uso de nuevos métodos y datos nativos o específicos del medio (indagado paralelamente sus estructuras, diseños y considerándolos parte del quehacer sociológico). Por su parte, el análisis de controversias (Marres y Rogers, 2005) implicaría el uso de técnicas computacionales para visibilizar y analizar distintas formas de contestación y discusión pública sobre asuntos de relevancia social y actualidad. Finalmente, la analítica cultural (Manovich, 2013, 2020) se centrarían en el análisis computacional de datos culturales (como fotografías, vídeos o canciones), con un enfoque especialmente sensible a las producciones de medios visuales.

Este tipo de propuestas han tratado de solventar de diferentes modos -y con diversas limitaciones- las dificultades de las ciencias sociales para integrar los datos transaccionales en sus análisis y métodos. En cierto sentido, se trata de propuestas intermedias que, aceptando las dificultades anteriormente mencionadas, y aspiran a acortar distancias con el Big Data, incorporándolo en su propuesta (aunque de manera parcial y limitada). Sin embargo, dicho posicionamiento implica una renuncia implícita: la aceptación de que son las corporaciones privadas las que

detentan el gobierno del 'dato' a través de procesos de concentración y monopolio de los *datasets*. Así, en la práctica, nuestros abordajes y herramientas se limitarían a adaptarse a las pequeñas parcelas de información a las que permiten acceder.

Pongamos un ejemplo ilustrativo. Twitter es uno de los medios sociales predilectos para el análisis social y político y uno de los medios que facilitan la extracción masiva de información a través de sus diferentes APIs (*Application Programming Interfaces*). Dependiendo del tipo de búsqueda, la información que se obtiene a través de esta interfaz rápidamente puede exceder los límites de almacenamiento de un ordenador estándar, algo extremadamente raro cuando se analizan encuestas o materiales cualitativos. Además, el formato más utilizado para el almacenamiento y gestión de la información, JSON, no se corresponde con ningún tipo de dato habitualmente empleado. En este caso, Twitter funciona como el conjunto de población a quien se le realizan las preguntas de investigación y la forma que tiene de 'respondernos' es a través de una descarga de información más o menos masiva y desordenada: se encuentra información cualitativa (el texto del tweet o la imagen que contenía), cuantitativa (cantidad de re-tweets, favoritos, likes, menciones, etc.), y de tipo relacional (id. de los re-tweets, id de las menciones, usuarios que respondieron, etc.). Es decir, que la descarga de información contiene un conjunto diverso de datos susceptibles de ser analizados cuantitativa y cualitativamente, en sus dimensiones micro y/o macro, de manera situada y/o agregada. Un 'todo a la vez' que difumina completamente las dicotomías, distinciones y limitaciones de los tipos de análisis clásicos.

Ahora bien, a pesar de ser una de las plataformas que más facilidades y cantidad de descargas ofrece, lo que puede parecer una ingente cantidad de información no deja de ser, en muchas ocasiones, una parte ínfima del conjunto. Sus versiones más generosas permiten como máximo descargar diez millones de tweets al mes, en el caso de las cuentas académicas o el 1% del total de tweets en tiempo real. No cabe duda de que, dependiendo del objeto de estudio, estas cifras pueden ser más que suficientes, pero distan mucho de resultarlo para los grandes temas sociales y culturales a los que el análisis del Big Data prometía dar respuesta.

En definitiva, el hecho de que sean escasos los equipos y grupos de investigación con capacidad para gestionar estos flujos de información no entra en contradicción con la realidad de que la comunidad investigadora tenga una falsa sensación de abundancia respecto a los datos que maneja, su representatividad y alcance. Un simple ejemplo: a pesar de la vasta bibliografía sobre la proliferación de polarización política dentro de la plataforma, sólo el equipo de Twitter es capaz de constatar que su algoritmo de selección de contenido propaga en mayor medida producciones y contenidos vinculados a la derecha que a la izquierda política (Chowdhury y Belli, 2021). Así, el énfasis de gran parte de la literatura centrada en la algoritmia sólo capta de forma parcial una realidad más compleja. Es la capacidad de acceso a una realidad concreta la que determina, en primera instancia, la posibilidad de analizarla. Por tanto, una de las cuestiones que la comunidad académica debería plantearse es la pertinencia de perseverar en la construcción de modelos a partir de conjuntos de datos siempre parciales, migajas, que además pueden llegar a ser efímeras ante un

hipotético (pero muy real) cambio de las condiciones de acceso o políticas concretas de la empresa propietaria de la herramienta tecnológica de turno.

4. Conclusión: ¿Y ahora? Posiciones para investigar desde la generación artificial de escasez de información del capitalismo informacional

En el paradigma del Big-Data la forma de acceder a la realidad es a través de los datos previamente almacenados con un objetivo en principio diferente al de registrar de manera delimitada una realidad social. Por tanto, no se trata de datos organizados u ordenados con anterioridad. De hecho, dentro del capitalismo cognitivo es un lugar común registrar absolutamente toda la información que sea posible, cuestión que produce un segundo cambio: la eliminación de dicotomías estancas y clásicas como las investigaciones cuantitativas *vs*. cualitativas, micro *vs.* macro, etc. (Venturini et al. 2017).

Dichas cuestiones plantean problemas epistemológicos inherentes tanto al alcance como a la concepción del Big Data. Porque, en definitiva, si bien se puede estar dando todos los rodeos posibles en torno a cuestiones técnicas y metodológicas, uno de los principales problemas es el de la accesibilidad a los conjuntos de datos (Burrows y Savage, 2014; Venturini y Rogers, 2019). Aún más preocupante resulta que estos *datasets* sean controlados por unas pocas corporaciones, en un proceso de concentración progresiva de los repositorios de datos. Tan importante es interactuar críticamente con las fuentes de información como hacer campañas para acceder a dichos datos (Burrows y Savage, 2014).

En este sentido, Twitter podría calificarse como un caso dentro de los menos malos, y quizás por ello sea una plataforma tan ampliamente utilizada. Por ejemplo, en la actualidad es imposible acceder a datos de cualquiera de las aplicaciones de citas y los estudios realizados en torno al tema han sido parciales, como en el caso de Tinder (Duportail, 2019). En definitiva, sólo Google tiene acceso al conjunto de palabras (con sus errores ortotipográficos típicos) que se teclean en los móviles Android; únicamente Amazon tiene acceso al conjunto real de datos sobre los pedidos de productos; y solamente VISA y Mastercard pueden realizar el perfilado de clientes. El ejemplo inicial del INE pone de manifiesto que el manido discurso sobre los miedos y reticencias de la población a que el Estado tenga determinado acceso a los datos, típico de las sociedades de control foucaultianas y con cierta preocupación por la privacidad, ha pasado. Parafraseando a Foucault (2019), más que en sociedades disciplinarias nos encontramos en el siglo deleuziano de las sociedades del control. La incentivación y compulsión permanentes a la interacción y producción de contenidos produce trazas que son registradas por todo tipo de agentes diferentes a los Estados, relegados a un plano más que secundario.

Por ello, dentro de la concepción Big Data como un fenómeno sociotécnico es común encontrar llamamientos a cuestionar críticamente sus supuestos, potencialidades y sesgos. Es urgente entender los datos masivos como un fenómeno cultural, producto de la unión de conocimiento y tecnología que no está exenta de cierta aura de veracidad y objetividad, que provoca a su vez toda suerte de retóricas (Boyd y Crawford, 2012). Quizás, paradójicamente, esa sea la mayor derrota de la

sociología: haber llegado tarde a la construcción 'desde dentro' del mayor aparataje de medición social jamás construido. El contexto actual de un Internet y *datasets* cada vez más centralizados y concentrados, con unas instituciones estatales cada vez más debilitadas, deja gran parte de las posibilidades del conocimiento sociológico excesivamente limitadas y con un tratamiento parcial.

5. Declaración de la contribución por autoría

Héctor Puente: Conceptualización, Metodología, Investigación, Análisis formal, Visualización, Redacción –borrador original, Redacción – revisión y edición.
Diego de Haro: Conceptualización, Metodología, Investigación, Análisis formal, Visualización, Redacción –borrador original, Redacción – revisión y edición.
Sergio D'Antonio: Conceptualización, Metodología, Investigación, Análisis formal, Visualización, Redacción –borrador original, Redacción – revisión y edición.

6. Referencias

Boletín Oficial del Estado (2020). *Anuncio de formalización de contratos de: Instituto Nacional de Estadística INE. Objeto: Estimación de los flujos de turismo interno y turismo receptor a partir del posicionamiento de los teléfonos móviles (Vodafone). Expediente: 2020N0056003.* BOE-B-2020-42528. https://www.boe.es/diario_boe/txt.php?id=BOE-B-2020-42528

Boletín Oficial del Estado (2020). *Anuncio de formalización de contratos de: Instituto Nacional de Estadística INE. Objeto: Estimación de los flujos de turismo interno y turismo receptor a partir del posicionamiento de los teléfonos móviles (Telefónica). Expediente: 2020N0056002.* BOE-B-2020-46213. https://www.boe.es/diario_boe/txt.php?id=BOE-B-2020-46213

Boletín Oficial del Estado (2020). *Anuncio de formalización de contratos de: Instituto Nacional de Estadística INE. Objeto: Estimación de los flujos de turismo interno y turismo receptor a partir del posicionamiento de los teléfonos móviles (Orange). Expediente: 2020N0056001.* BOE-B-2020-47419. https://www.boe.es/diario_boe/txt.php?id=BOE-B-2020-47419

Boletín Oficial del Estado (2021). *Anuncio de formalización de contratos de: Instituto Nacional de Estadística INE. Objeto: Análisis de la movilidad durante la pandemia por Covid-19 (Estudio de movilidad EM-4) a partir de la base de datos Flux Vision. Expediente: 2021N0060001. BOE-B-2021-18242* https://www.boe.es/diario_boe/txt.php?id=BOE-B-2021-18242

Boletín Oficial del Estado (2021). *Anuncio de formalización de contratos de: Instituto Nacional de Estadística INE. Objeto: Análisis de la movilidad durante la pandemia por Covid-19 (Estudio de movilidad EM-4) a partir de la base de datos Vodafone Analytics. Expediente: 2021N0060003.* BOE-B-2021-25336. https://www.boe.es/diario_boe/txt.php?id=BOE-B-2021-25336

Boyd, Danah y Crawford, Kate (2012). CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon.

Information, Communication & Society, *15*(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

Burrows, Roger y Savage, Mike (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*, *1*(1), 205395171454028. https://doi.org/10.1177/2053951714540280

Chowdhury, Rumman y Belli, Luca (2021, 21 de octubre). Examining algorithmic amplification of political content on Twitter. *Twitter company blog*. https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent

Christoffersen, Peter F. y Diebold, Francis X. (2000). How relevant is volatility forecasting for financial risk management? *Review of Economics and Statistics*, *82*(1), 12-22. https://www.jstor.org/stable/2646668

Conte, Rosaria, Gilbert, Nigel, Bonelli, Giulia, Cioffi-Revilla, Claudio, Deffuant, Guillaume, Kertesz, Jannos, Loreto, Vittorio, Moat, Suzy, Nadal, J.-P, Sanchez, Anxo, Nowak, Andrezj, Flache, Andreas, San Miguel, Maxi y Helbing, Dirk (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, *214*(1), 325-346. https://doi.org/10.1140/epjst/e2012-01697-8

Diebold, Francis X. (2012). On the origin(s) and development of the term «Big Data». *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2152421

Duportail, Judith (2019). *El algoritmo del amor: Un viaje a las entrañas de Tinder*. Contra.

Gómez-Cruz, Edgar (2020, 4 de septiembre). *Etnografía Digital: Del Ciberespacio a la cultura algorítmica*. Youtube. https://www.youtube.com/ watch?v =us 648 G3 XAfE

Foucault, Michel (2019). Theatrum philosophicum. En Donald F. Bouchard (Ed.), *Language, counter-memory, practice*. (pp. 165-196). Cornell University Press. https://doi.org/10.1515/9781501741913-009

Ginsberg, Jeremy, Mohebbi, Matthew H., Patel, Rajan S., Brammer, Lynnette, Smolinski, Mark S. y Brilliant, Larry (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012-1014. https://doi.org/10.1038/nature07634

Gómez-Cruz, Edgar (2007). *Las metáforas de Internet*. https://dialnet.unirioja.es/servlet/libro?codigo=614871

Gómez-Cruz, Edgar (2020, septiembre 4). *Etnografía Digital: Del Ciberespacio a la cultura algorítmica*. Youtube. https://www.youtube.com/watch?v=us648G3XAfE

Hine, Christine (2000). *Virtual ethnography*. Sage.

Hine, Christine (2015). *Ethnography for the Internet: Embedded, embodied and everyday*. Routledge.

INE (2019). Estadística Piloto sobre Movilidad a partir del posicionamiento de teléfonos móviles (Censos de Población y Viviendas 2021). (p. 11) INE. https://ine.es/censos2021/movilidad_proyecto.pdf

INE (2021). *Estudios de movilidad a partir de la telefonía 2020-2021 Proyecto técnico* (p. 9) [Proyecto técnico] INE. https://www.ine.es/experimental/movilidad/exp_em_proyecto.pdf

Jones, Steve (1999). *Doing internet research: Critical issues and methods for examining the net*. Sage.

Kitchin, Rob y McArdle, Gavin (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, 205395171663113. https://doi.org/10.1177/2053951716631130

Latour, Bruno (2021). *La esperanza de pandora: Ensayos sobre la realidad de los estudios de la ciencia*. Editorial Gedisa.

Lazer, David, Kennedy, Ryan, King, Gary y Vespignani, Alessandro (2014). the parable of google flu: Traps in Big Data analysis. *Science*, *343*(6176), 1203-1205. https://doi.org/10.1126/science.1248506

Manovich, Lev (2013). *Software takes command: Extending the language of new media*. Bloomsbury.

Manovich, Lev (2020). *Cultural analytics*. The MIT Press.

Marr, Bernard (2014, 6 de marzo). Big Data: The 5 Vs Everyone Must Know. [LinkedIn] *LinkedIn Personal Blog*. https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know

Marres, Noortje (2015). Why map issues? On controversy analysis as a digital method. *Science, Technology, & Human Values*, *40*(5), 655-686. https://doi.org/10.1177/0162243915574602

Marres, Noortje y Rogers, Richard (2005). Recipe for tracing the fate of issues and their publics on the web. En B. Latour y P. Weibel (Eds). Making Things Public: Atmospheres of Democracy (pp. 922-935). MIT Press. https://research.gold.ac.uk/id/eprint/6548/

Mayer-Schönberger, Viktor, Cukier, Kenneth y Iriarte, Antonio (2013). *Big Data: La revolución de los datos masivos*. Turner.

McNulty, Eileen (2014, mayo 22). *Understanding Big Data: The seven v's - dataconomy*. https://dataconomy.com/2014/05/seven-vs-big-data/

Pink, Sarah, Horst, Heather, Postill, John, Hjorth, Larissa, Lewis, Tania y Tacchi, Jo (2015). *Digital ethnography: Principles and practice*. (1st edition) Sage Publications Ltd.

Rheingold, Howard (1993). *The virtual community: Finding commection in a computerized world*. Addison-Wesley Longman Publishing Co., Inc.

Rice, Ronald E. (1989). Issues and Concepts in Research on Computer-Mediated Communication Systems. *Annals of the International Communication Association*, *12*(1), 436-476. https://doi.org/10.1080/23808985.1989.11678731

Rodrigo, Borja (2019, noviembre 20). *Es falso que el INE nos «catalogue e intervenga» el teléfono mediante el estudio de movilidad que arrancó el lunes* Newtral. https://www.newtral.es/es-falso-que-el-ine-nos-catalogue-e-intervenga-el-telefono-mediante-el-estudio-de-movilidad-que-arranco-el-lunes/20191120/

Rogers, Richard (2013). *Digital methods*. MIT Press.

Sánchez Holgado, Patricia (2021). *La comunicación de la ciencia de datos en España*. https://doi.org/10.14201/gredos.149448

Savage, Mike y Burrows, Roger (2007). The Coming Crisis of Empirical Sociology. *Sociology*, *41*(5), 885-899. https://doi.org/10.1177/0038038507080443

Savage, Mike y Burrows, Roger (2009). Some Further Reflections on the Coming Crisis of Empirical Sociology. *Sociology*, *43*(4), 762-772. https://doi.org/10.1177/0038038509105420

Thomas, Tom, Geurs, Karst T., Koolwaaij, Johan y Bijlsma, Marcel (2018). Automatic Trip Detection with the Dutch Mobile Mobility Panel: Towards Reliable Multiple-Week Trip Registration for Large Samples. *Journal of Urban Technology*, *25*(2), 143-161. https://doi.org/10.1080/10630732.2018.1471874

Turkle, Sherry (1995). *La vida en la pantalla: La construcción de la identidad en la era de internet*. Paidós Ibérica.

Venturini, Tommaso (2010). Diving in magma: how to explore controversies with actor-network theory. *Public Understanding of Science*, *19*(3), 258-273. https://doi.org/10.1177/0963662509102694

Venturini, Tommaso (2012). Building on faults: How to represent controversies with digital methods. *Public Understanding of Science*, *21*(7), 796-812. https://doi.org/10.1177/0963662510387558

Venturini, Tommaso, Jacomy, Mathieu, Meunier, Axel y Latour, Bruno (2017). An unexpected journey: A few lessons from sciences Po médialab's experience. *Big Data & Society*, *4*(2), 205395171772094. https://doi.org/10.1177/2053951717720949

Venturini, Tommaso y Latour, Bruno (2009, mayo). The Social Fabric. *Futur En Seine 2009*. https://hal-sciencespo.archives-ouvertes.fr/hal-01293394

Venturini, Tommaso y Munk, Anders Kristian (2021). *Controversy mapping: A field guide*. (1st edition) Polity.

Venturini, Tommaso y Rogers, Richard (2019). "API-Based Research" or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge Analytica Data Breach. *Digital Journalism*, *7*(4), 532-540. https://doi.org/10.1080/21670811.2019.1591927

Watts, Duncan J. (2014). Computational Social Science: Exciting Progress and Future Directions. En *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2013 Symposium*. (pp. 17-24). National Academies Press. http://www.nap.edu/catalog/18558

Weiss, Sholom M. y Indurkhya, Nitin (1998). *Predictive data mining: a practical guide*. Morgan Kaufmann.